

Xin-She Yang  
R. Simon Sherratt  
Nilanjan Dey  
Amit Joshi *Editors*

# Proceedings of Eighth International Congress on Information and Communication Technology

ICICT 2023, London, Volume 2



Springer

# **Lecture Notes in Networks and Systems**

Volume 694

## **Series Editor**

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,  
Warsaw, Poland

## **Advisory Editors**

Fernando Gomide, Department of Computer Engineering and Automation—DCA,  
School of Electrical and Computer Engineering—FEEC, University of  
Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering,  
Bogazici University, Istanbul, Türkiye

Derong Liu, Department of Electrical and Computer Engineering, University of  
Illinois at Chicago, Chicago, USA

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of  
Alberta, Alberta, Canada

Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering,  
KIOS Research Center for Intelligent Systems and Networks, University of Cyprus,  
Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong,  
Kowloon, Hong Kong



The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose ([aninda.bose@springer.com](mailto:aninda.bose@springer.com)).

Xin-She Yang · R. Simon Sherratt · Nilanjan Dey ·  
Amit Joshi  
Editors

# Proceedings of Eighth International Congress on Information and Communication Technology

ICICT 2023, London, Volume 2

*Editors*

Xin-She Yang  
Department of Design Engineering  
and Mathematics  
Middlesex University London  
London, UK

Nilanjan Dey  
Department of Computer Science  
and Engineering  
Techno International Newtown  
Chakpachuria, West Bengal, India

R. Simon Sherratt  
Department of Biomedical Engineering  
University of Reading  
England, UK

Amit Joshi  
Global Knowledge Research Foundation  
Ahmedabad, India

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-981-99-3090-6

ISBN 978-981-99-3091-3 (eBook)

<https://doi.org/10.1007/978-981-99-3091-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.  
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

The Eighth International Congress on Information and Communication Technology will be held during February 20–23, 2023, in a hybrid mode, physically at London, UK, and Digital Platform: Zoom. ICICT 2023 organized by Global Knowledge Research Foundation and managed by GR Scholastic LLP. The associated partners were Springer and InterYIT IFIP. The conference will provide a useful and wide platform both for the display of the latest research and for the exchange of research results and thoughts. The participants of the conference will be from almost every part of the world, with backgrounds of either academia or industry, allowing a real multinational multicultural exchange of experiences and ideas.

A great pool of more than 1300 papers were received for this conference from across 113 countries among which around 361 papers were accepted and will be presented physically at London and Digital Platform: Zoom during the 4 days. Due to the overwhelming response, we had to drop many papers in the hierarchy of the quality. Total of 46 technical sessions will be organized in parallel in 4 days along with a few keynotes and panel discussions in hybrid mode. The conference will be involved in deep discussion and issues which will be intended to solve at global levels. New technologies will be proposed, experiences will be shared, and future solutions for design infrastructure for ICT will also be discussed. The final papers will be published in four volumes of proceedings by Springer LNNS Series. Over the years, this congress has been organized and conceptualized with the collective efforts of a large number of individuals. I would like to thank each of the committee members

and the reviewers for their excellent work in reviewing the papers. Grateful acknowledgments are extended to the team of Global Knowledge Research Foundation for their valuable efforts and support.

I look forward to welcoming you to the 8th Edition of this ICICT Congress 2023.

Amit Joshi, Ph.D.  
Organising Secretary, ICICT 2023  
Director—Global Knowledge Research  
Foundation  
Ahmedabad, India

# Contents

<b>Development of a Method for Reducing the Impact of Metal Interconnection Parameters on the Speed of VLSI</b> .....	1
Narek Avdalyan and Armen Petrosyan	
<b>Deep Learning-Based Arrhythmia Detection Using RR-Interval Framed Electrocardiograms</b> .....	11
Song-Kyoo Kim, Chan Yeob Yeun, Paul D. Yoo, Nai-Wei Lo, and Ernesto Damiani	
<b>How Predictive Software Engineering Addresses Issues in Custom Software Development and Boosts Efficiency and Productivity</b> .....	23
Boris Kontsevoi, Sergey Kizyan, and Irina Dubovik	
<b>Using Conceptual Chunking to Support Information Processing While Solving Complex Industrial Tasks</b> .....	33
Anja Klichowicz, Tina Morgenstern, and Franziska Bocklisch	
<b>A Mobile Application Innovation for Public Healthcare Supply Chain Coordination</b> .....	51
Marcia Mkansi and Tshililo Ramovha	
<b>A Social Critical Analysis on Philippine Higher Education in the Time of COVID-19 Pandemic Toward a Framework on Flexible Learning</b> .....	61
Alvin A. Sario, Elcid A. Serrano, and Ramon L. Rodriguez	
<b>M-HEALTH System for Detecting COVID-19 in Chest X-Rays Using Deep Learning and Data Security Approaches</b> .....	73
Johnny Delgado, Luis Clavijo, Carlos Soria, Juan Ortega, and Sebastian Quevedo	

<b>Interpolated Solutions of Abel Integral Equations Using Barycentric Lagrange Double Interpolation .....</b>	<b>87</b>
E. S. Shoukralla and B. M. Ahmed	
<b>Procurement of the Future: Investing Today in the Technologies of Tomorrow .....</b>	<b>97</b>
Elizabeth Koumpan and Anna W. Topol	
<b>Development and Validation of a Health Information System to Improve Prenatal Controls in Guatemala .....</b>	<b>109</b>
Ignacio Prieto-Egido, Aitor Garrido Madrigal, and Cristina Barrena García	
<b>Adapting Atmospheric Chemistry Components for Efficient GPU Accelerators .....</b>	<b>129</b>
Christian Guzman Ruiz, Matthew Dawson, Mario C. Acosta, Oriol Jorba, Eduardo Cesar Galobardes, Carlos Pérez García-Pando, and Kim Serradell	
<b>Frequency Interleaved DAC System Design: Fundamental Problems and Compensation Methods .....</b>	<b>139</b>
Nagito Ishida, Koji Asami, Shogo Katayama, Anna Kuwana, and Haruo Kobayashi	
<b>Neural Network Models for Time Series Analysis and Estimation .....</b>	<b>159</b>
Louay Al Nuaimy	
<b>An Approach for Test Impact Analysis on the Integration Level in Java Programs .....</b>	<b>171</b>
Muzammil Shahbaz	
<b>ANN-Based Modeling and Control of a Pick and Place Manipulator .....</b>	<b>189</b>
Mohamed Essam Mostafa, Aya Essam Mostafa, Hossam Hassan Ammar, and Raafat Shalaby	
<b>Toward Learning Analytics in a Distributed Learning Environment .....</b>	<b>205</b>
Dijana Oreski, Vjeran Strahonja, and Darko Androcec	
<b>Influence and Optimization of Power Grid ERP System Permission Management on Enterprise Internal Control .....</b>	<b>215</b>
Zhu Zuoping, Zhang Wei, Huang Yao, and Chen Tianxiao	
<b>Ensemble Feature Selection and Classification of Medical Dataset Using K-Nearest Classifier with Swarm Intelligence .....</b>	<b>231</b>
Ebtesam Shadadi, Saahira Banu Ahamed, Latifah Alamer, Mousa Khubrani, Iman Mohammad Alqahtani, and Aisha Sumaili	



<b>Yet Another Parallelism Within the “Hobby Time Training”</b> .....	245
Milen Loukantchevsky	
<b>An Improved Apriori Algorithm for Interestingness of Association Rules: A Case Study on the Mushroom Dataset</b> .....	255
Huynh Anh Duy, Bui Trong Vinh, and Phan Duy Hung	
<b>Hybrid Network Anomaly Detection Based on Weighted Aggregation Using Endpoint Parameters</b> .....	269
L. Y. Dobkacz, S. A. Sakulin, A. N. Alfimtsev, and Y. A. Kalgin	
<b>IoT Infrared Imaging of Livestock Tissues Using a One-Eyed Bandit Technique</b> .....	279
Stefan Rizanov, Peter Yakimov, and Dimitar Todorov	
<b>The Digital Survival Game to Enhance the Digital Quotient of Lower Secondary Students</b> .....	293
Amornphong Suksen and Nutteerat Pheeraphan	
<b>An Experimental Analysis of Benchmarking Tools for Smart Contract-Based Blockchain Application</b> .....	309
Deepa Kumari, Chirag Jain, Aman Saxena, Pranjali Gupta, Ashay Netke, and Subhrakanta Panda	
<b>Digital Twins in Agriculture as an Internet of Things Paradigm: The Case of Azerbaijan</b> .....	321
Fuad Ibrahimov, Ulviyya Rzayeva, and Rasul Balayev	
<b>Theoretical Fundamentals of Criteria for Evaluation of Efficiency, Quality and Optimization of Complex Informatology Systems</b> .....	329
Volodymyr Kulivnuk, Ivan Kuzmin, Oleksandr Hladkyi, Alexander Gertsy, Tetiana Tkachenko, and Tetiana Shparaga	
<b>Detection of Structure Changes in Lightweight Concrete with Nanoparticles Using Computer Vision Methods in the Construction Industry</b> .....	339
Roman Mysiuk, Volodymyr Yuzevych, Bohdan Koman, Yuriy Tyrkalo, Oleksandra Farat, Iryna Mysiuk, and Lyudmyla Harasym	
<b>Mild Cognitive Impairment Screening System by Multiple Daily Activity Information—A Method Based on Daily Conversation</b> .....	349
Ayaka Yamanaka, Ikuma Sato, Shuichi Matsumoto, and Yuichi Fujino	
<b>System Models of a Single Information Space of Digital Twins</b> .....	361
Mykola Korablyov and Sergey Lutsyy	
<b>Creating a Happy Life Through Body Sensations</b> .....	373
Shuichi Fukuda	

<b>Virtual Training System for the Autonomous Navigation of an Omnidirectional Traction Robot</b> .....	383
De La Cruz Aida, Tapia Edison, and Víctor H. Andaluz	
<b>NFTs: Inside the Twitter Discussion</b> .....	397
Victor Hernández-Manrique, Rodrigo Carmona-Herrera, Francisco J. Cantú-Ortiz, and Héctor G. Ceballos-Cancino	
<b>Integrating Analog PIR Sensor Telemetry with TinyML Inference for On-The-Edge Classification of Moving Objects</b> .....	405
Ritha M. Umutooni, Marvin Ogore, Damien Hanyurwimfura, and Jimmy Nsenga	
<b>Advanced Signaling Mechanisms for Assurance of User Service Continuity in 4G/5G Mobile Network</b> .....	417
Diep Pham Quang, Hung Nguyen Tai, Hoan Nguyen Dac, and Tu Le Minh	
<b>The BB84 Quantum Key Distribution Protocol and Potential Risks</b> .....	429
Maria E. Sabani, Ilias K. Savvas, Dimitrios Poulakis, George C. Makris, and Maria A. Butakova	
<b>All Vaccinated: Open-Source Web System for the Control of Vaccination Processes in Health Centers</b> .....	439
Lucrecia Llerena, Nancy Rodríguez, Ana Osorio, Rino Arias, and John W. Castro	
<b>Centralized Tasks Scheduling and Load Balancing on a Cloudlet</b> .....	451
Manoj Subhash Kakade, Anupama Karuppiah, Samarth Agarwal, Mudigonda Sreevastav, Obulreddigari Gayathri, V. Ranjith, Sista Kasi Vishwanath, and Gaurav Basu	
<b>A Digital Twin Enabled Decision Support Framework for Ship Operational Optimisation Towards Decarbonisation</b> .....	467
Antonis Antonopoulos, Bill Karakostas, Takis Katsoulakos, Anargyros Mavrakos, Theodosia Tsaousis, and Stathis Zavvos	
<b>Financial Sustainability of Automotive Software Compliance and Industry Quality Standards</b> .....	477
Pavle Dakić, Vladimir Todorović, and Valentino Vranić	
<b>A Novel Multiband Patch Antenna Based on the Modification of a Rectangular Design</b> .....	489
Rafael B. Méndez-Vásquez, Marcelo D. Lojano-Angamarca, Luis F. Guerrero-Vásquez, Jorge O. Ordoñez-Ordoñez, and Paul A. Chasi-Pesantez	

<b>Labour Conditions and Their Impact on the Development of Green Economies in 2020</b> .....	499
María Fernanda Romo-Fuentes, Francisco J. Cantú-Ortiz, and Héctor G. Ceballos-Cancino	
<b>A Review of Deep Learning Techniques of Chest X-ray Analysis for Thoracic Disorders</b> .....	509
Pawan Sharma, S. Gurunarayanan, and Anupama Karuppiah	
<b>Multi-task Learning Method Using Emoji Prediction as Auxiliary Task for Sentiment Analysis</b> .....	521
Haruki Asano and Masafumi Matsuhara	
<b>Smart Cities Improving Government Management Systems with Blockchain Technology</b> .....	535
Marciele Berger Bernardes, Francisco Pacheco de Andrade, and Lucas Cortizo	
<b>Simple Moving Average (SMA) Investment Strategy During COVID-19 Pandemic</b> .....	545
Juan P. Licona-Luque, Luis F. Brenes-García, Francisco J. Cantú-Ortiz, and Héctor G. Ceballos-Cancino	
<b>Yield Prediction of Maize Using Random Forest Algorithm</b> .....	557
Jane Kristine G. Suarez and Luisito Lolong Lacatan	
<b>Enhancement of Prototype Driving Simulator Using Available AI-Based Game Technology</b> .....	569
Yun-Quan Cheng, Sarina Mansor, Ji-Jian Chin, Hezerul Abdul Karim, and Ban Kar-Weng	
<b>The QOM Toolbox: An Object-Oriented Python Framework for Cavity Optomechanical Systems</b> .....	581
Sampreet Kalita and Amarendra K. Sarma	
<b>Preserving Filipino Native Dishes Using Android-Based Application: A Heritage Cooking Tutorial</b> .....	591
Aries M. Gelera, Alyssa Joi A. Gonzales, Bryan James V. Torres, and Marvin G. Sison	
<b>Development of a Web-Based Graduate Tracer Information System with Data Analytics</b> .....	601
Karlo Jose E. Nabablit and Edgardo S. Dajao	
<b>Distance Education Opportunities for the Elderly in Thailand: Opportunity to Access Distance Education and Factors Affecting Such Opportunity</b> .....	613
Phisit Nadprasert, Chanoknart Boonwatthanakul, Supanita Sudsaward, Duangbhorn Sapphayalak, and Likkhasit Putkhiao	

**Motivation Prediction for Persuasive Intervention at Appropriate Timing to Promote Exercises ..... 629**  
Tomoya Yuasa, Fumiko Harada, and Hiromitsu Shimakawa

**Implications of 3D Printing on Physical Distribution in Logistics and Supply Chain Management ..... 641**  
Patrick Brandtner, Robert Zimmermann, and Jessika Allmendinger

**Towards Prototyping Single-modal and Multimodal Interactions in Mixed Reality Games ..... 655**  
Logan LaMont, Ged Fuller, Pratheep Kumar Paranthaman, Thomas Poteat, Dhvani Toprani, Qian Xu, and Nikesh Bajaj

**Possibility of Utilising Information Technology to Promote Local Production for Local Consumption of Agricultural Products and Future Challenges ..... 667**  
Tomoko Kashima, Takashi Hasuike, and Shimpei Matsumoto

**A Data Analysis of Video Game Reviews on Steam ..... 683**  
Shuyao Cai, Sunyi Zhang, Lin Zhu, and Yanxia Jia

**Non-destructive Technique for Agricultural Seed Classification Using Deep Learning ..... 695**  
Deepali B. Koppad, K. V. Suma, N. Nethra, and C. S. Sonali

**A Hybrid Strategy for DoS Attacks Detection and Mitigation on SDN Enabled Real Scenarios ..... 705**  
Jaime Vergara, Christian Garzón, and Juan Felipe Botero

**1001 Games a Night—Continuous Evaluation of an Intelligent Multi-agent-Based System ..... 715**  
Eicke Godehardt, Mohamed Amine Allani, Alexander Julian Vieth, and Thomas Gabel

**Real-Time Hand Action Detection and Classification Based on YOLOv7 from Egocentric Videos ..... 723**  
Van-Hung Le

**Interaction-Driven Design: A Case Study of Interactive Lighting ..... 733**  
Cun Li and Qiao Liang

**Sustainable Technologies for Environment-Friendly and Ecological Resilience ..... 745**  
Paul M. Cabacungan, Khim Cathleen M. Saddi, Maria Theresa Joy G. Rocamora, Reymond P. Cao, Salvador P. Granada, Paul Ryan A. Santiago, Neil Angelo M. Mercado, Carlos M. Oppus, Cristina F. Gonzales, Nathaniel Joseph C. Libatique, Emma E. Porio, and Gregory L. Tangonan

**Preliminary Investigation into a Security Approach  
for Infrastructure as Code ..... 763**  
Ammar Zeini, Ruth G. Lennon, and Patrick Lennon

**Use of Artificial Intelligence in the Digital Marketing Strategy  
of Latvian Companies ..... 785**  
Jelena Salkovska, Anda Batraga, Liene Kaibe, and Katrina Kellerte

**SR-OIR-SSD: Super-Resolved Eyes in the Sky ..... 799**  
Raghav Sharma and Rohit Pandey

**Participatory Design as an Audiovisual Strategy in Brand  
Manuals ..... 811**  
Carlos Borja-Galeas and Hugo Arias-Flores

**Citizen Engagement on Government Social Media: Validation  
of Measurement Items ..... 819**  
Ari Wedhasmara, Samsuryadi, and Ab Razak Che Hussin

**Shared Parking Concept in the Smart City Environment ..... 833**  
Zuzana Špitálová, Lucia Mandová, and Martin Opatovský

**Applying Machine Learning Techniques to the Analysis  
and Prediction of Financial Data ..... 843**  
Pablo Flores-Siguenza, Darío Espinoza-Saquicela,  
Marlon Moscoso-Martínez, and Lorena Siguenza-Guzman

**Time Series Analysis of Public Opinion on Work from Home  
During and After COVID-19 Pandemic ..... 855**  
Gabriela G. Mendoza-Leal, Jorge A. Mendez-Vargas,  
Francisco J. Cantú-Ortiz, and Héctor G. Ceballos-Cancino

**Determination of Air Quality with Unmanned Vehicles  
in Cement Plants ..... 867**  
Diego Verdugo-Ormaza, Jean P. Mata-Quevedo,  
Ricardo Romero Gonzalez, and Luis Serpa-Andrade

**The Determinants of ICT Use by University Professors ..... 879**  
Mounir Elatrachi and Samira Oukarfi

**Load Capacity Study on the Flora Path of the Manglares  
Churute Ecological Reserve ..... 895**  
Miriam Vanessa Hinojosa-Ramos, Marcelo Leon, Paulina Leon,  
Viviana Tomala, and José Maldonado-Quezada

**Convergent Fuzzy Cognitive Modelling of Regional Youth Policy  
Strategy ..... 911**  
Aleksandr Raikov

**Realistic Modeling of Computer Systems in Gem5 Simulator ..... 923**  
Amit Mankodi

**Construction Method of Operational Concept Model Based on Architecture Framework** ..... 939  
Jing An, Lei Zhang, Miaoting Zeng, and Xu Han

**Representation Learning with Attention for Spatial Reuse Optimization in Dense WLANs** ..... 949  
Stephen Azeez and Shagufta Henna

**A New Ultralightweight Authentication Protocol for IoTs: MFRAP** ..... 961  
Umar Mujahid and Binh Tran

**Development and Implementation of a Scalable and Replicable Industrial Environment at Low Cost to Control an Industrial Process** ..... 971  
Serpa-Andrade Luis, Mata-Quevedo Paul, Guerrero-Vasquez Fernando, Garcia-Velez Roberto, and Gonzalez-Gonzalez Sandro

**Graph Embedding of Chronic Myeloid Leukaemia K562 Cells Gene Network Reveals a Hyperbolic Latent Geometry** ..... 979  
Paola Lecca, Angela Re, Giulia Lombardi, Roberta Valeria Latorre, and Claudio Sorio

**Trend of M-Health Research in the Self-management of Chronic Illness: Bibliometric Analysis** ..... 993  
Ade Komariah and Erna Rochmawati

**The Readiness of a Private Hospital Toward Smart Hospital in Indonesia** ..... 1003  
Nur Hidayah, Qurratul Aini, and Gofur Ahmad

**Coping with the Business Ethics Issues in the Era of the Internet of Things** ..... 1015  
Indah Fatmawati

**Sentiment Analysis: Predicting the Position of Islamic Political Parties in Indonesia in the Next Election** ..... 1027  
Hasse Jubba, Tawakkal Baharuddin, Zuly Qodir, and Suparto Iribaram

**Digital Leadership in the Development of Digital Competencies in Voter Education Service** ..... 1035  
Titin Purwaningsih, Bambang Eka Cahya Widodo, Moch Edward Trias Pahlevi, and Azka Abdi Amrurrobbi

**Methodology for the Implementation of FPGA in Technological Applications** ..... 1047  
Coronel-Villavicencio Edison, Serpa-Andrade Luis, and Garcia-Velez Roberto

**Technical Requirements Survey on Multimodal Biometric  
Selection for Deployment in Governments ..... 1057**  
Mapula Elisa Maeko and Dustin Van Der Haar

**A Path Recommender System for Enjoyment Improvement  
of the Cultural Heritage ..... 1075**  
Francesco Colace, Dajana Conte, Maria Pia D’Arienzo,  
Domenico Santaniello, Alfredo Troiano, and Carmine Valentino

**Artificial Intelligence Applications in Healthcare ..... 1085**  
Usman Ahmad Usmani, Ari Happonen, Junzo Watada,  
and Jayden Khakurel

**Maintaining Performance with Less Data: Understanding  
Useful Data ..... 1105**  
Dominic Sanderson and Tatiana Kalganova

**Author Index ..... 1129**



# Editors and Contributors

## About the Editors

**Xin-She Yang** obtained his DPhil in Applied Mathematics from the University of Oxford. He then worked at Cambridge University and National Physical Laboratory (UK) as Senior Research Scientist. Now he is Reader at Middlesex University London, Fellow of the Institute of Mathematics and its Application (IMA), and a Book Series co-Editor of the Springer Tracts in Nature-Inspired Computing. He was also the IEEE Computational Intelligence Society task force chair for Business Intelligence and Knowledge Management (2015–2020). He has published more than 25 books and more than 400 peer-reviewed research publications with over 78,600 citations, and he has been on the prestigious list of highly cited researchers (Web of Sciences) for seven consecutive years (2016–2022).

**R. Simon Sherratt** was born near Liverpool, England, in 1969. He is currently Professor of Biosensors at the Department of Biomedical Engineering, University of Reading, UK. His main research area is signal processing and personal communications in consumer devices, focusing on wearable devices and health care. He received the first place IEEE Chester Sall Memorial Award in 2006, the second place in 2016, and the third place in 2017.

**Nilanjan Dey** is an Associate Professor at the Department of Computer Science and Engineering, Techno International New Town, India. He is the Editor-in-Chief of the *International Journal of Ambient Computing and Intelligence*; a Series Co-Editor of Springer Tracts in Nature-Inspired Computing (STNIC), Data-Intensive Research (DIR), Springer Nature; and a Series Co-Editor of *Advances in Ubiquitous Sensing Applications for Healthcare*, Elsevier. He is a fellow of IETE and a Senior Member of IEEE.

**Amit Joshi** is currently the Director of Global Knowledge Research Foundation, and also an Entrepreneur and Researcher who has completed his master's and research in

the areas of cloud computing and cryptography in medical imaging. He has an experience of around 10 years in academic and industry in prestigious organizations. He is an active member of ACM, IEEE, CSI, AMIE, IACSIT-Singapore, IDES, ACEEE, NPA, and many other professional societies. Currently, he is the International Chair of InterYIT at International Federation for Information Processing (IFIP, Austria). He has presented and published more than 50 papers in national and international journals/conferences of IEEE and ACM. He has also edited more than 40 books which are published by Springer, ACM, and other reputed publishers. He has also organized more than 50 national and international conferences and programs in association with ACM, Springer, IEEE to name a few across different countries including India, UK, Europe, USA, Canada, Thailand, Egypt, and many more.

## Contributors

**Mario C. Acosta** Barcelona SuperComputing Center, Barcelona, Spain

**Samarth Agarwal** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India

**Saahira Banu Ahamed** Department of Computer Science, College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia

**Gofur Ahmad** Master of Management, Universitas Muhammadiyah Jakarta, Jakarta, Indonesia

**B. M. Ahmed** Faculty of Engineering and Technology, FUE in Egypt, Cairo, Egypt

**Qurratul Aini** Master of Hospital Administration, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Louay Al Nuaimy** Oman College of Management and Technology, Halban, Sultanate of Oman

**Latifah Alamer** Department of Information Technology and Security, College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia

**A. N. Alfimtsev** Bauman Moscow State Technical University, Moscow, Russia

**Mohamed Amine Allani** Frankfurt University of Applied Sciences, Frankfurt, Germany

**Jessika Allmendinger** University of Applied Sciences Upper Austria, Steyr, Austria

**Iman Mohammad Alqahtani** Computer and Information Systems Department, King Khalid University, Abha, Saudi Arabia

**Hossam Hassan Ammar** School of Engineering and Applied Science, NU, Giza, Egypt

**Azka Abdi Amrullohi** Komite Independen Sadar Pemilu (KISP), Yogyakarta, Indonesia

**Jing An** National Defense University, Beijing, China

**Víctor H. Andaluz** Universidad de Las Fuerzas Armadas ESPE, Sangolquí, Ecuador

**Darko Androcec** Faculty of Organization and Informatics, University of Zagreb, Varazdin, Croatia

**Antonis Antonopoulos** Konnecta, Newbridge, Ireland

**Hugo Arias-Flores** Centro de Investigación en Mecatrónica Y Sistemas Interactivos (MIST), Universidad Tecnológica Indoamérica, Quito, Ecuador

**Rino Arias** Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador

**Koji Asami** Advantest Corporation, Tokyo, Japan

**Haruki Asano** Iwate Prefectural University, Iwate, Japan

**Narek Avdalyan** National Polytechnic University of Armenia, Yerevan, Armenia

**Stephen Azeez** Department of Computing, Atlantic Technological University, Donegal, Ireland

**Tawakkal Baharuddin** Government Studies, Universitas Muhammadiyah Makassar, Makassar, Indonesia

**Nikesh Bajaj** Imperial College London, London, UK

**Rasul Balayev** Azerbaijan State University of Economics, Baku, Azerbaijan

**Cristina Barrena García** Fundación EHAS, Madrid, Spain

**Gaurav Basu** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India

**Anda Batraga** University of Latvia, Riga, Latvia

**Marciele Berger Bernardes** Escola de Direito, Universidade do Minho, Braga, Portugal

**Franziska Bocklisch** Chemnitz University of Technology, Chemnitz, Germany

**Chanoknart Boonwatthanakul** School of Educational Studies of Sukhothai, Thammathirat Open University, Pakkret, Thailand

**Carlos Borja-Galeas** Facultad de Administracion de Empresas, Universidad Tecnológica Indoamerica, Quito, Ecuador

**Juan Felipe Botero** Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

**Patrick Brandtner** University of Applied Sciences Upper Austria, Steyr, Austria

**Luis F. Brenes-García** Monterrey Institute of Technology (ITESM), Monterrey, NL, Mexico

**Maria A. Butakova** Smart Materials Research Institute, Southern Federal University, Rostov, Russia

**Paul M. Cabacungan** Ateneo de Manila University, Quezon City, Philippines

**Shuyao Cai** Arcadia University, Glenside, PA, USA

**Francisco J. Cantú-Ortiz** Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico;  
Tecnológico de Monterrey, Monterrey, Mexico;  
Monterrey Institute of Technology (ITESM), Monterrey, NL, Mexico

**Reymond P. Cao** Ateneo de Manila University, Quezon City, Philippines

**Rodrigo Carmona-Herrera** Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico

**John W. Castro** Universidad de Atacama, Copiapó, Chile

**Héctor G. Ceballos-Cancino** Monterrey Institute of Technology (ITESM), Monterrey, NL, Mexico;  
Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico;  
Tecnológico de Monterrey, Monterrey, Mexico

**Paul A. Chasi-Pesantez** Universidad Politécnica Salesiana, Cuenca, Ecuador

**Yun-Quan Cheng** Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia

**Ji-Jian Chin** Faculty of Science and Engineering, University of Plymouth, Plymouth, UK

**Luis Clavijo** Universidad Católica de Cuenca, Cuenca, Ecuador

**Francesco Colace** DIIN University of Salerno, Fisciano, Italy

**Dajana Conte** DIPMAT University of Salerno, Fisciano, Italy

**Lucas Cortizo** Escola de Direito, Universidade do Minho, Braga, Portugal

**Hoan Nguyen Dac** Viettel High Technology Corporation, Hanoi, Vietnam

**Edgardo S. Dajao** Graduate School of Engineering, Pamantasan Ng Lungsod Ng Maynila, City of Manila, Philippines

**Pavle Dakić** Faculty of Informatics and Computing, Singidunum University, Belgrade, Serbia;  
Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Bratislava, Slovakia

**Ernesto Damiani** Center for Cyber-Physical Systems (C2PS), Khalifa University, Abu Dhabi, UAE;  
Department of EECS, Khalifa University, Abu Dhabi, UAE

**Maria Pia D'Arienzo** DISUFF University of Salerno, Fisciano, Italy

**Matthew Dawson** National Center for Atmospheric Research (NCAR), Boulder, CO, USA

**Johnny Delgado** Universidad Católica de Cuenca, Cuenca, Ecuador

**Francisco Pacheco de Andrade** Escola de Direito, Universidade do Minho, Braga, Portugal

**L. Y. Dobkacz** Bauman Moscow State Technical University, Moscow, Russia

**Irina Dubovik** Intetics Inc, Naples, FL, USA

**Huynh Anh Duy** FPT University, Hanoi, Vietnam

**Tapia Edison** Universidad de Las Fuerzas Armadas ESPE, Sangolquí, Ecuador

**Coronel-Villavicencio Edison** Universidad Politecnica Salesiana GIHEA, Cuenca, Ecuador

**Mounir Elatrachi** LARMIG Laboratory, Hassan II University – FSJES Ain Sebaâ, Casablanca, Morocco

**Darío Espinoza-Saquicela** Institute of Sectional Regime Studies of Ecuador, Universidad del Azuay, Cuenca, Ecuador

**Oleksandra Farat** Lviv Polytechnic National University, Lviv, Ukraine

**Indah Fatmawati** Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Guerrero-Vasquez Fernando** Universidad Politécnica Salesiana, Grupo de Investigación en Hardware Embebido Aplicado GIHEA, Cuenca, Ecuador

**Pablo Flores-Siguenza** Department of Applied Chemistry and Systems of Production, Faculty of Chemical Sciences, Universidad de Cuenca, Cuenca, Ecuador

**Yuichi Fujino** Department of Media Architecture, Future University Hakodate, Hakodate, Japan

**Shuichi Fukuda** Keio University, Yokohama, Japan

**Ged Fuller** Elon University, Elon, NC, USA

**Thomas Gabel** Frankfurt University of Applied Sciences, Frankfurt, Germany

**Eduardo Cesar Galobardes** Universitat Autònoma de Barcelona, Bellaterra, Spain

**Carlos Pérez García-Pando** Barcelona SuperComputing Center, Barcelona, Spain

**Aitor Garrido Madrigal** Universidad Rey Juan Carlos, Madrid, Spain

**Christian Garzón** Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

**Obulreddigari Gayathri** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India

**Aries M. Gelera** Department of Computer Studies, Cavite State University-CCAT Campus, Cavite, Philippines

**Alexander Gertsy** State University of Infrastructure and Technologies, Kyiv, Ukraine

**Eicke Godehardt** Frankfurt University of Applied Sciences, Frankfurt, Germany

**Alyssa Joi A. Gonzales** Department of Computer Studies, Cavite State University-CCAT Campus, Cavite, Philippines

**Cristina F. Gonzales** Ateneo de Manila University, Quezon City, Philippines

**Ricardo Romero Gonzalez** Universidad Católica de Cuenca, Azogues, Ecuador

**Salvador P. Granada** Ateneo de Manila University, Quezon City, Philippines

**Luis F. Guerrero-Vásquez** Universidad Politécnica Salesiana, Cuenca, Ecuador

**Pranjal Gupta** BITS-Pilani Hyderabad Campus, Secunderabad, India

**S. Gurunarayanan** Birla Institute of Technology and Science, Pilani, Hyderabad Campus, Telangana, India

**Xu Han** National Defense University, Beijing, China

**Damien Hanyurwimfura** African Center of Excellence in Internet of Things (ACEIoT), University of Rwanda, Kigali, Rwanda

**Ari Happonen** LUT University, Lappeenranta, Finland

**Fumiko Harada** Research Organization of Science and Technology, Ritsumeikan University, Shiga, Japan

**Lyudmyla Harasym** Ukrainian National Forestry University, Lviv, Ukraine

**Takashi Hasuike** Waseda University, Shinjuku City, Japan

**Shagufta Henna** Department of Computing, Atlantic Technological University, Donegal, Ireland

**Victor Hernández-Manrique** Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico

**Nur Hidayah** Master of Hospital Administration, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Miriam Vanessa Hinojosa-Ramos** Instituto Superior Tecnológico Vicente Rocafuerte, Guayaquil, Ecuador

**Oleksandr Hladkyi** Kyiv National University of Trade and Economics, Kyiv, Ukraine

**Phan Duy Hung** FPT University, Hanoi, Vietnam

**Ab Razak Che Hussin** Azman Hashim International Business School (AHIBS), Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia

**Fuad Ibrahimov** Public Association “Center for Socio-Economic and Environmental Research”, Baku, Azerbaijan

**Suparto Iribaram** Islamic Studies, Institut Agama Islam Negeri Fattahul Muluk, Jayapura, Papua, Indonesia

**Nagito Ishida** Gunma University, Gunma, Japan

**Chirag Jain** BITS-Pilani Hyderabad Campus, Secunderabad, India

**Yanxia Jia** Arcadia University, Glenside, PA, USA

**Oriol Jorba** Barcelona SuperComputing Center, Barcelona, Spain

**Hasse Jubba** Department of Islamic Politics, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Liene Kaibe** University of Latvia, Riga, Latvia

**Manoj Subhash Kakade** Department of Electrical and Electronics Engineering, BITS Pilani, Pune, India

**Tatiana Kalganova** Brunel University London, London, UK

**Y. A. Kalgin** Bauman Moscow State Technical University, Moscow, Russia

**Sampreet Kalita** Indian Institute of Technology Guwahati, Guwahati, Assam, India

**Bill Karakostas** Inlecom Systems, Brussels, Belgium

**Hezerul Abdul Karim** Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia

**Anupama Karuppiah** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India;  
Birla Institute of Technology and Science, Pilani, K. K. Birla Goa Campus, Goa, India

**Ban Kar-Weng** Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia



**Tomoko Kashima** Kindai University, Hiroshima, Japan

**Shogo Katayama** Gunma University, Gunma, Japan

**Takis Katsoulakos** Inlecom Systems, Brussels, Belgium

**Katrina Kellerte** University of Latvia, Riga, Latvia

**Jayden Khakurel** University of Turku, Turku, Finland

**Mousa Khubrani** Department of Computer Science, College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia

**Song-Kyoo Kim** Faculty of Applied Sciences, Macao Polytechnic University, Macau, Macao

**Sergey Kizyan** Intetics Inc, Naples, FL, USA

**Anja Klichowicz** Chemnitz University of Technology, Chemnitz, Germany

**Haruo Kobayashi** Gunma University, Gunma, Japan

**Bohdan Koman** Ivan Franko National University of Lviv, Lviv, Ukraine

**Ade Komariah** Master in Nursing Program, Universitas Muhammadiyah Yogyakarta, Tamantirto Kasihan Bantul, Indonesia

**Boris Kontsevoi** Intetics Inc, Naples, FL, USA

**Deepali B. Koppad** Department of Electronics and Communication, Ramaiah Institute of Technology, Bengaluru, India

**Mykola Korablyov** Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

**Elizabeth Koumpan** IBM Consulting, Markham, ON, Canada

**Volodymyr Kulivnuk** Vinnytsia National Pirogov Memorial Medical University, Vinnytsia, Ukraine

**Deepa Kumari** BITS-Pilani Hyderabad Campus, Secunderabad, India

**Anna Kuwana** Gunma University, Gunma, Japan

**Ivan Kuzmin** Vinnytsia National Pirogov Memorial Medical University, Vinnytsia, Ukraine

**De La Cruz Aida** Universidad de Las Fuerzas Armadas ESPE, Sangolquí, Ecuador

**Luisito Lolong Lacatan** Pamantasan ng Cabuyao, Cabuyao, Philippines

**Logan LaMont** Elon University, Elon, NC, USA

**Roberta Valeria Latorre** Department of Medicine, University of Verona, Verona, Italy

**Tu Le Minh** Viettel High Technology Corporation, Hanoi, Vietnam

**Van-Hung Le** Tan Trao University, Tuyen Quang, Vietnam

**Paola Lecca** Faculty of Computer Science, Smart Data Factory Laboratory, Free University of Bozen-Bolzano, Bolzano, Italy;

Member of National Group for Mathematical Analysis, Probability and their Applications, Francesco Severi's National Institute of High Mathematics, Rome, Italy

**Patrick Lennon** Atlantic Technological University, Letterkenny, Ireland

**Ruth G. Lennon** Atlantic Technological University, Letterkenny, Ireland;

Lero—The Irish Software Engineering Research Centre, University of Limerick, Limerick, Ireland

**Marcelo Leon** Universidad ECOTEC, Samborondon, Ecuador

**Paulina Leon** University of Malaga, Malaga, Spain

**Cun Li** School of Design, Jiangnan University, Wu Xi Shi, Jiang Su Sheng, China

**Qiao Liang** School of Design, Jiangnan University, Wu Xi Shi, Jiang Su Sheng, China

**Nathaniel Joseph C. Libatique** Ateneo de Manila University, Quezon City, Philippines

**Juan P. Licona-Luque** Monterrey Institute of Technology (ITESM), Monterrey, NL, Mexico

**Lucrecia Llerena** Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador

**Nai-Wei Lo** Department of Info. Mgt, NTUST, Taipei, Taiwan

**Marcelo D. Lojano-Angamarca** Universidad Politécnica Salesiana, Cuenca, Ecuador

**Giulia Lombardi** Department of Mathematics, University of Trento, Trento, Italy

**Milen Loukantchevsky** University of Ruse, Ruse, Bulgaria

**Serpa-Andrade Luis** Universidad Politécnica Salesiana, Grupo de Investigación en Hardware Embebido Aplicado GIHEA, Cuenca, Ecuador

**Sergey Lutskyy** Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

**Mapula Elisa Maeko** Academy of Computer Science and Software Engineering, University of Johannesburg, CNR University Road and Kingsway Avenue, Auckland Park, Gauteng, South Africa

**George C. Makris** Department of Digital Systems, University of Thessaly, Larissa, Greece

**José Maldonado-Quezada** Universidad Nacional de Loja, Loja, Ecuador

**Lucia Mandová** Faculty of Informatics and Information Technology, Institute of Computer Engineering and Applied Informatics, Slovak University of Technology, Bratislava, Slovakia

**Amit Mankodi** Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India

**Sarina Mansor** Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia

**Jean P. Mata-Quevedo** Universidad Católica de Cuenca, Azogues, Ecuador

**Masafumi Matsuhara** Iwate Prefectural University, Iwate, Japan

**Shimpei Matsumoto** Hiroshima Institute of Technology, Hiroshima, Japan

**Shuichi Matsumoto** Japan Cable Laboratories, Tokyo, Japan

**Anargyros Mavrakos** Inlecom Systems, Brussels, Belgium

**Jorge A. Mendez-Vargas** Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico

**Rafael B. Méndez-Vásquez** Universidad Politécnica Salesiana, Cuenca, Ecuador

**Gabriela G. Mendoza-Leal** Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico

**Neil Angelo M. Mercado** Ateneo de Manila University, Quezon City, Philippines

**Marcia Mkansi** University of South Africa, Pretoria, Gauteng, Republic of South Africa

**Tina Morgenstern** Chemnitz University of Technology, Chemnitz, Germany

**Marlon Moscoso-Martínez** Faculty of Sciences, Escuela Superior Politécnica de Chimborazo (ESPOCH), Riobamba, Ecuador

**Aya Essam Mostafa** School of Information Technology and Computer Science, NU, Giza, Egypt

**Mohamed Essam Mostafa** School of Engineering and Applied Science, NU, Giza, Egypt

**Umar Mujahid** Georgia Gwinnett College, Lawrenceville, GA, USA

**Iryna Mysiuk** Ivan Franko National University of Lviv, Lviv, Ukraine

**Roman Mysiuk** Ivan Franko National University of Lviv, Lviv, Ukraine

**Karlo Jose E. Nabablit** Department of Computer Studies, Cavite State University-CCAT Campus, Cavite, Philippines

**Phisit Nadprasert** Office of Educational Technology of Sukhothai, Thammathirat Open University, Pakkret, Thailand

**N. Nethra** National Seed Project, University of Agricultural Sciences, Bengaluru, India

**Ashay Netke** BITS-Pilani Hyderabad Campus, Secunderabad, India

**Jimmy Nsenga** African Center of Excellence in Internet of Things (ACEIoT), University of Rwanda, Kigali, Rwanda

**Marvin Ogore** African Center of Excellence in Internet of Things (ACEIoT), University of Rwanda, Kigali, Rwanda

**Martin Opatovský** Faculty of Informatics and Information Technology, Institute of Computer Engineering and Applied Informatics, Slovak University of Technology, Bratislava, Slovakia

**Carlos M. Oppus** Ateneo de Manila University, Quezon City, Philippines

**Jorge O. Ordoñez-Ordoñez** Universidad Politécnica Salesiana, Cuenca, Ecuador

**Dijana Oreski** Faculty of Organization and Informatics, University of Zagreb, Varazdin, Croatia

**Juan Ortega** Universidad Católica de Cuenca, Cuenca, Ecuador

**Ana Osorio** Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador

**Samira Oukarfi** LARMIG Laboratory, Hassan II University – FSJES Ain Sebaâ, Casablanca, Morocco

**Moch Edward Trias Pahlevi** Komite Independen Sadar Pemilu (KISP), Yogyakarta, Indonesia

**Subhrakanta Panda** BITS-Pilani Hyderabad Campus, Secunderabad, India

**Rohit Pandey** Hughes Systique Corporation, Gurugram, India

**Pratheep Kumar Paranthaman** Elon University, Elon, NC, USA

**Mata-Quevedo Paul** Universidad Católica de Cuenca, Azogues, Ecuador

**Armen Petrosyan** Synopsys Armenia CJSC, Yerevan, Armenia

**Nutteerat Pheeraphan** Srinakharinwirot University, Bangkok, Thailand

**Emma E. Porio** Ateneo de Manila University, Quezon City, Philippines

**Thomas Poteat** Elon University, Elon, NC, USA

**Dimitrios Poulakis** Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece

**Ignacio Prieto-Egido** Universidad Rey Juan Carlos, Madrid, Spain

**Titin Purwaningsih** Doctoral Program of Government Affairs and Administration, Postgraduate Faculty, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Likkhasit Putkhiao** School of Educational Studies of Sukhothai, Thammathirat Open University, Pakkret, Thailand

**Zuly Qodir** Department of Islamic Politics, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Diep Pham Quang** Viettel High Technology Corporation, Hanoi, Vietnam

**Sebastian Quevedo** Universidad Católica de Cuenca, Cuenca, Ecuador;  
Electrical and Computer Science Engineering Department, Escuela Superior Politécnica del Litoral—ESPOL University, Guayaquil, Ecuador

**Aleksandr Raikov** National Supercomputer Centre in Jinan, Shandong, China;  
Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia;  
MIREA – Russian Technological University, Moscow, Russia

**Tshililo Ramovha** University of South Africa, Pretoria, Gauteng, Republic of South Africa

**V. Ranjith** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India

**Angela Re** Department of Applied Science and Technology, Politecnico di Torino, Turin, Italy

**Stefan Rizanov** Faculty of Electronic Engineering and Technologies, Technical University of Sofia, Sofia, Bulgaria

**Garcia-Velez Roberto** Universidad Politécnica Salesiana, Grupo de Investigación en Hardware Embebido Aplicado GIHEA, Cuenca, Ecuador

**Maria Theresa Joy G. Rocamora** Ateneo de Manila University, Quezon City, Philippines

**Erna Rochmawati** Universitas Muhammadiyah Yogyakarta, Tamantirto Kasihan Bantul, Indonesia

**Ramon L. Rodriguez** National University, Manila, Philippines

**Nancy Rodríguez** Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador

**María Fernanda Romo-Fuentes** Tecnológico de Monterrey, Estado de México, Mexico

**Christian Guzman Ruiz** Barcelona SuperComputing Center, Barcelona, Spain

**Ulviyya Rzayeva** Azerbaijan State University of Economics, Baku, Azerbaijan

**Maria E. Sabani** Department of Digital Systems, University of Thessaly, Larissa, Greece

**Khim Cathleen M. Saddi** Ateneo de Naga, Camarines Sur, Philippines

**S. A. Sakulin** Bauman Moscow State Technical University, Moscow, Russia

**Jelena Salkovska** University of Latvia, Riga, Latvia

**Samsuryadi** Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

**Dominic Sanderson** Brunel University London, London, UK

**Gonzalez-Gonzalez Sandro** Independent Researcher, Azogues, Ecuador

**Domenico Santaniello** DIIN University of Salerno, Fisciano, Italy

**Paul Ryan A. Santiago** Ateneo de Manila University, Quezon City, Philippines

**Duangbhorn Sapphayalak** Office of Educational Technology of Sukhothai, Tham-mathirat Open University, Pakkret, Thailand

**Alvin A. Sario** University of Santo Tomas, Legazpi, Philippines

**Amarendra K. Sarma** Indian Institute of Technology Guwahati, Guwahati, Assam, India

**Ikuma Sato** Department of Media Architecture, Future University Hakodate, Hakodate, Japan

**Ilias K. Savvas** Department of Digital Systems, University of Thessaly, Larissa, Greece

**Aman Saxena** BITS-Pilani Hyderabad Campus, Secunderabad, India

**Luis Serpa-Andrade** Universidad Politécnica Salesiana GIHEA, Cuenca, Ecuador

**Kim Serradell** Barcelona SuperComputing Center, Barcelona, Spain

**Elcid A. Serrano** Mapua University, Manila, Philippines

**Ebtesam Shadadi** Department of Computer Science, College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia

**Muzammil Shahbaz** Thales UK Ltd Cheadle Heath, Stockport, UK

**Raafat Shalaby** SESC Center, School of Engineering and Applied Science, NU, Giza, Egypt;  
Faculty of Electronic Engineering, Menofia University, Menouf, Egypt

**Pawan Sharma** Birla Institute of Technology and Science, Pilani, Pilani Campus, Pilani, India

**Raghav Sharma** Hughes Systique Corporation, Gurugram, India

**Hiromitsu Shimakawa** Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

**E. S. Shoukralla** Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt

**Tetiana Shparaga** Taras Shevchenko Kyiv National University, Kyiv, Ukraine

**Lorena Siguenza-Guzman** Department of Computer Sciences, Faculty of Engineering, Universidad de Cuenca, Cuenca, Ecuador;  
Research Centre Accountancy, Faculty of Economics and Business, KU Leuven, Leuven, Belgium

**Marvin G. Sison** Department of Computer Studies, Cavite State University-CCAT Campus, Cavite, Philippines

**C. S. Sonali** Department of Electronics and Communication, Ramaiah Institute of Technology, Bengaluru, India

**Carlos Soria** Universidad Católica de Cuenca, Cuenca, Ecuador

**Claudio Sorio** Department of Medicine, University of Verona, Verona, Italy

**Zuzana Špitálová** Faculty of Informatics and Information Technology, Institute of Computer Engineering and Applied Informatics, Slovak University of Technology, Bratislava, Slovakia

**Mudigonda Sreevastav** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India

**Vjeran Strahonja** Faculty of Organization and Informatics, University of Zagreb, Varazdin, Croatia

**Jane Kristine G. Suarez** Bulacan State University, Malolos, Philippines

**Supanita Sudsaward** Office of Educational Technology of Sukhothai, Thammathirat Open University, Pakkret, Thailand

**Amornphong Suksen** Srinakharinwirot University, Bangkok, Thailand

**K. V. Suma** Department of Electronics and Communication, Ramaiah Institute of Technology, Bengaluru, India

**Aisha Sumaili** Department of Information Technology and Security, College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia

**Hung Nguyen Tai** School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam

**Gregory L. Tangonan** Ateneo de Manila University, Quezon City, Philippines

**Chen Tianxiao** State Grid, Xiangtan Power Supply Company, Xiangtan, China

**Tetiana Tkachenko** Kyiv National University of Trade and Economics, Kyiv, Ukraine

**Dimitar Todorov** Faculty of Electronic Engineering and Technologies, Technical University of Sofia, Sofia, Bulgaria

**Vladimir Todorović** Faculty of Business Studies and Law, MB University, Belgrade, Serbia



**Viviana Tomala** Universidad Cesar Vallejo, Trujillo, Perú

**Anna W. Topol** IBM Research—Watson, New York, USA

**Dhvani Toprani** Elon University, Elon, NC, USA

**Bryan James V. Torres** Department of Computer Studies, Cavite State University-CCAT Campus, Cavite, Philippines

**Binh Tran** Georgia Gwinnett College, Lawrenceville, GA, USA

**Alfredo Troiano** NetCom Group, Napoli, Italy

**Theodosis Tsaousis** Inlecom Systems, Brussels, Belgium

**Yuriy Tyrkalo** Lviv Polytechnic National University, Lviv, Ukraine

**Ritha M. Umutoni** African Center of Excellence in Internet of Things (ACEIoT), University of Rwanda, Kigali, Rwanda

**Usman Ahmad Usmani** Universiti Teknologi Petronas, Seri Iskandar, Perak, Malaysia

**Carmine Valentino** DIIN University of Salerno, Fisciano, Italy

**Dustin Van Der Haar** Academy of Computer Science and Software Engineering, University of Johannesburg, CNR University Road and Kingsway Avenue, Auckland Park, Gauteng, South Africa

**Diego Verdugo-Ormaza** Universidad Católica de Cuenca, Azogues, Ecuador

**Jaime Vergara** Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

**Alexander Julian Vieth** Frankfurt University of Applied Sciences, Frankfurt, Germany

**Bui Trong Vinh** Hanoi Procuratorate University, Hanoi, Vietnam

**Sista Kasi Vishwanath** Department of Electrical and Electronics Engineering, BITS Pilani, Zuarinagar, Goa, India

**Valentino Vranić** Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Bratislava, Slovakia

**Junzo Watada** Waseda University, Heidelberg, Japan

**Ari Wedhasmara** Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

**Zhang Wei** State Grid, Xiangtan Power Supply Company, Xiangtan, China

**Bambang Eka Cahya Widodo** Undergraduate Program of Government Affairs and Administration, Faculty of Social and Political Sciences, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

**Qian Xu** Elon University, Elon, NC, USA

**Peter Yakimov** Faculty of Electronic Engineering and Technologies, Technical University of Sofia, Sofia, Bulgaria

**Ayaka Yamanaka** Graduate School of System Information Science, Future University Hakodate, Hakodate, Japan

**Huang Yao** State Grid, Xiangtan Power Supply Company, Xiangtan, China

**Chan Yeob Yeun** Center for Cyber-Physical Systems (C2PS), Khalifa University, Abu Dhabi, UAE;  
Department of EECS, Khalifa University, Abu Dhabi, UAE

**Paul D. Yoo** CSIS, Birkbeck College, University of London, London, UK;  
Cranfield School of Defence and Sec, Defence Academy of the UK, Shrivenham, UK

**Tomoya Yuasa** Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

**Volodymyr Yuzevych** Karpenko Physico-Mechanical Institute of the NAS of Ukraine, Lviv, Ukraine

**Stathis Zavvos** VLTN, Antwerp, Belgium

**Ammar Zeini** Atlantic Technological University, Letterkenny, Ireland;  
Lero—The Irish Software Engineering Research Centre, University of Limerick, Limerick, Ireland

**Miaoting Zeng** National Defense University, Beijing, China

**Lei Zhang** National Defense University, Beijing, China

**Sunyi Zhang** Arcadia University, Glenside, PA, USA

**Lin Zhu** Arcadia University, Glenside, PA, USA

**Robert Zimmermann** University of Applied Sciences Upper Austria, Steyr, Austria

**Zhu Zuoping** State Grid, Xiangtan Power Supply Company, Xiangtan, China

# Development of a Method for Reducing the Impact of Metal Interconnection Parameters on the Speed of VLSI



Narek Avdalyan  and Armen Petrosyan

**Abstract** The development of modern very large-scale integration circuits is aimed increasing the degree of integration, which will provide higher speeds. For that, it is necessary to reduce the size of all the logical elements and increase the number of metal layers, which in its turn will lead to an increase in the parasitic capacitance and inductance of the transmission lines (interconnections). As a result, we will have a signal delay in the transmission lines, which will affect the speed of the system.

**Keywords** VLSI · High performance · High-speed · Interconnections · Intersection · Overlap metal layers

## 1 Introduction

In modern integrated circuits (ICs), the number of metal layers reaches 15 and more during the interconnect tracing. It is more important the correct choose of metal, which is conditioned by the appropriate level of metal.

Quite a few parameters must be taken into account (mutual overlap, time delay due to the length of the metal). The placement of elements with high density interconnections in the IC can become quite a serious problem.

It is necessary to calculate the capacitances and resistances of metals for different technological norms depending on the different densities of the project.

The result of this research can be very useful for design planning, which will allow the designer to set the right rules to avoid deterioration of power and time parameters.

---

N. Avdalyan (✉)

National Polytechnic University of Armenia, Teryan 105, 009 Yerevan, Armenia  
e-mail: [narekius@gmail.com](mailto:narekius@gmail.com)

A. Petrosyan

Synopsys Armenia CJSC, Arshakunyats 41, 0026 Yerevan, Armenia  
e-mail: [armenp@synopsys.com](mailto:armenp@synopsys.com)

Figure 1 shows an example of a modern IC where metal layers and interconnects are shown. Metal interconnections are divided into two main groups: local and global. Local interconnections are due to transistors (gates), and global connections are higher metal connections that are thicker metals, connecting elements at the macro level.

One of the modern technological problems is the increasing of the integration degree, which allows to distribute as many logical elements as possible in a unit area. As a result, the number of elements per unit area increases and leads to an increase in speed. But it is hampered by an increase in the number of metal interconnects. This leads to increased delays due to an increase in the coupling capacity of interconnect metal lines. Figure 2 shows the main components of metal capacities due to metal line overlap.

As can be seen from Fig. 2, the main capacitance that can affect the signal is the  $C_2$ . It describes the capacitance due to two parallel metals ( $M1$  metal layer) at a certain distance.

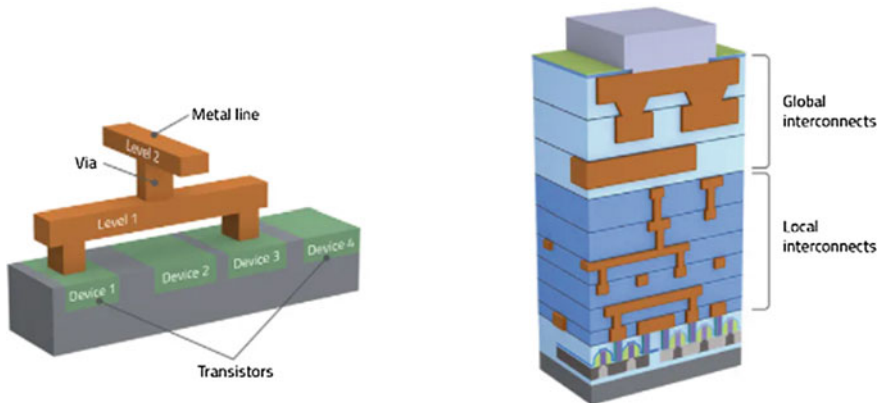


Fig. 1 Metal layers in IC

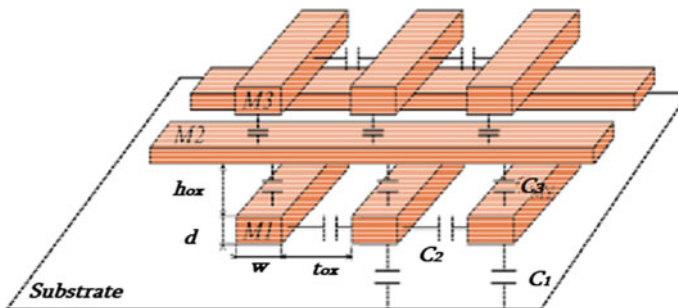


Fig. 2 Capacitance due to metal line overlap

$C_1$  describes the capacitance between the first metal layer and substrate.  $C_3$  describes the capacitance due to overlap of different metal layers and  $h_{ox}$  characterizes the distance between two metal layers. The capacity of metal layers is calculated by the following formula [1–3]:

$$C = \frac{Q}{\iint \frac{E dS}{\epsilon_0 \epsilon_r}} \quad (1)$$

where  $Q$  is the charge,  $E$  is the vector of the electric field potential V/m,  $\overline{dS}$  is the integral of the outer surface of the metal layer, which depends on the thickness and width of the metal layer,  $m^2$ .

To calculate the capacities, it is necessary to have the width, the length of the metal layers and the distance between those metal layers and the area of their overlap.

$$C_{\Sigma} = C_1 + 2C_2 + C_3 \quad (2)$$

The area of the overlap is conditioned by the design of the metal layers and how the given metal layer passes in comparison with other metal layers.

From the expression (1), it follows that long parallel connections in the same metal layer should be avoided, which is especially critical at short distances. In these cases, it is necessary to change the metal layer to avoid a large increase in capacity. The change of the metal layer, which can be a higher level of metal, leads to another type of problem, which is the increase of resistance. It is due to the thickness of the high metal layers, as the high metal layers are thicker than the low metal layers. The resistance of metal layers is calculated by the following formula:

$$R = \frac{\rho_v}{d} \frac{l}{w} = \rho_s \frac{l}{w} \quad (3)$$

where  $\rho_v$  is the resistance of the material, Ohm/m,  $\rho_s$  is the resistance per unit area Ohm/m.

The signal delay in a metal conductor is determined by the following formula:

$$\tau_{RC} = RC \quad (4)$$

where  $R$  is the resistance of the metal interconnect and  $C$  is the capacity.

The inductance of metal layers is calculated by the following formula [4]:

$$L = \frac{\Phi_B}{I} = \frac{\iint \overline{B} d\overline{S}}{I} \quad (5)$$

where  $\Phi_B$  is the magnetic field flux, Wb,  $I$ —current, A, and  $\overline{B}$ —the magnetic field vector, T.

If we do not have active resistance in the system, then the delay in the interconnection lines is calculated by the following formula:

$$\tau_{LC} = \sqrt{LC} \quad (6)$$

where  $L$  is the inductance,  $H$ ,  $C$ —the capacitance of metal sheets,  $F$ .

In the general case, the delay in the system will be expressed as follows:

$$\tau_{RLC} = \tau_{LC} + \tau_{RC} \quad (7)$$

In the general case, the delay in the system will be expressed as follows:

$$\tau_{RLC} = 1.047\tau_{LC} + 1.4\tau_{RC} \quad (8)$$

In order to obtain values closer to reality, expression (7) requires correction factors [5–10]. In modern VLSI, the length of metal lines reaches microns. Therefore, the inductance component of the metal line can be neglected, as it is very small.

## 2 Problem Setting and Justification of the Methodology

From the above, it becomes obvious that we are dealing with a reciprocal overlapping problem, for the solution of which it is necessary to offer a compromise solution. The compromise will be conditioned by the “change of the level of the metal layer” of two parallel metal layers.

Thus, it becomes obvious that in not all cases it is expedient to avoid from parallel metal interconnects for small distances, which increases the capacity, which in its turn leads to an increase in delay times. Because even in the case of a high metal layer, the resistance increases, which can also lead to an increase in the delay times according to expression (4)

The aim of the work is to fix by modeling the cases in which it is expedient to change the metal level, i.e., it is not necessary to take into account the components of resistance and capacitance, which can affect the speed. A method has been developed to solve the problem, which will allow finding the best compromise solution to the above problem. The sequence of algorithmic steps of the method can be presented as follows:

1. It is necessary to model the metal layers of the system.
2. Assess the resistance and capacitance values of a given system, which are represented by arrays  $(R_1, R_2, \dots, R_n$  and  $C_1, C_2, \dots, C_n)$ :
3. Classify resistance and capacitance values in descending order.
4. Find the mean mathematical expectation for the values of resistance and capacitance.
5. If the average mathematical expectation of the capacitance is greater than the corresponding value of the resistance, it is necessary to change the metal layers and start comparing the average mathematical expectation again.

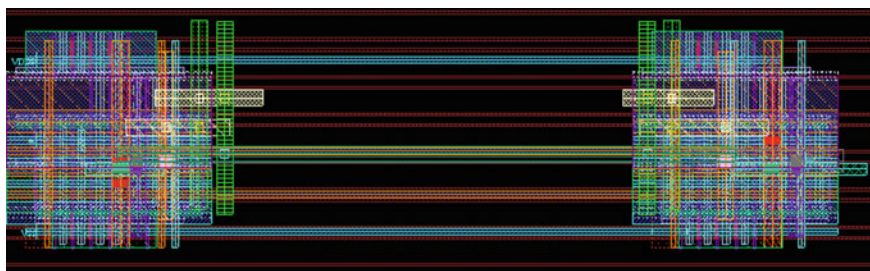
6. In case of reverse fulfillment of the condition in point 5, one should try to reduce the high metal layers and compare again.
7. 5–6 points should be done as many times as possible to get the most centralized values for both resistance and capacitance.

This algorithm allows to obtain resistance and capacitance values for metal layers with the smallest possible centralized parameters and will provide the reduction of the resistance and capacitance values due to the increase of metal layers in the system.

### 3 Simulation Results

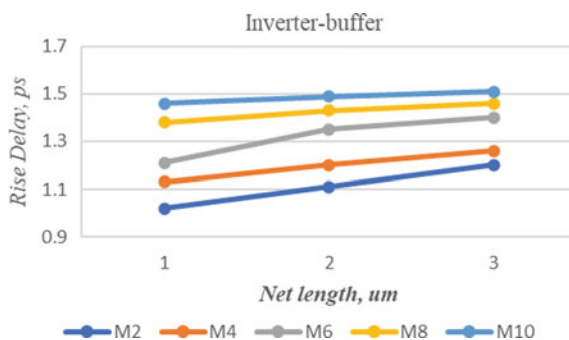
The research was performed using 10 metal layers for inverter-buffer, inverter-NAND, inverter-NOR nodes. The distance between the nodes was chosen to be 1, 2 and 3  $\mu\text{m}$ .

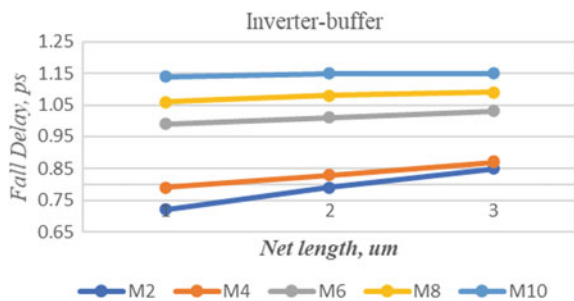
The modeling was performed according to 14 nm technological norms in case of 0.9 V supply voltage, 25 °C temperature, standard process, 50 ps fronts of the input signal, with Custom compiler and HSPICE software package. The simulation results and layout views are presented in Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13 [11, 12].



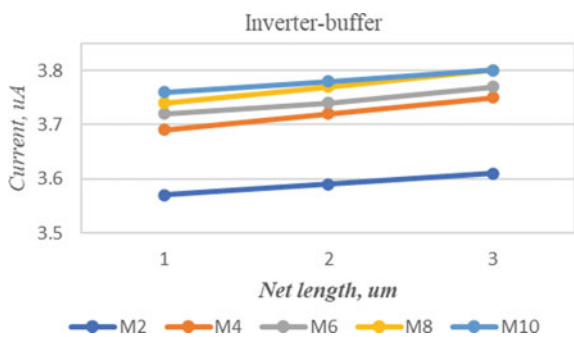
**Fig. 3** Layout view of inverter connect to buffer

**Fig. 4** Distance dependence of the rise front time of the inverter-buffer node

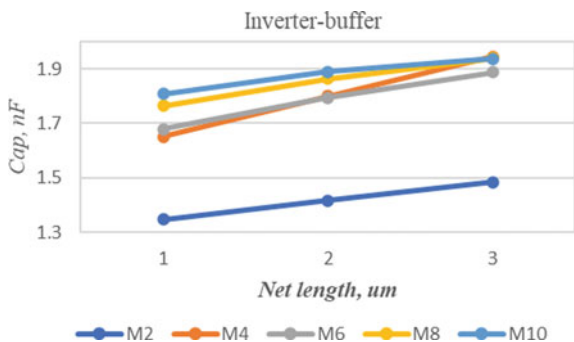




**Fig. 5** Dependence of the fall front time of the inverter-buffer unit on the distance

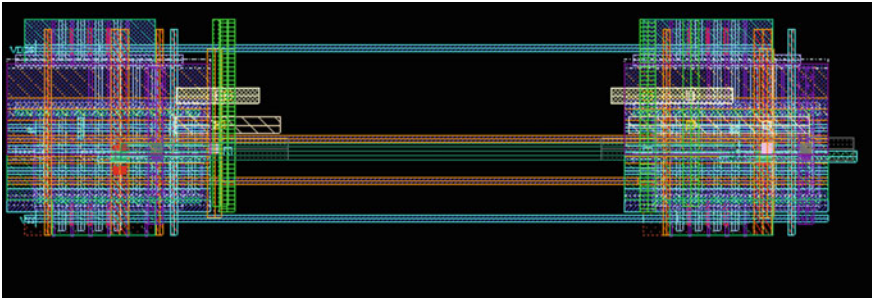


**Fig. 6** Dependence of current consumption of inverter-buffer unit on distance



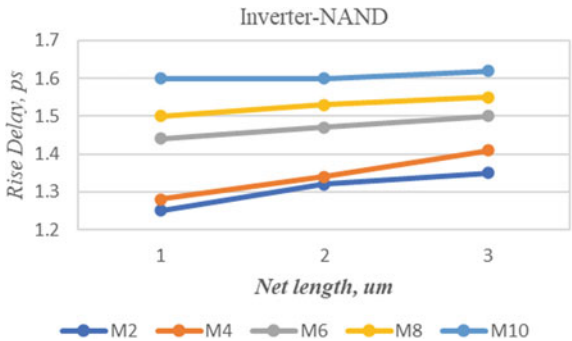
**Fig. 7** Dependence of inverter-buffer line capacity on distance



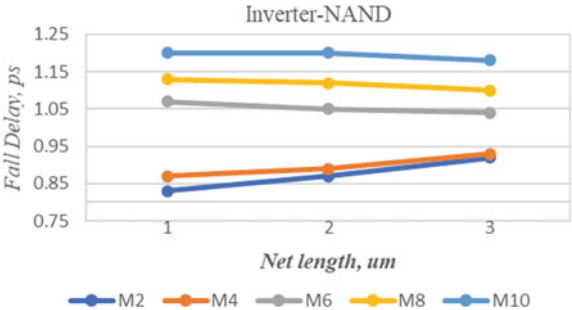


**Fig. 8** Layout view of inverter connects to NAND

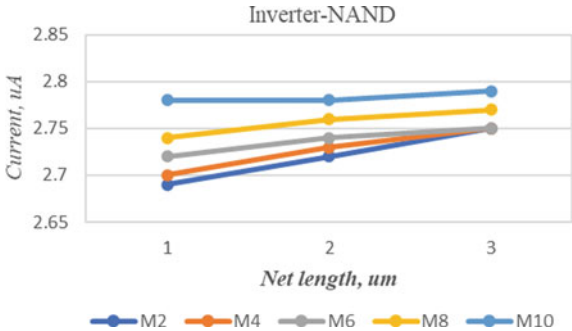
**Fig. 9** Inverter-NAND node rise front time dependence on distance

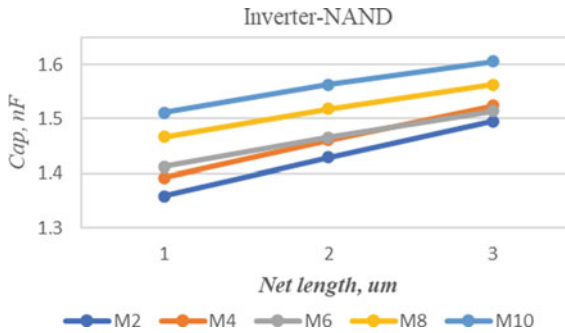


**Fig. 10** Inverter-NAND node fall front time dependence on distance

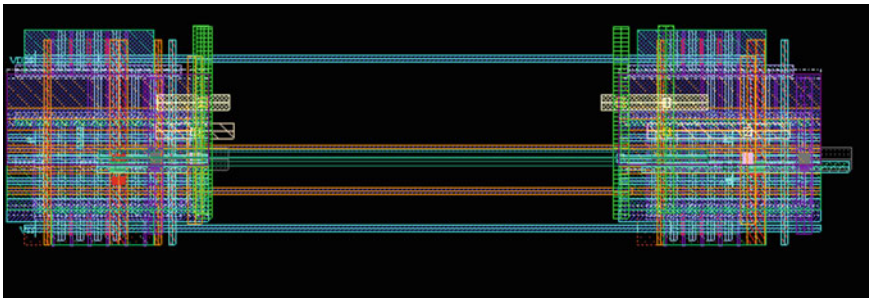


**Fig. 11** Inverter-NAND node consumption current dependence on distance





**Fig. 12** Inverter-NAND node line capacity dependence on distance



**Fig. 13** Layout view of inverter connects to NOR

## 4 Conclusion

1. The developed method allows to evaluate the capacitances and resistances of metal layers of any design.
2. The developed method allows to reduce the capacitance and resistance values of metal layers.
3. The work considers simple examples of logic elements; in which case the applied algorithm ensures a reduction of delay times by up to 8–10%. It is obvious that in large systems it can provide better results given the large number of interconnection metals.
4. The developed method does not carry out the reduction in a fully automated way and the user needs certain design knowledge to apply it correctly.

## References

1. Kumar A (2012) Study the performance analysis of carbon nanotube as a VLSI interconnect. Patiala: Thapar University, p 101. <http://hdl.handle.net/10266/1898>
2. Jaing Y (2006) Modeling and optimization of VLSI interconnects. USA: Nevada Las Vegas. Department Elect and Computer Engineering, University of Nevada Las Vegas, p 100
3. The international technology roadmap for semiconductors. URL: <http://www.itrs2.net> (20.07.2018)
4. Ismail I, Friedman EG (2002) Effects of inductance on the propagation delay and repeater insertion in VLSI circuits: a summary. *Circ Syst Soc Outstand Young Author Award* 8(2):195–206
5. Abinash R, Noha M (2007) Effects of coupling capacitance and inductance on delay uncertainty and clock skew. USA: Illinois Chicago. Department of Elect. and Computer Engineering, University of Illinois at Chicago, pp 184–187
6. Ramadass U (2013) A novel interconnect structure for Elmore delay model with resistance – capacitance – conductance scheme/Ramadass U, Krishnappriya, J. Ponnian. *Am J Appl Sci* 10(8):881–892. <https://doi.org/10.3844/ajassp.2013.881.892>
7. Yang XD, Cheng CK, Ku WH et al (2002) Hurwitz stable reduced order modeling for RLC interconnect trees. *IEEE J Anal Integr Circ Signal Process* 31:222–228
8. Ismail YI (Dec. 2002) On-chip inductance cons and pros. *IEEE Trans VLSI Syst* 685–694
9. Kahng AB, Muddu S, Vidhani D (Jan 2004) Noise and delay uncertainty studies for coupled RC interconnects. In: *Proceedings IEEE international conference on VLSI design*, pp 431–436
10. Chen YH et al. (Jun. 2020) Ultra high density SoIC with sub-micron bond pitch. In: *Proceedings IEEE 70th electronic components technology conference (ECTC)*, pp 576–581
11. Yan J-T, Tseng Y-J, Yen C-H (Dec. 2014) Efficient micro-bump assignment for RDL routing in 3DICs. In: *Proceedings 21st IEEE international conference on electronics, circuits and systems (ICECS)*, pp 195–198
12. Panth S, Samadi K, Du Y, Lim SK (Oct. 2017) Shrunk-2-D: a physical design methodology to build commercial-quality monolithic 3-D ICs. *IEEE Trans Comput Aided Design Integr Circuits Syst* 36(10):1716–1724

# Deep Learning-Based Arrhythmia Detection Using RR-Interval Framed Electrocardiograms



Song-Kyoo Kim, Chan Yeob Yeun, Paul D. Yoo, Nai-Wei Lo,  
and Ernesto Damiani

**Abstract** Deep learning applied to electrocardiogram (ECG) data can be used to achieve personal authentication in biometric security applications, but it has not been widely used to diagnose cardiovascular disorders. We developed a deep learning model for the detection of arrhythmia in which time-sliced ECG data representing the distance between successive R-peaks are used as the input for a convolutional neural network (CNN). The main objective is developing the compact deep learning-based detect system which minimally uses the dataset but delivers the confident accuracy rate of the arrhythmia detection. This compact system can be implemented in wearable devices or real-time monitoring equipment because the feature extraction step is not required for complex ECG waveforms, only the R-peak data is needed. The 10 hidden layers of the CNN detect arrhythmias using a novel RR-interval framing (RRIF) approach. Two testing processes were implemented, the first during the training and validation of the CNN algorithm and the second using different datasets for testing under realistic conditions. The results of both tests indicated that the Compact Arrhythmia Detection System (CADS) matched the performance of conventional systems for the detection of arrhythmia in two consecutive test runs.

**Keywords** Deep learning · Convolutional neural network · Arrhythmia · Electrocardiogram · Time-sliced data · MATLAB

---

S.-K. Kim (✉)

Faculty of Applied Sciences, Macao Polytechnic University, Macau, Macao

e-mail: [amang@mpu.edu.mo](mailto:amang@mpu.edu.mo)

C. Y. Yeun · E. Damiani

Center for Cyber-Physical Systems (C2PS), Khalifa University, Abu Dhabi, UAE

Department of EECS, Khalifa University, Abu Dhabi, UAE

P. D. Yoo

CSIS, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Cranfield School of Defence and Sec, Defence Academy of the UK, Shrivenham, UK

N.-W. Lo

Department of Info. Mgt, NTUST, Taipei, Taiwan

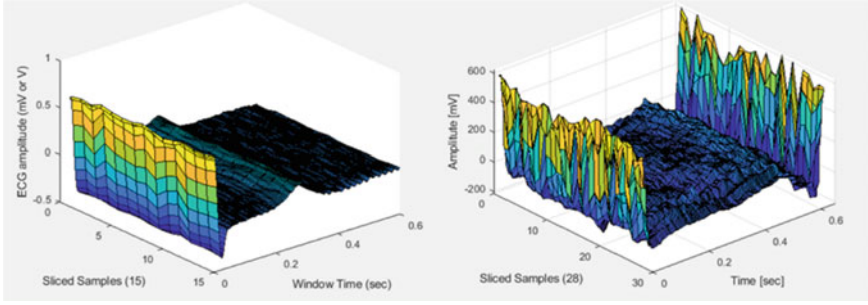
# 1 Introduction

The use of time-sliced electrocardiogram (ECG) data was originally developed for biometric applications [1–3]. ECGs are usually encountered in a medical setting [4, 5] but are also useful for security because they offer a novel way to identify individuals [6–11]. ECGs monitor the electrical activity of the heart, and can be used to capture analog signal profiles that allow personal identification and authentication when converted into digital data [12]. The components of ECG signals are often used by cardiologists to diagnose cardiovascular problems [13–15]. However, the manual analysis of ECG signals is challenging because it is difficult to categorize the different waveforms and signal morphologies accurately [16]. Automated ECG signal analysis has been introduced to avoid human errors and reduce costs, recently including the use of artificial intelligence [16–18].

Machine learning (ML) is a subset of artificial intelligence in which computer systems learn to perform a specific task without explicit instructions [1]. ML algorithms build a mathematical model of training data to make predictions or decisions without direct programming [19]. ML has been applied in many areas in recent decades, including the analysis of ECG data [19–26]. ML techniques applied to ECG analysis include multi-layered perceptron (MLP) and various forms of deep learning, such as artificial neural networks (ANNs) or convolutional neural networks (CNNs), either for biometric security applications [7–11] or the diagnosis of cardiovascular diseases such as arrhythmia [26–31]. The performance of deep learning models often depends on the research objectives, and the design of effective deep learning models tailored for specific research topics is a research topic in itself [32–34]. A CNN is chosen because it is widely applied for various ML applications but it is usually required heavy amount of data and high computing power which is not generally suitable for low performance devices (IoT devices). But a CNN could be applied even for low performance devices if training dataset are compact but value enough for developing accurate machine learning systems. Time-sliced ECG data is highly flexible because it can be mixed with other training inputs and is compatible with various ML methods without the need to categorize all the featured waveforms: only the R-peaks are required. The sliced data are used as the input parameters to train neural networks.

Two examples are shown in Fig. 1, one based on a fixed-time window and the other based on RR-interval framing (RRIF). Unlike conventional time slicing method [1, 2], the proposed RRIF method uses ECG data based on the interval between heartbeats instead of the slicing window time [3]. In this article, we use RRIF as a new approach to provide time-sliced ECG data for ML systems, allowing the detection of arrhythmia. We have designed a Compact Arrhythmia Detection System (CADS) based on a CNN with binary outputs, which requires only short-range ECG data to achieve detection, allowing it to be used with conventional wearable devices or real-time monitoring equipment.

This article consists of five sections. Section 2 describes the RRIF concept, the CADS deep learning model based on a 10-layer CNN, and the training and testing



**Fig. 1** Fixed-time window (left) versus RR-interval frame (right) for the collection of time-sliced ECG data [34]

procedures. Section 3 discusses the testing and validation results, including the confusion matrices for both testing procedures. Section 4 summarizes the new functions of the MATLAB toolbox used to evaluate the performance of the CADs. Finally, our conclusions and a discussion of how our work on the CADs contributes to the wider field are provided in Sect. 5.

## 2 Compact Arrhythmia Detection System Based on Deep Learning

The analysis and classification of ECG data using various ML techniques, including deep learning, has been studied widely [27–31]. CNNs are often used in the context of deep learning, and CNN variants have proven highly successful in classification tasks across different domains [23–25, 32–34]. However, the learning and detection capabilities of CNNs may be insufficient if there is a lot of redundant information, and the analysis of large-scale or highly dimensional data can be computationally demanding. For example, CNNs extract local features gradually from high-resolution feature maps and then combine these features into more abstract feature maps at lower resolution [35]. This is realized by alternating convolution and subsampling layers. The last few layers in the CNN use fully connected MLP-based neural network classifiers to produce the abstracted results. Assuming  $Q$  input feature maps and  $R$  output feature maps, and a feature map size of  $M \times N$ , the convolutional kernel size is  $K \times L$  and the computation in the convolution layer can be represented in a nested-loop description, as shown below:

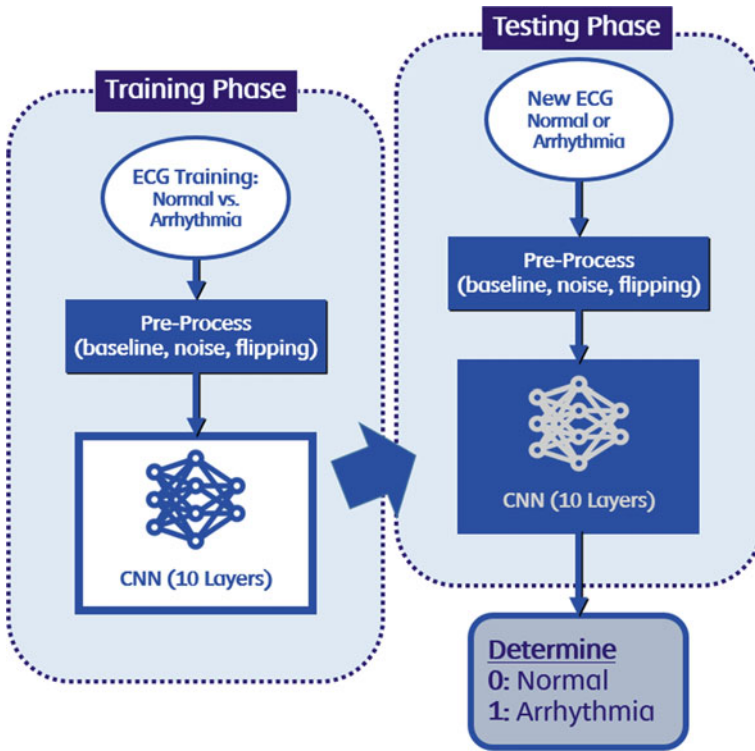
while (r = 0; r < R; r ++ )	//output feature map
while (q = 0; q < Q; q ++ )	//input feature map
while (m = 0; m < M; m ++ )	//row in feature map
while (n = 0; n < N; n ++ )	//column in feature map
while (k = 0; k < K; k ++ )	//row in convolution kernel
while(l = 0; l < L; l ++ )	//column in convolution kernel
$Z[r][m][n] += Y[r][q][k][l] * X[q][m + k][n + l];$	

The array  $X$  contains the input feature maps, and the array  $Z$  contains the output feature maps, which are initialized as zeros. The array  $Y$  contains the weights in the convolution kernels. The computational workload in the convolution layer alone is in the order of  $O(R \bullet Q \bullet M \bullet N \bullet K \bullet L)$ , whereas the computational workload in the next subsampling layer is in the order of  $O(Q \bullet M \bullet N)$ . An activation function is applied to each vector in the feature maps at the output of each layer to mimic neuron activation. However, we developed a CNN with 10 hidden layers to detect arrhythmias in short time ranges (typically < 30 s). This short sampling time allows the CADs to be implemented on portable wearable devices and real-time monitoring equipment. The pre-processing method was also adapted for the training and testing phases to improve the quality of the ECG data [1]. The process flow of the CADs is shown in Fig. 2. Among the standard features of ECG waveforms described by the American National Standards Institute and Association for the Advancement of Medical Instrumentation [36], the CADs uses only the R-peaks.

The process is adapted from previous research by authors and the proposed process for the CADs using CNN is shown in Fig. 2. The process starts onto the training phase to acquire corresponding ECG data from both regular and arrhythmia patients as the training dataset. The RRIF method is used on top of these data after filtering ones with higher quality. The RRIF data is used as the input of a CNN engine as a core process mechanism to generate a CADs evaluation model. The core process in this paper supports the trained CNN engine as the evaluation model for a CADs. The NN Engine is generated when the training phase is completed. In the testing phase, the CADs requests associated with newly received ECG data is generated and the RRIF method after filtering is applied to obtain enhanced data. Then the enhanced data are sent to validation process as the input to check with the NN engine for the final decision on this arrhythmia detection request.

## 2.1 RR-Interval Framed ECG

The new RRIF method slices ECG data based on R-peaks, making each RR-interval a single frame of the sliced data (Fig. 1). Conventional ECG data slicing uses a fixed-time sliding window, an approach known as segmentation. This is widely used to detect major ECG features, such as the P-wave, QRS complex and T wave, allowing the identification of various waveforms for wave feature extraction [18, 36, 37].



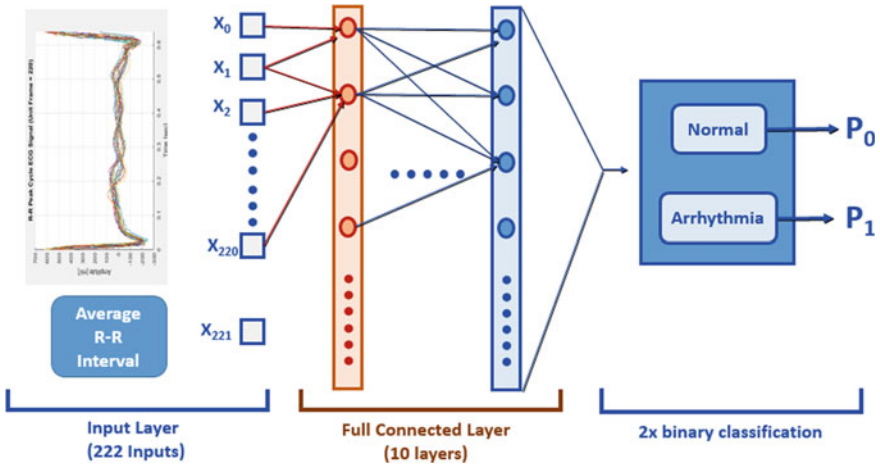
**Fig. 2** Process of the compact arrhythmia detection system (CADS) based on deep learning

During segmentation-based ECG analysis [4, 38–41], it is necessary to find the starting point of the *P*-wave to generate one standard ECG signal [37, 38]. The start of the *P*-wave is considered as the start of the ECG signal cycle. The sliced ECG data are layered by anchoring the R-peak as the base line [2, 17, 40]. R-peak anchoring is effective for ECG analysis that requires no additional feature waveform detection other than the R-peaks, as is the case for biometric authentication [1–3]. Each RR-interval is considered as one cycle of the time-sliced ECG data, and each sliced ECG has a fixed sampling size regardless of the RR-interval. Unlike conventional fixed-time sliding windows, the RRIF method only fixes the sampling size in the RR-interval and does not fix the duration of the R-peaks. The sampling size for each RR-interval can be arbitrary, but it should be consistent when used as input data for a CNN. In the other words, all sliced ECG data should have the same sampling size whereas the cycle time between R-peaks can vary. The sampling values at each position within one RR-interval are used as input parameters for the CNN. In this study, the size of each RRIF was fixed as 220 by referencing a standard ECG signal [37, 38, 42]. However, the actual number of input parameters from the sliced ECG becomes 221 when the potential minimum value of 0 is included, and 222 when we also consider the average RR-interval.



## 2.2 Deep Learning Design for the Detection of Arrhythmia

Based on the sampling data described above, the CNN architecture used in this study has 222 inputs and a binary output that determines whether the input data represent arrhythmia (positive) or normal cardiac behavior (negative). The structure of the CNN is shown in Fig. 3. This particular architecture does not follow conventional ECG classification standards [36], but focuses solely on the detection of arrhythmia using the RRIF method. The ECG sampling data from the RRIF and the average RR-intervals are used as the input parameters for the CNN. The CNN training dataset was gathered from two sources and comprised 20 normal ECG samples with a high sampling frequency ( $> 300$  Hz) from the Diabetes Complications Research Initiative [43] and 20 arrhythmia ECG samples from the PhysioBank database [44]. Following the CNN training phase, another dataset was used for realistic testing. This involved the same number of samples (20 normal and 20 arrhythmia ECG samples) from different individuals, selected randomly from the ECG testing sample to simulate a realistic situation. The proposed CNN was automatically generated using the MATLAB ML function (*nntraintool*). The scaled conjugate gradient method [44] was used for CNN training, and cross-entropy was used to measure performance [45]. We used the WFDB package for MATLAB [46] and the *amgecg* toolbox [1] to develop the framework of the time-sliced RRIF method and CNN.



**Fig. 3** Design of a deep learning approach for the compact arrhythmia detection system (CADS)

### 3 Neural Network Validation and Testing

The training dataset was validated during CNN training. The training samples comprised 1018 sliced data with 222 input parameters, and the binary output followed the Bernoulli probability distribution. The progress measures of the CNN are shown in Table 1. These values were measured automatically using the MATLAB DL function. The best cross-entropy performance was 0.071 at 150 epochs.

The confusion matrix for the training dataset is shown in Table 2. The training dataset was divided into three sections for training (70%), internal validation (15%) and testing (15%). This indicates that the training performance (detection of arrhythmia) was around 96%, which is reasonable compared to a conventional neural network.

Realistic testing was conducted by the random sampling of 20 normal and 20 arrhythmia ECGs to be collected form the different timeslots in the dataset [47] which are not overlapped with the training dataset, thus simulating a realistic testing scenario. The confusion matrix for the testing dataset is shown in Table 3 and the result of using the testing dataset is 100% which is revealing that the results were very competitive even when the ECG sampling time is relatively short and based solely on R-peaks without categorizing any additional waveforms.

**Table 1** Progress of CNN training for arrhythmia detection

Epoch	156 iterations
Time	0.1 s
Performance	0.0594
Gradient	$1.0 \times 10^{-6}$
Validation check	6

**Table 2** Confusion matrix for the CNN training phase

1108 sliced pieces (40 persons)		Actual ECG data		
		Normal	Arrhythmia	
Predicted ECG data	Normal	528 (51.9%)	38 (3.7%)	93.3% 6.7%
	Arrhythmia	6 (0.6%)	446 (43.8%)	98.7% 1.3%
		98.9% 1.1%	92.1% 7.9%	95.7% 4.3%

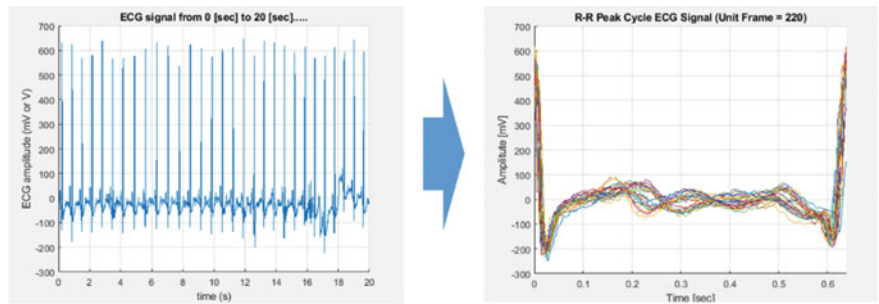
**Table 3** Confusion matrix for the realistic testing of CADs

100 trials (40 individuals)		Actual ECG data	
		Normal	Arrhythmia
Predicted ECG data	Normal	50	0
	Arrhythmia	0	50

4 MATLAB Toolbox: Medical Extension

The work described in this paper, including the MATLAB code, is based on the *amgecg* toolbox we have developed [1]. Here, we introduce the *amgecg* toolbox Medical Extension for medical applications. The previous sections describe the adaptation of core components for the CADs, and the proposed methods are implemented as MATLAB functions. The *amgecg* toolbox Medical Extension adds new functions to the basic *amgecg* toolbox [1], which researchers can use for their own ECG analysis projects. Some of the functions in this extension package are discussed below. Time-sliced RRIF is the core method described in this article, and this function requires some components of the basic *amgecg* toolbox as well as three components in the extension, namely *ecgrpeakframe*; *deltamean*; and *sqframechanger*. The MATLAB function *ecgrpeakframe* in the package generates the RRIF data and displays the sliced ECG data as either a 2D chart (Fig. 4) or a 3D graph (see Fig. 1b). This function provides the reference values of the sliced samples based on both mean and mode. Users can try demos of the new functions in the package, and functions such as *ecgrpeakframe* combine some basic and some integrated functions of the *amgecg* toolbox. The functions *deltamean* and *sqframechanger* are explained in MATLAB and can be accessed using the help menu.

The MATLAB source codes (*amgecg* Toolbox Medical Extension) are publicly available online and users can download either the full package (*amgecg* Toolbox + Medical Extension) or just the extension package, both of which feature sample demos. Furthermore, video clips of the demonstration are available on YouTube to help users become familiar with the toolbox, and allowing them to implement their own codes based on the extended package.



**Fig. 4** 2D graph generated using the *ecgrpeakframe* function

## 5 Conclusions

As new ECG sensors become portable for compatibility with smartphones and wearable devices, ECG-based analysis for healthcare applications will become more common. In this article, we have described a novel system (CADS) for the detection of arrhythmia which could be implemented on wearable and portable devices. Our approach involves a new way to build the input parameters for the training of a deep learning algorithm without the need to analyze complex ECG waveforms. Although a CADS is not targeted to handle huge training data for detecting various heart diseases, it is fast and easy to implement into even an IoT device which has limited resources and less computing powers. The RRIF makes this deep learning-based detection system flexible and competitive, achieving 96% accuracy during training and validation and 100% under realistic test conditions.

## References

1. Kim S-K, Yeun CY et al (2019) A machine learning framework for biometric authentication using electrocardiogram. *IEEE Access* 7:94858–94868
2. Alzaabi E, Kim S-K et al (2019) Electrocardiogram biometric authentication system by using machine learnings. *IEEE Access* 7:123069–123075
3. Kim S-K, Yeun CY et al (2019) An enhanced machine learning-based biometric authentication system using RR-interval framed electrocardiograms. *IEEE Access* 7:168669–168674
4. Luz EJ, Schwartz WR et al. (2016) ECG-based heartbeat classification for arrhythmia detection: a survey. *Comput Methods Programs Biomed* 127:144–164
5. Islam MR, Hossain R (2015) Arrhythmia detection technique using basic ECG parameters. *Int J Comput Appl* 119:11–15
6. Pummer C (2016) Continuous biometric authentication using electrocardiographic (ECG) data. <https://usmile.at/publications>. Accessed 1 Apr 2019
7. Zhang Q, Zhou D, Zeng X (2017) HeartID: a multi resolution convolution neural network for ECG-based biometrics human identification in smart health applications. *IEEE Access* 5:11805–11816
8. Pinto JR, Cardoso JS, Lourenco A (2018) Evolution, current challenges, and future possibilities in ECG biometrics. *IEEE Access* 6:34746–34776
9. Luz EIS, Moreira GJP, Oliveira LS et al. (2018) Learning deep off-the-person heart biometrics representations. *IEEE Trans Inf Forensics Secur* 13:1258–1270
10. Kim H, Chun SY (2019) Cancelable ECG biometrics using compressive sensing-generalized likelihood ratio test. *IEEE Access* 7:9232–9242
11. Guennoun M, Abbad N et al. (2009) Continuous authentication by electrocardiogram data. In: *IEEE Toronto international conference science and technology for humanity*, Toronto, ON, pp 40–42
12. Spach MS, Kootsey JM (1983) The nature of electrical propagation in cardiac muscle. *Am J Physiol Heart Circ Physiol* 244:3–22
13. McGraw R, Lord J et al. (2019) analysis and interpretation of the electrocardiogram, <https://meds.queensu.ca/central/assets/modules/ts-ecg/index.html>. Accessed 1 Apr 2019
14. Goldberger AL, Amaral LAN et al (2000) Physiobank, physio toolkit, and physioNnt: components of a new research resource for complex physiologic signals. *Circulation* 101:e215–e220
15. Gacek A, Pedrycz W (2012) ECG signal processing, classification and interpretation. Springer, New York, NY

16. Kachuee M, Fazeli S et al. (2018) ECG heartbeat classification: a deep transferable representation. In: 2018 IEEE international conference on healthcare informatics, New York, NY, 443–444
17. Bhatti AT, Kim JH (2015) R-peak detection in ECG signal compression for heartbeat rate patients at 1KHz using high order statistic algorithm. *J Multidisciplin Eng Sci Tech* 2:2509–2515
18. Zhong W, Liao L et al. (2019) A deep learning approach for fetal QRS complex detection. *Physiologic Measur* 39(4):045004
19. Bennett FH et al. (1996) Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: *Artificial Intelligence in Design*, vol 96. Springer, Dordrecht, pp 151–170
20. Mincholé A, Rodríguez B (2019) Artificial intelligence for the electrocardiogram. *Nat Med* 25:22–23
21. Tatara E, Cinar A (2002) Interpreting ECG data by integrating statistical and artificial intelligence tools. *IEEE Eng Med Biol Mag* 21:36–41
22. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances*. In: *Neural information processing systems*, pp 1097–1105
23. Yu H, Xie T et al. (2011) Comparison of different neural network architectures for digit image recognition. In: *Proceeding IC-HIS*. Yokohama, Japan, pp 98–103
24. Weiss SM, Kapouleas I (1989) An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In: *Proceedings IJCAI*. <https://www.ijcai.org/Proceedings/89-1/Papers/125.pdf>. Accessed 1 Apr 2019
25. Szandal T (2015) Comparison of different learning algorithms for pattern recognition with hopfield's neural network. *Proc Comput Sci* 71:68–75
26. Roopa CK, Harish BS (2017) A survey on various machine learning approaches for ECG analysis. *Int J Comput Appl* 163:25–33
27. Kiranyaz S, Ince T et al (2016) Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Tran Biomed Eng* 63:664–675
28. Rajpurkar R, Hannun AY et al (2019) Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* 25:65–69
29. Mitra M, Samanta RK (2013) Cardiac arrhythmia classification using neural networks with selected features. *Procedia Technol* 10:76–84
30. Mondejar-Guerra V, Novo J et al (2019) Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers. *Biomed Signal Process Contr* 47:41–48
31. Isina A, Ozdalilib S (2017) Cardiac arrhythmia detection using deep learning. *Proc Comput Sci* 120:268–275
32. Gerven M, Bohte S (2017) Artificial neural networks as models of neural information processing. *Front Comput Neurosci*. <https://www.frontiersin.org/articles/https://doi.org/10.3389/fncom.2017.00114/full>. Accessed 1 April 2019
33. Yann L (2019) LeNet-5, Convolutional neural networks. <http://yann.lecun.com/exdb/lenet/>. Accessed 1 Apr 2019
34. Yu W, Yang K et al. (2012) Visualizing and comparing convolutional neural networks. <https://arxiv.org/abs/1412.6631>. Accessed 1 Apr 2019
35. Taddei A, Distanti G et al (1992) The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. *Eur Heart J* 13:1164–1172
36. ANSI/AAMI (2008) Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms, American National Standards Institute, Inc. (ANSI), Association for the Advancement of Medical Instrumentation (AAMI), ANSI/AAMI/ISO EC57. 1998–(R)2008
37. Pater C (2005) Methodological considerations in the design of trials for safety assessment of new drugs and chemical entities. *Trials* 6(1):1–13
38. Cadogan M (2019) PR Interval. <https://litfl.com/pr-interval-ecg-library/>. Accessed 1 Apr 2019

39. Afonso VX, Tompkins WJ et al (1999) ECG beat detection using filter banks. *IEEE Trans Biomed Eng* 46:192–202
40. Hu YH, Tompkins WJ, Urrusti JL, Afonso VX (1993) App of artificial neural networks for ECG signal detection and classification, *J Eletrocardiol* 26:66–73
41. Xiang X, Lin Z, Meng J (2018) Automatic QRS complex detection using two-level convolutional neural network. *BioMed Eng OnLine* 17:13. <https://biomedical-engineering-online.biomedcentral.com/articles/https://doi.org/10.1186/s12938-018-0441-4>. Accessed 1 Jan 2020
42. Alarsan FI, Younes M (2019) Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *J Big Data* 6(8):1–15
43. Imam MH, Karmakar CK et al (2016) Detecting subclinical diabetic cardiac autonomic neuropathy by analyzing ventricular repolarization dynamics. *IEEE J Biomed Health Inf* 20:64–72
44. Moody GB, Mark RG (2001) The impact of the MIT-BIH arrhythmia database. *IEEE Eng Med Biol* 20:45–50
45. Straeter TA On the extension of the Davidon-Broyden class of rank one, Quasi-newton minimization methods to an infinite dimensional Hilbert space with applications to optimal control problems. NASA Technical Reports Server. NASA
46. Viana M (2019) Loss Functions. [https://ml-cheatsheet.readthedocs.io/en/latest/loss\\_functions.html](https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html). Accessed 1 Apr 2019
47. Aspuru J, Ochoa-Brust A et al. (2019) Segmentation of the ECG signal by means of a linear regression algorithm. *Sensors* 19(4):775. <https://doi.org/10.3390/s19040775>

# How Predictive Software Engineering Addresses Issues in Custom Software Development and Boosts Efficiency and Productivity



Boris Kontsevoi, Sergey Kizyan, and Irina Dubovik

**Abstract** Many of the bottlenecks facing businesses today can be overcome with an improved service delivery framework. This paper describes the concept of Predictive Software Engineering (PSE) and how it ensures transparency, controllability, and predictability in software development, mainly focusing on the seven principles that the framework is based on. Each principle offers a unique way of benefiting the company that implements it and is associated with its own set of strategies, best practices, methodologies, and important KPIs. Throughout the paper, the authors detail the advantages of using PSE in custom software development, including project transparency and budget management. Together, the principles can transform and streamline processes associated with custom products and help build successful businesses. The PSE framework draws on over 27 years of experience in software development and considers recognized and generally accepted practices.

**Keywords** Predictive Software Engineering · Agile · Disciplined agile delivery · Meaningful customer care · End-to-end control · Proven productivity · Efficient distributed team · Disciplined agile delivery (DAD) · Sound human development · Measurable Quality Management and Technical Debt Reduction System (MQM and TDR)

---

B. Kontsevoi · S. Kizyan · I. Dubovik (✉)  
Intetics Inc, Naples, FL, USA  
e-mail: [i.dubovik@intetics.com](mailto:i.dubovik@intetics.com)

B. Kontsevoi  
e-mail: [boris@intetics.com](mailto:boris@intetics.com)

S. Kizyan  
e-mail: [s.kizyan@intetics.com](mailto:s.kizyan@intetics.com)

## 1 Introduction

Predictive Software Engineering, or PSE, is a framework designed to eliminate or at least minimize the unpredictability in software development. With engineering made more precise and predictable, development teams can become better collaborators and produce deliverables in a more controlled manner. These benefits are made possible by key PSE principles.

The following sections will focus on the seven principles of Predictive Software Engineering and how they contribute to transparent, controllable processes in software engineering. Each section draws on over 27 years of experience in software development.

## 2 Meaningful Customer Care

A customer is put first in any type of organization—whether one sells directly to individual consumers, businesses, or other parties. By extension, the success of these organizations is measured in customer-related parameters—the number of customers, customer satisfaction, loyalty, etc.

Businesses can turn to customer feedback and satisfaction as a way to measure their performance and progress. The reason to use customer-related metrics as primary KPIs is the integral link between what customers think about an organization and their purchasing decisions. The more customers recognize the value of a product or service, the more sales an organization will have.

The opposite is also true: a dissatisfied customer does not recognize the value of a product or service, so they will not make further purchases or return to the business. If not addressed, this will ultimately cause the business to fail.

These eight strategies will help organizations deal with unhappy clients and create an even better relationship with satisfied customers:

1. Build a customer portal with self-service tools.
2. Map out a smooth customer onboarding process.
3. Set and follow the standards for transparency in governance.
4. Develop an efficient plan for customer escalation.
5. Prioritize early problem discovery.
6. Investigate customer complaints and act on the analysis.
7. Keep track of best practices for improving customer satisfaction.
8. Manage customer expectations and set realistic boundaries.



### 3 Transparent End-To-End Control

While full transparency can be hard to achieve, it is crucial to progress on an individual and a company-wide level. It is the basis of productive cooperation, trust, and openness. For organizations that make an effort to be transparent, the results justify the resources spent—it demonstrates your reliability to the customers, proves that there is minimal risk in working with you, and positions you ahead of less transparent competitors.

To build trust with external stakeholders:

1. Use agile project management.
2. Regularly discuss governance across teams.
3. Introduce workload management and time-tracking software.
4. Write project status reports.
5. Give customers access to project-tracking solutions.
6. Audit business processes.
7. Correlate site visits with major events—the project launch, major deliveries, any project issues, etc.
8. Manage proactively rather than reactively.

To create a culture of internal transparency:

1. Conduct check-in meetings with project managers.
2. Have the department head present at your meetings.
3. Improve relationships with team-building activities.
4. Get the talent management team involved.

### 4 Proven Productivity

Defining and measuring the productivity of software engineers is hard, if not impossible, to do accurately. However, every organization should still have productivity goals and the tools to monitor them. To measure the amount of work that a team can do and has already done, you can use a Measurable Quality Management and Technical Debt Reduction System (MQM and TDR) and KPIs for Software Development Efficiency.

The former tool will be covered shortly. As for the latter solution, there are 10 KPIs to focus on:

1. Planned versus actual hours
2. Static code analysis (the entire project)
3. Static code analysis (individual developers)
4. Bugs per line of code (per 1000 lines)
5. Templates missed
6. Bugs per feature
7. Algorithms missed

8. Standard libraries missed
9. Mishandled exceptions and errors
10. Security issues

Best practices for measuring productivity with KPIs:

- Have at least two code reviews from different developers.
- Document code reviews from multiple perspectives (author, review, and team).
- Have the team lead or project manager go over and approve the results.
- Automate the review process to save time and eliminate the human factor.
- Review the code incrementally (less than 500 lines at a time).

## 5 Efficient Distributed Team

Modern organizations often integrate the distributed team model, which spreads them out across different locations and allows them to grow a global talent pool. The model benefits organizations in three main vectors:

- **Cost savings:** When companies hire remote developers from abroad, it helps them save money on office rent, operating costs, and, often, labor costs.
- **Time efficiency:** Distributed teams allow companies to cover different time zones and operate for longer periods of time. Strategic placement of development hubs can even provide you with 24/7 service. With one or two teams always available, your operations can continue non-stop, and customer support can be accessible at all times.
- **Diverse workforce:** The distributed model gives access to a global talent pool with virtually limitless hiring opportunities [1].

The obvious drawbacks of distributed tech teams are communication issues and inefficiencies caused by them. With the team members spread out over great distances, some of them may never meet the others in person, and even real-time communication can be rare. To avoid misunderstandings and workflow disruptions:

1. Take into account workloads, local and industry-wide regulations, and time constraints when setting tasks.
2. Clearly define roles and responsibilities for all team members.
3. Organize proper communication procedures and practice closed-loop communication to avoid errors of omission.
4. Minimize instances when the workload of one employee rests on the delivery of other people.
5. Introduce mutual monitoring to improve team performance [2].

You should also account for cultural differences, which can create a mismatch of communication styles and hinder team cohesion. To bring together developers from vastly different areas, a business needs strong leadership, which will facilitate group dynamics, motivate employees, and provide coaching and mentoring when needed.

It also helps to find shared ideas, interests, or beliefs to make developers feel more connected to each other. If possible and when appropriate, the team should work and train in a shared environment.

## 6 Disciplined Agile Delivery Process

Agile software development is common across organizations, but there are thousands of methodologies under this umbrella. Based on our past projects, we found that disciplined agile delivery (DAD) has the highest organizational effectiveness.

DAD supports a different delivery life cycle compared to other agile methods and calls out three development phases: inception, construction, and transition [3].

### A. Inception

- State the project vision.
- Ensure the stakeholders agree with the vision.
- Write down the project plan, requirements, and tech strategy.
- Build the initial team.

### B. Construction

- Address the changes in stakeholders' needs.
- Develop a potentially releasable product.
- Bring the product closer to deployment.
- Achieve or maintain quality performance.
- Manage the biggest risks.

### C. Transition

- Prepare the solution for deployment.
- Present the product to the stakeholders.
- Deploy.

## 7 Measurable Quality Management and Technical Debt Reduction

The goal of the Measurable Quality Management and Technical Debt Reduction System (MQM&TDR) is to measure the quality, technical debts, and economic efficiency of the software.

MQM and TDR is relevant for everyone involved in the product—developers and testers receive performance reviews, managers understand how the supply for the project matches the demand, and users get access to a high-quality product. It also allows investors to assess risk and set a fair market value.

The system evaluates the following software components:

- Source code quality
- Usability
- Security
- Performance
- Business logic
- Solution architecture and data model
- Data quality
- Use of third-party code [4]

The benefits of the system for development teams, product owners, and investors are:

1. Allows to easily measure performance and implement changes as needed.
2. Keeps technical debts under control.
3. Ensures lower support and maintenance expenditures.
4. Helps forecast economic efficiency.
5. Makes it possible to conduct a comprehensive quality analysis.
6. Provides a product feature analysis.
7. Conducts a compliance check.
8. Offers improvement suggestions.

MQM and TDR is suitable for all software development projects, as confirmed by numerous clients.

## 8 Sound Human Development

Adopt a more human-centered approach to HR. Rather than treating employees as assets, view them as agents with their own goals, skills, and decision-making abilities.

The change of perspective on human development does more than simply improve the productivity of the workforce. It also builds employer-employee loyalty and helps employees find fulfillment in work [5].

### A. Benefits

On top of standard employee benefits (medical insurance, free food, and gym membership), consider providing other types of compensation (language lessons, professional training, etc.).

### B. Mentorship

Mentoring programs provide learning opportunities from real projects, converting theoretical knowledge into practical skills under tactful guidance. Such programs aren't solely focused on technical skills—they also encourage participants to advance

their soft skills, like communication and business processes. By the end of the program, the best trainees can join the team officially.

### C. Performance Review and Career Growth

Formulate individual development plans for employees to establish their personal and career goals. Schedule a meeting with each employee and the talent management team and plan together. For an even more personalized approach, allow employees to conduct their own performance reviews once or twice a year. Discuss the scores with the employee and determine what needs to be accomplished during the next cycle for better results.

### D. Skill Improvement

Support ongoing education for developers, whether it is training, courses, or programming certifications. Some of the best skill improvement initiatives are Centers of Excellence, internal and online training, and outside experts to advise your team.

## 9 Practical Approach to PSE

Introducing PSE into a company can be complicated. Companies that have previously made a transition to agile methodology may be familiar with the challenge of changing old habits and mindsets. Similarly, as you transition to PSE, you should foster an environment where your developers can unlearn and relearn. That means being accepting of hiccups during a contention period in the workplace.

There needs to be involvement and effort from all levels of company management. Without the help of top management and the CEO, transformative actions will not take root. Or worse, the new methodology can be detrimental to the business—it may be adopted unevenly, fail to consider unique business needs, or misalign with future plans and strategies.

In some sense, PSE builds on agile processes, which should make implementation relatively smooth. The first step is to have the management on board with the transition. Next, the approach needs to be adapted to fit the company in question and its goals. Finally, once the PSE processes are built, they need to be standardized across the organization.

In other words, PSE implementation will differ from one company to the other but should be internally uniform across each of the companies.

Suppose two companies need to hire ten new employees each. For one company, it may mean working with a manager internally; for another company, it may mean working with a recruiting agency. Neither case contradicts the principles of PSE.

## 10 Risk that PSE Covers

Predictive Software Engineering and the principles within the framework help reduce the following risks in software development:

- Poor project scope management (transparent control)
- Lack of executive involvement (meaningful care and transparent control)
- Estimation pitfalls (disciplined agile)
- Resource scarcity (distributed teams)
- Uncertainty in decision-making (meaningful care)
- Inadequate training (sound human development)
- Project delays (transparent control)

## 11 Predictability as a Solution to Outsourcing Problems

Development services commonly operate on the collaboration and teamwork principles of scrum and agile. But considering the rising demand for a predictable timeline in software development projects, the methodologies do not hold up.

Predictive Software Engineering improves on the preceding methodologies. While agile development does not offer a clear vision of the “final product,” PSE introduces a much higher degree of predictability for the vision, deadlines, and other aspects. It involves multiple stakeholders, too; however, it’s still possible to quantify the full extent of the required efforts. The focus on precision makes PSE a more suitable model in projects where reliability is more important than flexibility.

What is more, PSE does not fail to take into account the bigger business goals. It aims to cover the full spectrum of activities surrounding software development, from the technical to the economic.

**Acknowledgements** Predictive Software Engineering is a much-needed framework that addresses bottlenecks in custom software development, ultimately establishing a reliable, systematic delivery of development services.

PSE does not view programming as a creative, unstable undertaking. Instead, it brings software engineering to what it should be: a predictable and precise operation.

The framework consists of seven principles, with each contributing to transparency, controllability, and predictability. PSE is a proprietary framework by Intetics, which draws on over 27 years of providing custom development services.

The PSE framework can be taken further! We are looking forward to working with other development companies and invite you to join us. Predictive Software Engineering goes beyond Intetics—it aims to transform the industry and programming technologies as a whole. Let us move past competing and join forces to enhance our professional practice.

This paper was written under the leadership of Intetics’ President and CEO, Boris Kontsevoi, and Sergei Kizyan, Delivery Director, Sandbox. The principles are based on recognized and generally accepted practices.

## References

1. Kontsevoi B, Kizyan S (2022) Predictive software engineering: transform custom software development into effective business solutions. *J Econ Finan Manage Stud* 5(01):73–77. <https://doi.org/10.47191/jefms/v5-i1-09>
2. Larsson A, Törlind P, Karlsson L, Mabogunje A, Leifer L, Larsson T, Elfström B-O (2003) Distributed team innovation – a framework for distributed product development. In: 14th International conference on engineering design 2003, ICED'03, vol 322. Design Society, Stockholm, pp 1–10. ISBN: 1-904670-00-8
3. Ambler SW, Lines M (2013) Introduction to disciplined agile delivery. *Crosstalk J* 40:7–11
4. Kontsevoi B, Terekhov S (2021) TETRA™ techniques to assess and manage the software technical debt. *Adv Sci Technol Eng Syst J* 6(5):303–309. <https://doi.org/10.25046/aj060534>
5. United Nations Development Programme Human Development Report (2016) <https://www.undp.org/publications/human-development-report-2016>. Last accessed 10 Oct 2018

# Using Conceptual Chunking to Support Information Processing While Solving Complex Industrial Tasks



Anja Klichowicz , Tina Morgenstern, and Franziska Bocklisch

**Abstract** Monitoring and controlling human–machine systems in intelligent industrial manufacturing becomes more and more complex. As human skills and expert knowledge are valuable, operators and cyber-physical systems (CPS) should complement one another as team partners. Efficient teaming requires a deeper understanding of human cognition, such as memory processes. Conceptual chunking is one strategy to optimize working memory performance by integrating a number of small information units and their interrelations to a larger one. Graphical visualizations (e.g., in industrial control panels) can support teaming and understanding of complex interactions by highlighting these relationships. The aim of the present study was to investigate whether graphical design elements enhance conceptual chunking. In an experiment ( $N = 40$ ), graphical design elements (i.e., coordinate systems vs. side-by-side bar graphs) were used to induce or inhibit conceptual chunking. Response accuracies, response times, gaze data, and solving strategies were assessed. Results reveal that participants rely more often on graphical design elements when relations between variables are presented (coordinate system). If conceptual chunking was induced successfully, participants showed more correct answers and needed less time for information search. Graphic presentations displaying relations between variables seem to be suitable to support the understanding of complex tasks. This indicates the high potential for teaming between humans and machines in intelligent industrial manufacturing systems.

**Keywords** Eye tracking · Human machine teaming · Information processing · Memory · Thermal spraying

---

A. Klichowicz (✉) · T. Morgenstern · F. Bocklisch  
Chemnitz University of Technology, Erfenschlager Strasse 73, 09125 Chemnitz, Germany  
e-mail: [anja.klichowicz@mb.tu-chemnitz.de](mailto:anja.klichowicz@mb.tu-chemnitz.de)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_4](https://doi.org/10.1007/978-981-99-3091-3_4)

33



## 1 Introduction

Understanding human information processing is an important issue for cognitive psychologists to get a deeper insight into human thinking and to improve the interaction with the environment in everyday life. One example, in which the understanding of cognitive processes is enormously important, is the field of human–machine interaction in the industrial context. Many control tasks, such as supervisory control of high-risk industrial systems [1, 2], pose high requirements on the human cognitive system in order to operate machines efficiently, keep technical processes stable and avoid accidents and risks for human operators. Mostly, the operator has to control and adjust multiple interacting variables at any given time in complex and dynamically changing technical systems. Therefore, technical systems should be designed in accordance with the strength of the cognitive system to allow teaming. The constitution of a mixed agent team of humans and cyber-physical systems (CPS) that collaboratively pursues goals unachievable or at least not wholly achievable individually ([3]; p. 5) is meaningful because of the interdependence relationship determined by the joint working task [4]. Teaming as a process reflects the dynamic (re-) arrangement of the elements of the system into a fluid team structure and enables human-oriented automation with flexible function allocation and control [5] by clarifying “what” is done by “whom” and “when” [6]. One of the key challenges that arise from this conceptualization of human–machine-teaming is assuring transparency in machine operations; thus, facilitating human understanding and supporting situation awareness [7, 8].

### *1.1 Using Conceptual Chunking for Understanding Complex Machine Processes*

In simplified terms, understanding a technical system means to connect new information (e.g., changes in operating state or fault signals) with what is already known (e.g., based on prior experiences or professional expertise). Hence, collected information is linked to knowledge from long-term memory and integrated into a mental representation of the current situation [9], allowing the human operator to take the machines’ perspective [10]. This representation is located in working memory, which is the system that coordinates all processes that are needed for thought and action [11]. Although memory processes have long been the subject of intense research (e.g., [11–13]), this knowledge is not yet implemented in technical systems and their interfaces to a satisfying degree. For example, working memory is very limited in its capacity and cognitive operations (especially in human interactions with CPS) can be rather complex [13, 14]. How information is processed and how many resources are required for this, is determined by three features [15]. The first feature is the person that is processing the information. If the operator is an expert (i.e., the information is not entirely new to her/him), the information can be understood within an

elaborate reference frame retrieved from long-term memory. Hence, less working memory capacity is needed. Thus, familiar information is “easier” to process [15]. The second feature is information complexity and the third feature is the type of information presentation [12, 14]. In the industrial context, it is not realistic that all operators have the same degree of expertise. However, there are different strategies to help operators with different levels of expertise to transfer new (and complex) information into long-term memory, and thereby, allowing them to use that information in the ongoing process with reduced demands on working memory capacity.

For example, chunking of information is beneficial to reduce working memory load [16]. Chunking was first described by Miller [17] as construction of larger information units by closely integrating smaller ones [18]. Thereby, an (theoretically unlimited) amount of information can be stored at the same costs as a single information when organized in a chunk [19]. This is because only a pointer to this large chunk remains in working memory, whereas the chunk itself is transferred into long-term memory [12]. Within a chunk, elements or items are strongly associated [19]. The pointer acts as a direction sign to allow quick retrieval from long-term memory when the information is needed [20]. Thus, chunking is able to optimize performance on working memory tasks [21].

However, in the industrial context, processed information are often rather complex. For example, in thermal spraying—a manufacturing technology of fundamental importance for key industrial and medical application [22, 23]—human operators have to monitor and control complex technical systems (spraying units). During the coating process, material, such as powder particles, is melted by high-energy sources (e.g., plasma) and sprayed onto prepared work pieces for surface composition (e.g., corrosion-preventive coating). It is important that the spraying process has the exact temperature, which is dependent on other variables, such as the mixtures of processing gases (e.g., hydrogen and argon) needed for plasma generation. If the process fails, the resulting surface is of low quality. Consequently, the operator does not only need to know *which* variables are involved, but also *how* they interact. Hence, saving the items of the chunk in long-term memory might not be enough when those items are interacting. If the operator creates a chunk that also includes such interactions, the process is called *conceptual chunking* [24]. Thereby, only an abstract, one-dimensional picture of the information remains in working memory. All other dimensions and their interactions can be accessed with the pointer from long-term memory. In the example, the operator would only keep the temperature (dimension 1) in her/his working memory representation, but could unfold temperature as a function of the proportion of substance 1 (dimension 2, e.g., hydrogen) to substance 2 (dimension 3, e.g., argon) from long-term memory, whenever needed.

Visualizations, such as graphical elements, can easily be implemented in machine displays and address the cognitive strength of the human team partner, who acquires most information through vision [25]. Humans are able to interpret visual input without training and to process data properties that could not be extracted from written or spoken language [25]. However, graphs are a flexible form of communication as the reader can control which information is extracted [26]. Therefore, the graphic presentation needs to actively communicate the intended content. Not

all graphic presentations are suitable to facilitate conceptual chunking. For example, side-by-side bar graphs illustrate distributions across a number of categories [25]. Data is broken down according to these categories making it more difficult to judge relationships and proportions. Hence, bar graphs elicit *perceptual* chunks. That is, information is summarized by category if several bars are organized in one category. In contrast, coordinate systems are able to represent one dimension as a function of another, and hereby, show interrelations and dependencies. Coordinate systems allow compressing a number of variables and their interactions in one abstract dimension, and thus, induce *conceptual* chunking.

## 1.2 Present Research

The aim of the current study was to investigate whether conceptual chunking can facilitate the understanding of complex machine processes, and thus, improve teaming in human-machine systems. In order to extract and systematically investigate comprehension processes out of the high number of cognitive processes (e.g., perception, storage or evaluation of information), a standardized lab setup was used. Thus, we aimed to bridge the gap between research on cognition and the applied context, allowing us to induce conceptual chunking under controlled conditions and enabling a transfer to real thermal spraying applications.

In the present experiment, conceptual chunking was induced using graphic visualizations as they can enhance the understanding of complex interactions [25, 27]. The present study comprised three parts and introduced eye tracking as a tool for assessing information search. First, during an instruction phase, participants gained knowledge regarding the application of thermal spraying and the involved variables and their relationships through written explanations. This was accompanied by either a bar chart (unchunked condition) or a coordinate system (chunked condition). During this phase, fixations were tracked to investigate information acquisition (graphical elements vs. written explanations). Second, in a practice phase, the task relevant knowledge gained in the instruction phase was tested. Third, during the test phase participants solved ten different single-choice tasks. Here, eye tracking was used to (1) account for information search and (2) give insight into information retrieval. Presenting a placeholder to replace the graphical element only displaying the axis, but not the content, enables us to monitor retrieval without allowing participants to gain information. This so-called looking at nothing behavior is well investigated and a validated way to assess memory retrieval ([28, 29]; for an overview see [30, 31]). Further, eye movements support information retrieval [29, 32]. Therefore, they provide information on whether the operator is able to efficiently access the information through the pointer as part of the mental representation of the current system in working memory (see [11, 12, 24]) without the need of another helping strategy.

### **1.3 Hypotheses**

First, according to the previous explanations, we predicted that participants' understanding of complex machine processes benefits from information presented in a coordinate system compared to information presented in side-by-side bar charts (H1). This should be evident in a more extensive use of the graphical element during the instruction phase (H1a). To assess this, gaze data was used. Further, participants that learned the relations between variables in the instruction phase using a coordinate system (i.e., chunked condition) tend to use these relations to solve subsequent tasks more often than participants using visualizations not displaying relations (i.e., side-by-side bar charts, unchunked condition). Therefore, we assumed that more participants in the chunked condition are able to store and use the relationship between variables to solve the tasks (i.e., engage in conceptual chunking) than participants in the unchunked condition (H1b). To assess this, participants were asked to report their individual strategies for solving the tasks.

Second, we predicted that once participants engage in conceptual chunking, they are able to solve the tasks faster (H2a) and more accurate (H2b) than those, who did not engage in conceptual chunking. To address this hypothesis (H2), participants' performance parameters (i.e., response accuracy and response time) were calculated.

Third, we expected that participants are able to solve the task without additional information search or retrieval strategies, if participants used conceptual chunking as a strategy. To address this exploratory research question, we analyzed gaze data during the test phase.

## **2 Method**

### **2.1 Participants**

Forty participants (mainly students, 27 female, 25 years, ranging from 18 to 51) enrolled at Chemnitz University of Technology volunteered in the experiment in exchange for student course credit. As this was a controlled laboratory study, a student sample was used. Hence, it was important that all participants had the same prior knowledge (i.e., no knowledge) concerning thermal spraying and gained all used information only through the instructions. All participants had normal or corrected to normal vision. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Chemnitz University of Technology. Informed consent was obtained from each participant.

2.2 Apparatus and Material

Gaze data was recorded using a binocular IViewX RED eye-tracking-system from SensoMotoric Instruments with a sampling rate of 120 Hz. Stimuli were presented on a 22 inch computer screen using EPrime 2.0 software and with a resolution of 1680 × 1050 pixels. All subjects sat at a distance of 450–650 mm in front of the screen. BeGaze 3.0, Microsoft Excel 2007, IBM Statistics 28 (SPSS) and JASP.0.14.00 were used to analyze data.

The experiment consisted of an instruction, a practice and a test phase. Areas of Interest were assigned according to Fig. 1.

During the instruction phase, on the left hand side, the explained variables were displayed by a graph. For half of the participants, the variables were presented in side-by-side bar charts, not allowing to grasp interactions between the variables (unchunked condition; see Fig. 2a). The other half of the participants received visualizations depicting the variables in a coordinate system, indicating interactions between the variables (chunked condition, see Fig. 2b). During the test phase, the graphical element was replaced by its outline.

On the upper part of the right hand side, written information (i.e., explanation in the instruction phase; the current state of the system and answer alternatives in the test phase, see Fig. 1) was presented. For the present study, the process of thermal spray coating was used in a simplified manner. During thermal spray coating, the process is monitored and adjusted using the display of the spraying unit. The success of the process is mainly influenced by plasma temperature. For the present experiment, an optimal temperature of 10,000 °C was defined, which is to be accomplished by mixing the correct proportion of the process gases argon and hydrogen. More hydrogen leads to increasing plasma temperature, whereas more argon decreases the temperature of the plasma torch. Further, due to technical restrictions, the spraying system limits the amount of each gas that can be inserted into the system.

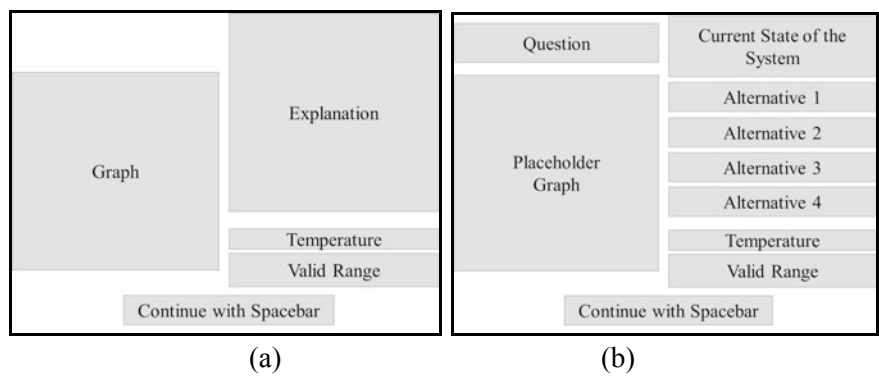
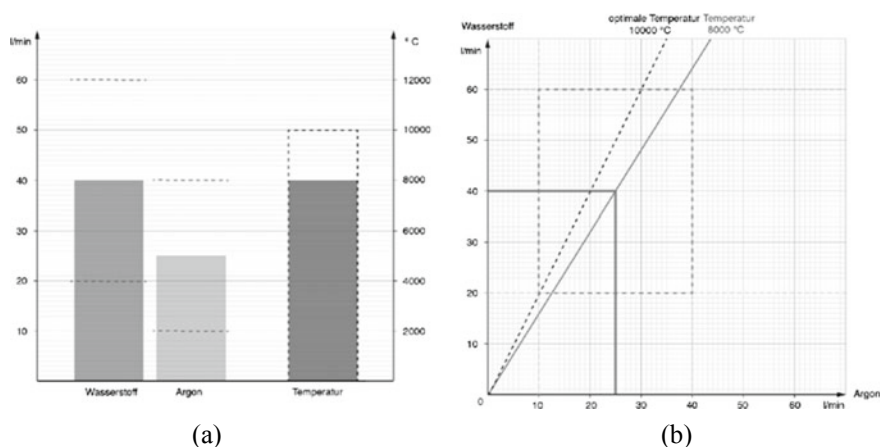


Fig. 1 Areas of interest (AOI) during instruction (a) and practice and test phase (b)



**Fig. 2** Graphic presentations in the unchunked (a) and chunked (b) condition

The lower part of the right hand side displayed permanent system information that did not change throughout the experiment, such as the optimal temperature of the plasma and the valid range of the gases (see Fig. 1).

When studying the graphic presentation and working through the explanations and examples in the instruction phase, it became evident, that hydrogen and argon had to be inserted into the system by a precise ratio of two to one. This ratio was easier to understand in the chunked condition than in the unchunked condition as it could be deduced from the coordinate system. Participants in the unchunked condition had no clear implication of the ratio in the graphic presentation as bar graphs are not fit to display relations. Still, participants had the chance to understand this relation from an explained example. Further, participants must ensure that the amount of gas inserted into the system is kept in the valid range, which was displayed throughout the entire time for each task.

Stimuli presenting participants' task were organized in a similar way. However, the graph was replaced with a placeholder only displaying the axis, but not the content. Above the placeholder, the tasks question ("What has to be done to adjust the system to its optimal state?") was displayed. On the upper part of right hand side, where explanations were presented before, the current state of the system was presented. Participants were instructed to choose one out of four options in order to adjust the proportion of the gases argon and hydrogen to achieve the optimal temperature. The bottom right part containing permanent system information stayed visible at that location throughout the entire experiment.

The short questionnaire in paper-pencil format was used to control for prior knowledge and asked participants to rate cognitive skills (i.e., mental calculation, recognizing interrelationships and dependencies, skills regarding relevant subjects, such as chemistry and physics) on a 5 point Likert scale ranging from "very bad" to "very good". Most importantly, the questionnaire asked whether a strategy was

used to solve the ten test trials. If so, participants had to describe their strategy in an open format. We used this format to make sure that no strategies were accidentally suggested.

## 2.3 Design and Procedure

After signing the informed consent and the privacy agreement, participants completed the demographic questionnaire. Afterward, the eye tracker was calibrated using a five-point calibration. Gaze data recording started with the beginning of the instruction phase.

During the instruction phase, the application of thermal spray coating was explained. Using an example, participants were introduced into their task of adjusting the amount of hydrogen and argon that is inserted into the system to keep the temperature at an optimum. As this experiment followed a two factor between subjects design, half of the participants received information presented in a bar graph (unchunked condition; see Fig. 2a) and the other half received information presented in a coordinate system (chunked condition; see Fig. 2b).

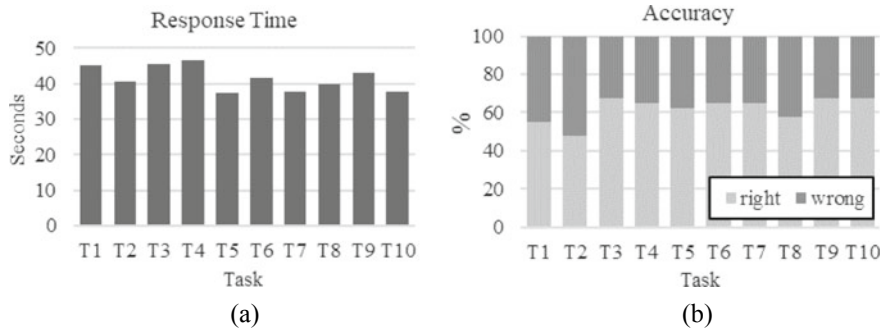
After reading the example, participants solved three practice trials with feedback (correct/incorrect). Participants were free to decide whether they wanted to repeat the explanations including the practice trials or not.

The test phase started with a new five-point calibration. Afterward, participants solved ten test trials without feedback. Gaze data, response accuracy and response time were recorded. Finally, participants answered the questionnaire including the question concerning the strategies used when solving the tasks.

## 3 Results

### 3.1 Analysis

To analyze participants gaze behavior in the instruction and test phase, we defined areas of interest (AOI) for all relevant areas in the stimuli (see Fig. 1). As the application was explained through a series of eleven screens, of which eight contained the graphic presentation, gaze data was summarized over these eight images. Fixation times were calculated as the sum of fixation times in a specific AOI. Fixation times were averaged over all screens of the instruction and test phase respectively. Response times were defined as the time from the start of task presentation until the answer was given via keyboard and was averaged over all trials. A repeated measures ANOVA revealed no differences in response times between tasks ( $F(351,9) = 1.82$ ,  $p = 0.06$ , and  $\eta_p^2 = 0.05$ , all Bonferroni pairwise comparisons  $p > 0.3$ ; see Fig. 3a). Response accuracy was calculated as a proportion of right answers for the ten test



**Fig. 3** Overview over response times (a) and task accuracy (b)

trials. A Cochran-Q-test revealed no differences in response accuracy between tasks ( $p = 0.54$ ; see Fig. 3b). Therefore, all test trials are comparable in their difficulty.

Twenty-one of the 40 participants repeated the instruction phase. There was no difference in accuracy ( $F(39,1) = 0.16, p = 0.69$ , and  $\eta_p^2 = 0.00$ ) and response time ( $F(39,1) = 0.230, p = 0.14$ , and  $\eta_p^2 = 0.06$ ) between those that did and those that did not repeat the instruction phase.

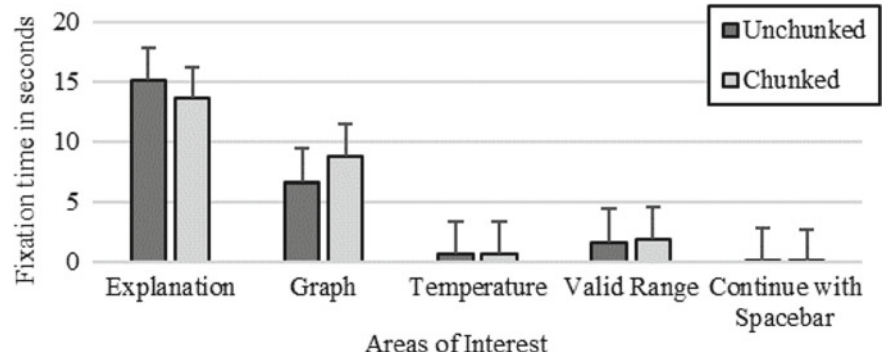
### 3.2 Inducing Chunking (H1)

To analyze whether visualizations displaying relations between variables (i.e., chunked condition) are more extensively used by participants for understanding than visualizations not displaying relations (unchunked condition), we first examined how much the participants used the graphic presentation compared to the written explanation. Figure 4 gives an overview over the aggregated fixation times to the AOIs of the stimulus material dependent on the format of the graphic presentation in the instruction phase.

The written explanation and the graph captured most of participants' attention ( $F(1, 38) = 60.00, p < 0.001$ , and  $\eta_p^2 = 0.61$ ). Over all AOIs, there are differences between the chunked and the unchunked condition ( $F(1, 38) = 0.10, p = 0.75$ , and  $\eta_p^2 = 0.003$ ). However, participants in the unchunked condition focused more on the explanation than those in the chunked condition. In turn, participants in the chunked condition looked more at the graph than those in the unchunked condition. This interaction was statistical meaningful ( $F(1, 38) = 4.74, p = 0.04$ , and  $\eta_p^2 = 0.11$ ), which supports H1a.

To analyze whether participants that learned the relations between variables in the instruction phase using visualizations displaying relations (chunked condition) tend to use these relations to solve subsequent tasks more often than participants using visualizations not displaying relations (unchunked condition), we asked participants about their strategies used for task solving. In total, 34 of the participants answered





**Fig. 4** Overview over fixation times in seconds to the elements of the stimuli material during the instruction phase. Error bars represent standard error

**Table 1** Overview over used strategies for each condition

	Strategy use			
	Yes	Strategy		No
Unchunked	17	Ratio	3	3
		Elimination	11	
		Other	4	
Chunked	17	Ratio	7	3
		Elimination	6	
		Other	4	

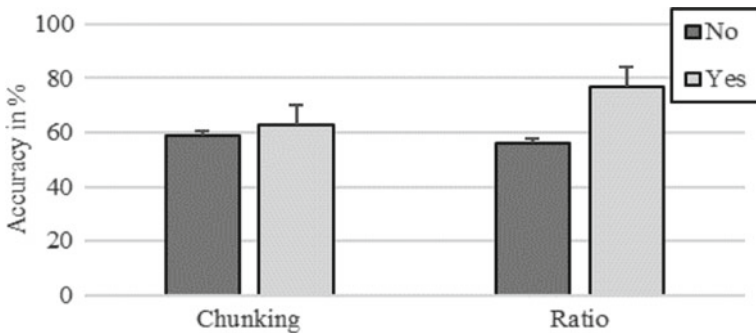
that they indeed used a strategy for task solving (90%, 17 in each condition, see Table 1). Out of the 34 participants that used a strategy, 17 reported that they eliminated wrong answers in any manner; most of them ( $N = 11$ ) were in the unchunked condition. Seven participants of the chunked condition stated that they used the actual ratio between the gases as a strategy, which is more than twice as much as in the unchunked condition ( $N = 3$ ). This is in line with H1b. However, statistical analysis failed to provide support for H1b ( $\chi^2 = 2.13$ ,  $p = 0.14$ , and  $\phi = 0.23$ ).

### 3.3 Performance Measures (H2)

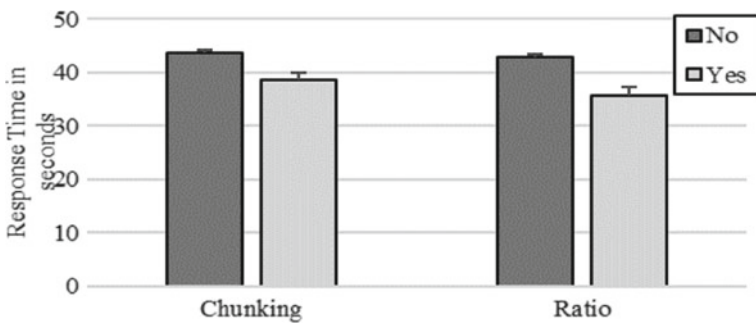
To analyze whether visualizations displaying relations between variables (i.e., chunked condition) lead to a deeper understanding than visualizations not displaying relations (unchunked condition), we used participants' mean accuracy rate in percentage (i.e., the percentage of correctly solved tasks; ACC) and participants' mean response time in seconds (i.e., the time for solving a task; RT). A Student's t-test revealed that, on average, participants in the chunked condition solved the test

trials with a higher accuracy and faster (see left side of Figs. 5 and 6) than those in the unchunked condition. However, these differences were not statistically significant ( $t_{ACC}(38) = -0.52, p = 0.61, d = -0.16$ ;  $t_{RT}(38) = 1.18, p = 0.25, d = 0.37$ ).

As it is possible that conceptual chunking was not successfully induced for all participants in the chunked condition, we further compared participants ( $N = 10$ ) that did consciously comprehend the relationship between the gases argon and hydrogen with those that did not ( $N = 30$ ) (see right side of Figs. 5 and 6). Analysis showed that the test tasks were more accurately solved, if the relationship was understood than if not ( $t_{ACC}(38) = -2.59, p = 0.01, d = -0.95$ ). However, participants that understood the relationship were not significantly faster than those that did not ( $t_{RT}(38) = 1.44, p = 0.16, d = 0.53$ ).



**Fig. 5** Participants’ response accuracy in the test trials. Bars are grouped according to the assignment to the chunking condition (left hand side; yes = chunking condition, no = unchunking condition) or the usage of the ratio as a strategy (right hand side; yes = the ratio was used, no = the ratio was not used). Error bars represent standard error



**Fig. 6** Participants’ response time in the test trials. Bars are grouped according to the assignment to the chunking condition (left hand side; yes = chunking condition, no = unchunking condition) or the usage of the ratio as a strategy (right hand side; yes = the ratio was used, no = the ratio was not used). Error bars represent standard error

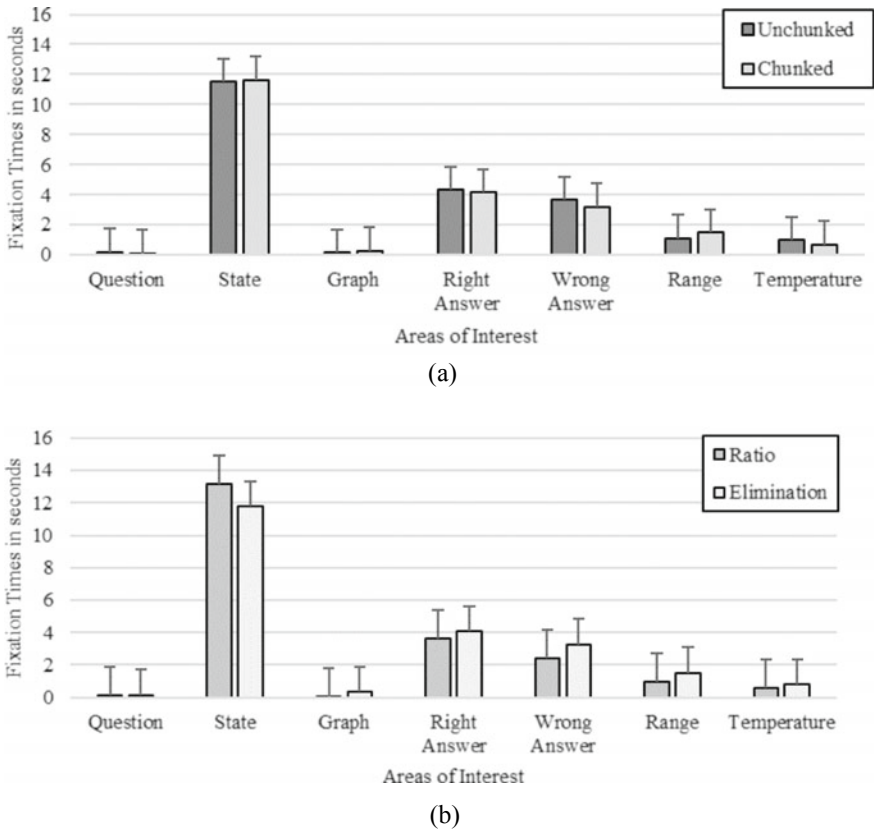
### 3.4 Gaze Data during the Test Phase

In Fig. 7a, summed fixation times to the relevant AOIs during the test phase dependent on the format of the graphic presentation are displayed. Participants in the unchunked condition looked significantly more often to the test question than participants in the chunked condition ( $t(38) = 1.72$ ,  $p = 0.047$ , and  $d = 0.54$ ). Moreover, descriptive data analysis yielded some interesting results. In both conditions, the current state of the system was the most important information ( $M_{\text{unchunked}} = 11.50$  s,  $SD = 5.50$ ;  $M_{\text{chunked}} = 11.63$  s, and  $SD = 6.23$ ). However, the placeholder, which only displayed the axis, but not the content, received only little attention by participants ( $M_{\text{unchunked}} = 0.17$  s,  $SD = 0.60$ ;  $M_{\text{chunked}} = 0.28$  s, and  $SD = 1.04$ ). In addition, the right answer alternative received, on average, more attention than each wrong answer alternative ( $M_{\text{unchunked}.r} = 4.36$  s,  $SD = 1.90$ ;  $M_{\text{chunked}.r} = 4.17$  s,  $SD = 2.08$ ;  $M_{\text{unchunked}.w} = 3.67$  s,  $SD = 1.64$ ; and  $M_{\text{chunked}.w} = 3.21$  s,  $SD = 1.69$ ). However, these differences were not statistically significant.

To draw a more comprehensive picture of information search and information processing during the test phase, we further contrasted the two most frequently used strategies for task solving. The strategy “ratio” means that participants consciously used the proportion of gases to adjust the state of the system. In that case, participants engaged in conceptual chunking. The strategy “elimination” means that participants eliminated wrong answer alternatives until only one alternative remained. As Fig. 7b shows, participants looked longer at the current state of the system, if the ratio between gases was used for task solving than if not. In contrast, participants that eliminated wrong answer alternatives focused more often on the alternatives themselves. Information, such as the valid range or optimal temperature, were more important when participants used elimination as a strategy, because those are elimination criteria. As more information have to be inspected, participants using elimination as a strategy looked significantly longer at the display of a given task before giving an answer than those using the ratio ( $M_{\text{ratio}} = 35.76$  s,  $SD = 10.15$ ;  $M_{\text{elimination}} = 43.24$  s,  $SD = 13.37$ ; and  $t(25) = -1.83$ ,  $p = 0.04$ ,  $d = -0.61$ ).

## 4 Discussion

Aim of the present study was to examine if conceptual chunking affects information processing when solving complex operating tasks. For this, graphic presentations (not) displaying relations between variables (i.e., bar graphs vs. coordinate systems) were used. Thus, conceptual chunking should be facilitated (chunked condition) or constrained (unchunked condition) during the process of understanding relations and interactions between variables.



**Fig. 7** Summed fixation times in seconds to the relevant AOIs during the test phase over conditions (a) and most frequently used strategies (b). Error bars represent standard error

### 4.1 Summary of the Results

For both conditions, gaze data showed that the written explanations were used most extensive. This might be due to the fact that reading information takes longer than encoding visual stimuli [25]. However, participants in the chunked condition used the graphic presentation more extensively than participants in the unchunked condition. This indicates that graphic presentations displaying relations between variables seem to be suitable to support the understanding of complex industrial tasks. In addition, participants in the chunked condition used the interrelation between variables (i.e., ratio between the gases) more than twice as often compared to participants in the unchunked condition, indicating that the coordinate systems support the understanding of complex industrial processes. However, this result was not statistically significant, which might be a consequence of our limited sample as discussed later.

Moreover, there was a tendency for higher response accuracy and lower response time when solving the tasks in the chunked condition compared to the unchunked condition. However, these effects also failed to reach statistical significance. When analyzed by strategy, a significantly higher accuracy for the participants that engaged in conceptual chunking was found compared to those that did not.

Further, we could show that conceptual chunking is an effective strategy to facilitate understanding while solving complex machining tasks. Participants, who engaged into conceptual chunking show much less information search behavior than those that used other strategies (such as evaluating wrong answer alternatives).

## ***4.2 Limitations and Further Research***

There were some limitations in the present study. First, due to the limitations resulting from the corona pandemic, our sample was rather small. Second, the results showed that there was a significant performance improvement once conceptual chunking was successfully induced. However, even though the coordinate system clearly supported the understanding of the relations between variables in our experiment (compared to the side-by-side bar chart), chunking could not be induced for all participants successfully by the coordinate system. This is evident by the fact that only seven out of 20 participants stated that they used the ratio between gases as a strategy. However, it is possible that more participants engaged in conceptual chunking, but did not report, because either they did not want to or they were not able to report (e.g., because it was not used consciously). Hence, the material should be optimized to induce conceptual chunking reliably.

Creating the best visualizations is an iterative process that grows with each study and the analysis of more fine-grained measures. As the coordinate system did not induce conceptual chunking very reliable (even though it was much better than the bar graph), other forms of visualizations have to be investigated. For instance, stacked bar graphs are an additional way to visualize relations, especially proportions, between variables [25].

## ***4.3 Implications***

The present research topic is enormously important for a human-centered development of Industry 4.0. An innovative integration of human cognition into CPSs that is based on a system perspective as well as complementarity and teaming principles will be fruitful for developing flexible situation- and application-oriented automation [33]. To date, many industrial systems are designed without considering the strengths and limitations of the human team partner. This research presents one way how knowledge about cognition can be used to shape teaming between humans and CPSs. This is done by an easier transfer of information into long-term memory using

visualizations. Thus, expert knowledge is gained, which leads to an improved reference frame for the processing of new information that also includes relations between variables. Thereby, this research focused on facilitating teaming by improving the human understanding of the technical “team partner”. The next step would be to also include processes of judgment and decision making into these considerations.

To conclude, the understanding of memory processes involved in monitoring and controlling industrial applications, the design of cognitively efficient visualizations and the utilization of suitable research methods (e.g., eye tracking) are small building blocks on the way to support human operators in their complex and challenging tasks in a responsible way.

**Acknowledgements** We thank Leonie Lude, Ben Herrmann, Elisabeth Chernenko and Chayenne Gläser for their contribution and help planning and executing the experiment as well as Chayenne Gläser for her help analyzing the data and Josef Schmidt for his help formatting the document. We also thank Thomas Lampke, Maximilian Grimm, Gerd Paczkowski, and Klaus Oberauer for constructive discussions. This research was partially funded by the research initiative “Instant teaming between humans and production systems” of Chemnitz University of Technology and co-financed by the Saxony State Ministry of Science and Art (grant SMWK3-7304/35/3-2021/48192).

## References

1. Miller CA, Parasuraman R (2007) Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control. *Hum Factors* 49(1):57–75. <https://doi.org/10.1518/001872007779598037>
2. Rasmussen J, Goodstein LP (1987) Decision support in supervisory control of high-risk industrial systems. *Automatica* 23(5):663–671. [https://doi.org/10.1016/0005-1098\(87\)90064-1](https://doi.org/10.1016/0005-1098(87)90064-1)
3. Madni AM, Madni CC (2018) Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. *Systems* 6(4):1–17. <https://doi.org/10.3390/systems6040044>
4. Johnson M, Bradshaw JM (2021) The role of interdependence in trust. *Trust Human-Robot Interact* 379–403. <https://doi.org/10.1016/B978-0-12-819472-0.00016-2>
5. Lee J, Moray N (1992) Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35(10):1243–1270. <https://doi.org/10.1080/00140139208967392>
6. Inagaki T (2003) Adaptive automation: sharing and trading of control. In: Hollnagel E (ed) *Handbook of cognitive task design*. CRC Press, pp 147–169
7. Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. *Hum Factors* 37(1):32–64. <https://doi.org/10.4324/9781315092898-13>
8. Tataschiere M, Bowden VK, Visser TAW, Michailovs SIC, Loft S (2020) The benefits and costs of low and high degree of automation. *Hum Factors* 62(6):874–896. <https://doi.org/10.1177/0018720819867181>
9. Wason PC, Green DW (1984) Reasoning and mental representation. *Q J Exp Psychol Sect A* 36:597–610. <https://doi.org/10.1080/14640748408402181>
10. von Salm-Hoogstraeten S, Müsseler J (2021) Human cognition in interaction with robots: taking the robot’s perspective into account. *Hum Factors* 63(8):1396–1407. <https://doi.org/10.1177/0018720820933764>

11. Cowan N (2017) The many faces of working memory and short-term storage. *Psychon Bull Rev* 24:1158–1170. <https://doi.org/10.3758/s13423-016-1191-6>
12. Oberauer K et al (2018) Benchmarks for models of short term and working memory. *Psychol Bull* 144(9):855
13. Baddeley AD, Hitch GJ (1974) Working memory. *Psychol Learn Motiv* 8:47–89
14. Johnson-Laird PN, Byrne RMJ, Schaeken W (1992) Propositional reasoning by model. *Psychol Rev* 99(3):418–439. <https://doi.org/10.1037/0033-295X.99.3.418>
15. Sweller J, van Merriënboer JJG, Paas F (2019) Cognitive architecture and instructional design: 20 years later. *Educ Psychol Rev* 31(2):261–292. <https://doi.org/10.1007/s10648-019-09465-5>
16. Akyürek EG, Kappellmann N, Volkert M, van Rijn H (2017) What you see is What you remember: Visual chunking by temporal integration enhances working memory. *J Cogn Neurosci* 29(12):2025–2036. <https://doi.org/10.1162/jocn>
17. Miller GA (1956) The magical number of seven, plus or minus two: some limit on our capacity for processing information. *Psychol Rev* 63(2):81–97. <https://doi.org/10.1177/001088049003100202>
18. Thalmann M, Souza AS, Oberauer K (2019) How does chunking help working memory? *J Exp Psychol Learn Mem Cogn* 45(1):37–55. <https://doi.org/10.1037/xlm0000578>
19. Cowan N (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav Brain Sci* 24(1):87–114. <https://doi.org/10.1017/S0140525X01003922>
20. Halford GS, Wilson WH, Phillips S (1998) Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behav Brain Sci* 21(6):803–864. <https://doi.org/10.1017/S0140525X98001769>
21. Nassar MR, Helmers JC, Michael JF (2017) Chunking as a rational strategy for lossy data compression in visual working memory. *Psychol Rev* 125(4):486–511. <https://doi.org/10.1037/rev0000101>
22. Killinger A, Gadow R (2021) Thermally sprayed materials for biomedical applications. In: Pomeroy M (ed) *Encyclopedia of materials: technical ceramics and glasses*. Elsevier, Amsterdam, pp 732–749
23. Markets and Markets (2021) Thermal spray coatings market by materials (ceramics and metals & alloys), process (combustion flame and electrical), end-use industry (aerospace, automotive, healthcare, agriculture, energy & power and electronics) and region - Globa. <https://www.marketsandmarkets.com/Market-Reports/thermal-spray-coating-market-181347083.html>. Accessed 06 Jul 2021
24. Halford GS, McCredden JE (1998) Cognitive science questions for cognitive development: the concepts of learning, analogy, and capacity. *Learn Instr* 8(4):289–308. [https://doi.org/10.1016/S0959-4752\(97\)00023-6](https://doi.org/10.1016/S0959-4752(97)00023-6)
25. Ware C (2021) *Information visualization*, 4th edn. Elsevier Inc., Cambridge, MA
26. Gillan DJ, Lewis R (1994) A componential model of human interaction with graphs: 1. linear regression modeling. *Hum Factors* 36(3):419–440. <https://doi.org/10.1177/001872089403600303>
27. Pinker S (1990) A Theory of Graph Comprehension. In: Freedle R (ed) *Artificial intelligence and the future of testing*. Taylor & Francis Inc., New York and London, pp 73–126
28. Richardson DC, Spivey MJ (2000) Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition* 76(3):269–295
29. Scholz A, Mehlhorn K, Krems JF (2016) Listen up, eye movements play a role in verbal memory retrieval. *Psychol Res* 80(1):149–158. <https://doi.org/10.1007/s00426-014-0639-4>
30. Ferreira F, Apel J, Henderson JM (2008) Taking a new look at looking at nothing. *Trends Cogn Sci* 12(11):405–410. <https://doi.org/10.1016/j.tics.2008.07.007>
31. Richardson DC, Altmann GTM, Spivey MJ, Hoover MA (2009) Much ado about eye movements to nothing: a response to Ferreira et al.: taking a new look at looking at nothing. *Trends Cogn Sci* 13(6):235–236. <https://doi.org/10.1016/j.tics.2009.02.006>
32. Scholz A, Mehlhorn K, Bocklisch F, Krems JF (2011) Looking at nothing diminishes with practice. In: *Proceedings of the 33rd annual conference of the cognitive science society*, pp 1070–1075

33. Bocklisch F, Paczkowski G, Zimmermann S, Lampke T (2022) Integrating human cognition in cyber-physical systems: a multidimensional fuzzy pattern model with application to thermal spraying. *J Manuf Syst* 63:162–176. <https://doi.org/10.1016/j.jmsy.2022.03.005>



# A Mobile Application Innovation for Public Healthcare Supply Chain Coordination



Marcia Mkansi  and Tshililo Ramovha

**Abstract** *Purpose* This study presents a mobile application (App) innovation that responds to a multi-embedded supply chain coordination problem linked to the unavailability, understock, and/or overstock in many public healthcare sectors in Africa. It also drives the transformation of the initial cost of investment for many information systems that are prohibitively high in developing countries, where the use of information technology infrastructure is still limited. *Research Approach* Using a design science methodology, a quantitative computer programming approach was used to incorporate the Internet of Things (IoT) model, barcode technology, Xamarin technology and the Microsoft dot net (.NET) framework. *Findings and Originality* The outcome of the experimental research strategy was a functional mobile application able to respond to the needs of the multi-embedded supply chain coordination problem related to artemisinin-based combination therapies (ACTs). *Research Impact* Theoretically, the innovation extends supply chain coordination frameworks beyond what might normally be expected of coordination theories and healthcare supply chain information systems. *Practical Impact* The mobile App equips small manufacturers and public healthcare centers, especially township and rural health centers, with the technology to improve efficiency in the deployment of drugs, to track supply across the value chain, and enables users of the system to effectively coordinate available supply to meet demands and react to outbreaks quickly. Subsequently, it limits the impact of infectious diseases to vulnerable groups.

**Keywords** Mobile application innovation · Public healthcare supply chain · Supply chain coordination · Design science

---

M. Mkansi (✉) · T. Ramovha  
University of South Africa, Pretoria 0002, Gauteng, Republic of South Africa  
e-mail: [mkansm@unisa.ac.za](mailto:mkansm@unisa.ac.za)

T. Ramovha  
e-mail: [38785374@mylife.unisa.ac.za](mailto:38785374@mylife.unisa.ac.za)

## 1 Introduction

The importance of access to good health and well-being, and to economic development can never be understated. It is the third highest sustainable development goal of the United Nations (UN), and is widely recognized by policy makers, professional funding bodies and scholars as one of the critical, interminable, and intractable global challenges that requires innovation and cohesive efforts from all community stakeholders [1, 2]. More so in Africa, where one of the greatest challenges for growth and development is the lack of national effort in the coordination of the industry, academic and corporations' contribution of a shared agenda toward the transformation of Africa, and expensive public services such as healthcare and education systems. The role of every segment of the society, including academics and industry, is invaluable, because governments budgets are shrinking, and non-government bodies are grappling with limited funds post-COVID-19 that has wreaked havoc on most African budgets and economies. Malaria can be taken as a default example, where the WHO [2] reports that sub-Saharan Africa accounts for 94% of world's malaria deaths, of which 78% of all deaths are children under five years old. Recognizing the transformational power of innovation and social-embedded solutions in creating social value for people of all age groups is a global, the African Union (AU) agenda 2063 and UN priority [3].

## 2 Background Research for the Mobile Application

Nagitta and Mkansi [4] addressed the theoretical problem of the multi-embedded supply chain coordination of artemisinin-based combination therapy (ACT). Their findings revealed several dimensions related to the micro, market and macro dimensions that are deemed critical to the availability of ACTs. The study by Nagitta and Mkansi [4] also revealed critical logistics dimensions that need to be addressed. According to several scholars [4, 5], the lack of the relevant Information and Communication Technology (ICT) systems, such as an appropriate decision-support system, has amplified the problem of the multi-embedded supply chain coordination of ACTs, and will require a huge investment to solve.

Although mobile applications (Apps) are available in some areas, they do not cover the end-to-end processes of the supply chain coordination of ACT drugs [6]. For example, the Apps are not integrated, and some of the processes are done manually [7]. Some examples of mobile application innovations for the healthcare supply chain are Pilldrop (enables patients to register as users and motorists to register as providers); iWander App (discreet monitoring device for caregiver to remotely monitor dementia patients); and Remote Area Medical (the use of drones for medication delivery to address the medical needs for underserved communities).

However, none of these innovations address the challenge of multi-embedded healthcare supply chain coordination. Hence, there is substantial potential for the

mobile application innovation for healthcare supply chain coordination to bridge the digital and healthcare gap. Healthcare inclusion and the availability of stock can help avert most outbreaks, build patients' trust, and result in a health workforce that will benefit all stakeholders and the economy. The mobile application innovation not only addresses Sustainable Developmental Growth (SDG) goal 3, but also offers fresh insight in terms of how academia can transform research into innovation that can help to shape the supply chain.

### 3 Design Methodology

Design science is described as the creation of innovative technological artifacts that broaden the capabilities of humans and organizations, alike [8, 9]. Literature provides a class of development methodologies that follows many approaches, including but not limited to, waterfall (systematic), iterative, spiral (lifecycle oriented), V-shaped (controlled focused), and agile (highly adaptive) [10–12]. While the latter scholars focus on different aspects of design science, the common emphasis is on the stages of the seven design methods, regardless of the development methodology employed, namely: problem relevance (concept design from research to *artifact*), design as a search process (critical evaluation of the suite of technologies to address the problem), design as an artifact (coding, development, and prototyping), design evaluation (testing and observation of the artifact), research rigor (accuracy and precision in line with research), research contribution (application to theory, practice and methods), and lastly, communication (publication of the app and research) which were implemented in the current study as discussed below.

#### 3.1 Problem Relevance

From the perspective of problem relevance, the study followed Bormane and Bērziša's [13] Business Analysis Book of Knowledge (BABOK) that offers a road map for stakeholder engagement to ensure that stakeholders agree on the requirements, system architecture and design for satisfactory system implementation. The market and macro environments are part of the major management environment, as discussed in detail by previous studies [14, 15]. Using the waterfall methodology, the requirements phase revealed the requirements that inform the technological features relevant to the current study, which were: a dashboard that allows the user to view and analyze the stock status and to respond to demand.

At the micro-environment level, employees from the various healthcare centers, namely, the district hospitals, referral hospitals, district medical offices and health centers, receive ACT drugs from either private pharmacies, national medical stores, or donors. The ACT drugs are scanned using the sensor devices that are in use at the distribution center, before the drugs are dispatched to the hospitals and health

**Table 1** Mapping of IoT model with high-level technological features

IoT layer	Environment	Features
Application	Micro	Dashboard, web portal to track and trace ACT drugs, dispensing of ACT drugs to patients, SMS notification
	Macro	Dashboard to view and analyze stock status, view the demand of drugs
	Market	Dashboard, view the demand of drugs
Network	Network infrastructure	
Sensor	Micro	Interface to scan ACT drugs when dispensing to patient and web portal for administration
	Market	Interface to scan ACT drugs when distributing to district office, hospital, and so on

centers. The App will automatically update the system and reflect on the dashboard. A similar principle will apply when an employee dispenses ACT drugs to patients, it will automatically inform patients by Short Message Service (SMS) (Table 1).

**3.2 Design as a Search Process**

In the design as a search process stage, the study followed Tătaru and Fleacă’s [16] guidance to elicit requirements from key stakeholders, and then to transform the requirements into functional tasks using the internet of things (IoT framework). Figure 1 shows the conceptual technological interpretation of the critical supply chain coordination across the micro, market, macro, and logistical activities, as implemented in the current study using the IoT model. For example, in the application layer of IoT, the study considered the web portal, dashboard, application programming interface (API), event processing and analytics. The App is to be linked to the sensing layer through the network layer’s infrastructure that includes the IoT, such as mobile phones, scanners and QR codes. Together, the three layers serve as the foundation, advanced as the conceptual models of analysis for evaluating a class of technologies across the different layers of IoT, developing process flows, use cases and inputs for the next stage of design as an artifact. Hailes [17] defines requirements analysis as the task that structures and organizes the requirements collected during the elicitation activities in the form of models.

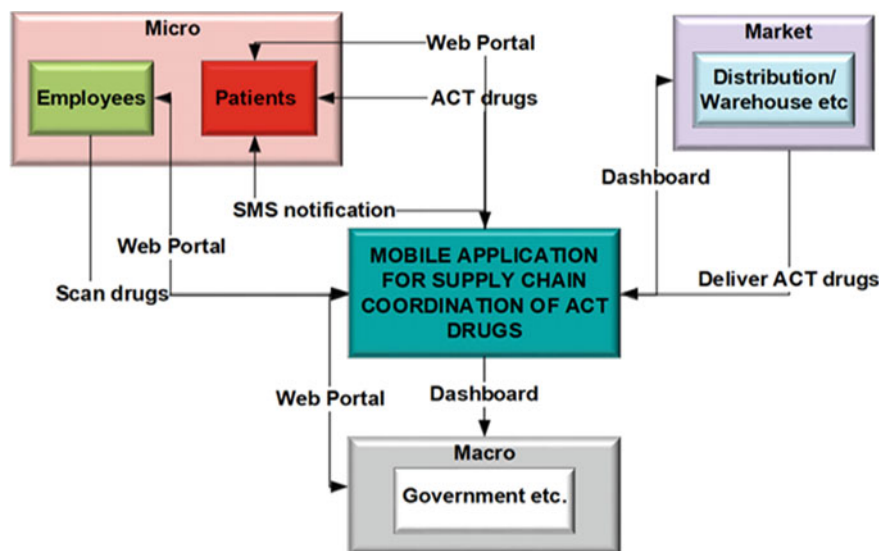


Fig. 1 Main functionality and stakeholders

### 3.3 Design as an Artifact

The design as an artifact stage deals with both design and implementation using unified modified modeling languages. Booch et al. [18] provide two common software design methods, the first being: structured design, defined as a method to convert requirements specifications, such as Data Flow Diagrams (DFD), into a system flowchart that can be implemented using a programming language; and the second being object-oriented design (OOD), defined as a mechanism that encompasses modularity, abstraction, and encapsulation as three important concepts in software design.

The study followed OOD because of its natural modular design structure that easily allows changes to be made to it without affecting the other modules. In this study, the modules that are integral to the object-oriented programming are application programming interfaces (using C# language), and mobile application (using the Xamarin platform, dot net, and C#). Although scholars and programmers rarely use OOD within a waterfall model [19], the integration of both structured waterfall principles and OOD programming was critical and possible in this study. The OOD programming is crucial for the flexible adjustment of the sensing and application layers which had several objects; hence, the waterfall model was necessary for the supply chain coordination flow of the micro, macro, and market stakeholders. The inputs of the design included logical design, physical design, system architecture and the entity relational diagram necessary to produce the desired output, design, and user documents.

The output from the requirements, analysis and design stages were used as input to develop the unified user interface, mobile application screens, Database Management System (DBMS) and coding of the app. The technical specifications were also used to build a cloud environment where the application will be running. The outputs of this stage are the different sub-systems and the DBMS.

## 4 Findings

As this study is based on development, the findings are the actual mobile application (App) that was developed, and which serves as the extension of the design methods that cover design evaluation (testing and observation of the artifact), and research rigor (accuracy and precision in line with research). The incorporation of the first five stages produced the App which is accessible on the Google Play Store and App Store under the title “UNISA ACT”.

Some of the core features that were deduced through problem relevance, design as a process, and design artifact using the waterfall methodology produced the unified User Interface (UI) displayed in Fig. 2. The app is encapsulated inside the unified UI and is made to look as if it is part of one system [20]. This is accomplished by hosting the ACT mobile app in a Xamarin shell flyout. According to Wulf and Blohm [21], the unified UI uses responsive web design principles to provide an optimal viewing and interaction experience for any screen size, device, or orientation.

In the post-implementation and integration stage, the entire system was tested for any faults and failures, prior to being deployed to the cloud environment or released. According to Nidhra and Dondeti [22], there are two fundamental methods for testing software behavior and performance, namely, white box testing and black box testing. Acharya and Pandya [23] define black box testing as a method that focuses on testing the requirements or specifications of the software entity under test. Khan and Khan [24] explained that white box testing uses the coding experience as part of the testing procedure.

This study used white box testing to test whether the system had been programmed according to the functional requirements, whether all the functionalities were covered, and whether the programs were handling input errors appropriately. The testing included: unit, integration, functional, system, usability, performance, and security testing. The application performed to the optimal level across computer experimentation. It is currently undergoing pilot testing in the real environment by African Applied Chemical (a small enterprise producing malaria products in South Africa).



Fig. 2 ACT mobile app high level

## 5 Discussions and Conclusions

The current study adopted the principles and technologies of IoT to develop a mobile solution that will encompass all the stakeholders. The IoT technology will be able to provide real-time monitoring of activities, such as the delivery of stock, stock levels and consumption levels. The current study extended the findings of Nagitta and Mkansi [4] by developing a mobile application, which was the main research objective of the study. It completes the last two stages of the design methodology

by outlining the research contribution (application to theory, practice, and methods), and lastly, communication (publication of the app and research).

### ***5.1 Practical Contribution***

In practice, the mobile application improves the processes of coordinating the supply and distribution of ACTs and addresses the issue of the non-availability of ACTs to patients who are in need, by developing an integrated mobile application solution that provides real-time monitoring of the stock delivery, stock level and consumption level at each health facility. It presents the value proposition of all-in-one centralized mobile innovation features by using unique features built into the supply chain principles of supply chain coordination management. At best, it provides an alternative to the high-cost information systems that are highly unattainable for most of the public health and Small and Medium Enterprises (SMEs) that have a tight budget.

Providing digital power to independent supplier chemists in rural areas will ultimately increase digital social inclusion, employment, and digital poverty reduction. Access to medicine will ultimately improve the well-being of society and ensure a healthier workforce. Economically, public services such as health care are very expensive [5, 25, 26]. The mobile application innovation technology is increasingly enabling and driving the transformation of these expensive services. The role and finances of governments are shrinking, and NGOs are struggling to raise funds, some of the social problems cannot be addressed without the participation of every segment of society, including corporations, business professionals and academics. The affordability of ACT technology for use by both suppliers of all sizes and governments in developing economies is an example of how the academic segment is contributing toward the transformation of Africa and the healthcare system. Most importantly, it is expected that the system will create a job for the developer of the system, and for the data analyst who will support the different market segments.

### ***5.2 Methodological Contribution***

This study serves as one of the blueprints of how to convert research findings into software design. It further provides a context and evidence of how OOD can complement the structured principles of the waterfall model in the IoT context because scholars and programmers rarely use OOD within a waterfall model [19]. This study developed the mobile application using the API to create independence between the mobile application and the database. The mobile application was developed using C# and other related object-oriented programming languages.



### 5.3 Theoretical Contribution

One of the critical challenges for African growth is to contribute to the development of a shared agenda for the transformation system of Higher Education and to re-position the institutions in this sector to play a more meaningful role in the transformation of Africa. The mobile healthcare supply chain coordination innovation for ACTs contributes to national and global universal health coverage (Goal 3 of the United Nations' Sustainable Development) Priority 1: Economic transformation and job creation, and Priority 7: A better Africa and world.

Socially, ACTs' innovation presents opportunities to create social value and to contribute to SDG goal 3 that aims to ensure the health and well-being of people of all age groups and to ensure better living standards. The expression of interest by the Minister of Health in Uganda and the small supplier, African Applied Chemical, in South Africa, is evidence of the social impact.

## References

1. Oluka NP, Mkansi M (2019) Exploring the supply chain coordination dimensions for artemisinin-based combination therapies in Uganda. *Int J Supply Chain Manage* 8(4):134–151
2. World Health Organization (2017) World malaria report 2017. World Health Organization. <https://apps.who.int/iris/handle/10665/259492>
3. African Union (2022) Goals & priority areas of agenda 2063. <https://au.int/en/agenda2063/goals>. Accessed 21 May 2022
4. Nagitta OP, Mkansi M (2019) Exploring the supply chain coordination dimensions for artemisinin-based combination therapies in Uganda. *Int J Supply Chain Manage* 8(4):134–151
5. Stanley EF, Cynthia W, Chad A, Gregory M (2009) Supply chain information sharing: benchmarking a proven path. *Benchmarking: Int J* 16(2):222–246
6. Mpimbaza A, Miles M, Sserwanga A, Kigozi R, Wanzira H, Rubahika D, Nasr S, Kapella BK, Yoon SS, Chang M, Yeka A (2015) Comparison of routine health management information system versus enhanced inpatient malaria surveillance for estimating the burden of malaria among children admitted to four hospitals in Uganda. *Am J Trop Med Hyg* 92(1):18–21
7. Khurana S, Chhillar N, Kumar V, Gautam S (2013) Inventory control techniques in medical stores of a tertiary care neuropsychiatry hospital in Delhi. *Health* 5(1):8–13
8. Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. *MIS Q* 28(1):75–105. <https://doi.org/10.2307/25148625>
9. Nunamaker JF, Chen M, Purdin TDM, Journal S, Systems I, Support M, Taylor P, Nunamaker JAYF et al (2016) Systems development in information. *Syst Res* 7(3):89–106
10. Afif AN, Noviyanto F, Sunardi S, Akbar SA, Aribowo E (2020) Integrated application for automatic schedule-based distribution and monitoring of irrigation by applying the waterfall model process. *Bull Electr Eng Inform* 9(1):420–426
11. Almeida F, Simões J (2019) Moving from waterfall to agile: perspectives from IT Portuguese companies. *Int J Serv Sci Manage Eng Technol* 10(1):30–43
12. Kramer M (2018) Best practices in systems development lifecycle: an analyses based on the waterfall model. *Rev Bus Finan Stud* 9(1):77–84
13. Bormane L, Bērziša S (2017) Role of 'bridge person' in software development projects. In: International conference on information and software technologies, pp 3–14, 12–14 Oct 2017, Druskininkai, Lithuania. Springer, Cham

14. Nieman G, Bennett A (2002) Business management a value chain approach. Van Schaik Publishers, Pretoria
15. Fahey L, Narayanan VK (1986) Macro environmental analysis for strategic management (The west series in strategic management). West Publishing Company, St. Paul, Minnesota
16. Tătaru IM, Fleacă E (2019) Technologies for modeling business processes. *FAIMA Bus Manage J* 7(2):31–41
17. Hailes J (2014) Business analysis based on BABOK® guide version 2—a pocket guide. Van Haren, VW's-Hertogenbosch, Netherlands
18. Booch G, Maksimchuk RA, Engle MW, Young BJ, Connallen J, Houston KA (2008) Object-oriented analysis and design with applications. *ACM SIGSOFT Softw Eng Notes* 33(5):29–29
19. Brindhya J, Vijayakumar V (2015) Analytical comparison of waterfall model and object-oriented methodology in software engineering. *Adv Nat Appl Sci* 9(12):7–11
20. Carter P, Mulliner C, Lindorfer M, Robertson W, Kirda E (2016) Curious droid: automated user interface interaction for android application analysis sandboxes. In: International conference on financial cryptography and data security, pp 231–249. Springer, Berlin, Heidelberg
21. Wulf J, Blohm I (2020) Fostering value creation with digital platforms: a unified theory of the application programming interface design. *J Manag Inf Syst* 37(1):251–281
22. Nidhra S, Dondeti J (2012) Black box and white box testing techniques-a literature review. *Int J Embed Syst Appl* 2(2):29–50
23. Acharya S, Pandya V (2012) Bridge between black box and white box-gray box testing technique. *Int J Electron Comput Sci Eng* 2(1):175–185
24. Khan ME, Khan F (2012) A comparative study of white box, black box and grey box testing techniques. *Int J Adv Comput Sci Appl* 3(6)
25. Rose J, Persson JS, Heeager LT, Irani Z (2015) Managing e-government: value positions and relationships. *Inf Syst J* 25(5):531–571
26. Pan G, Pan SL, Newman M, Flynn D (2006) Escalation and de-escalation of commitment: a commitment transformation analysis of an e-government project. *Inf Syst J* 16(1):3–21

# A Social Critical Analysis on Philippine Higher Education in the Time of COVID-19 Pandemic Toward a Framework on Flexible Learning



Alvin A. Sario, Elcid A. Serrano, and Ramon L. Rodriguez 

**Abstract** The global pandemic significantly affects the higher education institution's delivery of its mandate to educational stakeholders. The emergency shift from face-to-face to flexible modality causes different university problems. The need to evaluate the status quo of the delivery of education in the Philippines is essential to understand the present scenario and propose a framework that is anchored on a pedagogical perspective and educational philosophy. Several proposed models and theories were reviewed as a basis for the proposed framework, and the research employs social critical analysis as the method of the study. The study presents the status of flexible learning in the Philippines, its challenges, and possible recommendations based on the development education principle. The proposed framework on flexible learning intends to exemplify strategies for efficiency and effectiveness of the delivery considering the context of the Philippine higher education system, institutions, teachers, students, and other academic stakeholders.

**Keywords** Education technology · Flexible learning framework · Educational philosophy · Social critical methods

## 1 Introduction

The school system was greatly affected by the COVID-19 pandemic across the world. Educational institutions around the world modified the way they do their education programs and activities to respond to the various protocols and systems implemented

---

A. A. Sario  
University of Santo Tomas, Legazpi, Philippines

E. A. Serrano  
Mapua University, Manila, Philippines

R. L. Rodriguez (✉)  
National University, Manila, Philippines  
e-mail: [rlrodriguez@national-u.edu.ph](mailto:rlrodriguez@national-u.edu.ph)

to protect the public and still promote education as a social good [1]. Schools worldwide pushed through with the new school year, given that we have not yet reached herd immunity to counter and stop the spread of the SARS-COV2 [2]. The Philippines did the same. The private schools started in August 2020 and public schools in October 2020 to address requirements given the challenges of online learning and module learning. The Philippines, as a developing country, adopts flexible learning for all Filipino students [3]. With the shift to online learning and module learning, Philippine education faces a considerable challenge that is fully immersed in traditional modalities [4]. There is a need to create a framework for online learning in Philippine education during the COVID-19 pandemic based on educational praxis and philosophies of education [5]. The study is deemed significant because it provides an alternative perspective on Philippine education in consideration of various philosophies of education given the context of the COVID-19 pandemic [6, 7]. In effect, we are challenged to create a framework to address the said concerns [8].

The study intends to create a framework for online learning in Philippine higher education during the COVID-19 pandemic. There are three objectives of the study: first, to determine the status of flexible learning in Philippine education in the time of the COVID-19 pandemic; second, to provide challenges given the status; and third, to articulate principles to guide in addressing issues about flexible learning in our higher education system.

## 2 Theoretical Underpinnings and Conceptual Framework

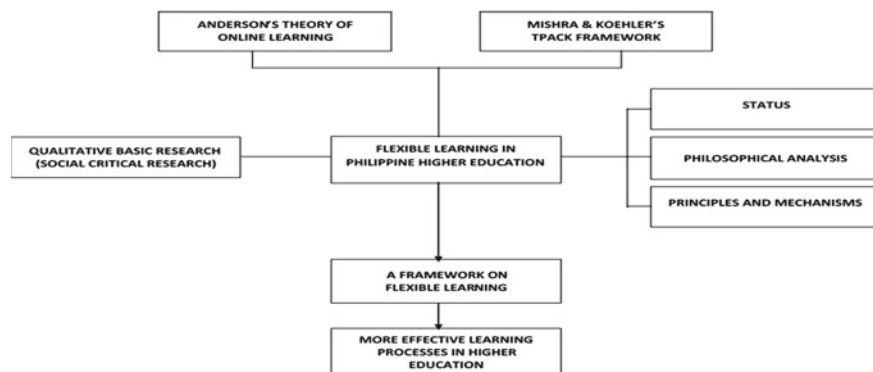
The study considers two perspectives to have a good analysis of flexible learning during this time of pandemic in the Philippine context. These are Terry Anderson's Theory of Online Learning and Punya Mishra and Matthew J. Koehler's TPACK Framework. The former provides the necessary dynamic interplay between and among teachers, students, and content to generate solid interfaces between and among them, resulting in collaborative learning, independent study, structured learning resources, and a community of inquiry. The latter provides the three necessary elements for successful online/flexible learning: content, pedagogy, and technology producing pedagogical content learning, technological content learning, and technological pedagogical knowledge leading to the core principle of technological pedagogical content knowledge [9]. It is fundamentally a model of e-learning demonstrating the interactions of learners and teachers relative to the knowledge/content. For Anderson, there are four overlapping components in effective learning environments. These are community-centered learning, knowledge-centered learning, learner-centered learning, and assessment-centered learning. Online learning communities can and should share a sense of belonging, trust, an expectation of learning, and commitment to participate in and contribute to such communities; therefore, students work together online to create new knowledge collaboratively.

Furthermore, whether the context is online or school-based, the internet provides vast information that provides opportunities to gather these resources. Therefore, learning becomes making connections with ideas, facts, people, and communities. It should not lose the premise that school stakeholders must support students strongly. The teacher, school, and society must create successful learner-centered environments and, in effect, create meaningful connections to the communities. With varying modalities, formative and summative evaluation have to be part of the entire schema to ensure that targets and competencies are reached, narrowing the gap between the traditional and alternative modalities. With these four lenses, education must recognize interactions' role in online learning. There has to be student-student interaction (which is collaborative learning), student-content interaction (which consists of interactive knowledge and customization of content), and student-teacher interaction (which defines the relationship between them in actual teaching and learning experience). The model, in general, is a multi-component learning environment that enhances the critical interaction in education in many creative ways. The Anderson Model is crucial in the paper for it highlights four components that underscore Philippine education contexts.

The TPACK model is an integration framework of technological, pedagogical, and content knowledge [10]. As a framework, it affirms three primary forms of human knowledge: content knowledge, pedagogical knowledge, and technological knowledge. This TPACK Model provides us with a complete understanding of teaching with technology. Content, pedagogy, and technology have to blend to provide a better learning experience for students. It centers on digital pedagogies focused on cooperative learning and the realization of desired learning outcomes. The TPACK model is significant to the paper for it explains the interaction of content, pedagogy, and technology.

The study on flexible learning in the Philippine higher education system is anchored on two relevant contemporary teaching and learning theories and flexible/online/distance education: Anderson's Theory of Online Learning and Mishra and Koehler's TPACK Framework. It focuses on determining the status of such flexible education in the country, subjecting the status to some philosophical analysis, and articulating principles and mechanisms that would result in a framework for flexible learning for Philippine higher education institutions. This is done through qualitative basic research using the social critical research approach. Figure 1 shows the conceptual framework model anchored on the theory of online learning and the TPACK framework as the basis for the proposed flexible learning framework to have a more effective learning process in higher education in the Philippines.

Some models have been proposed concerning students' learning activities involving technology in higher education. With the recent global crises, the technology enable model is essential to ensure the continuity of education everywhere. The contextual facilitators for learning activities involving technology or the Cb-model [11] emphasize student learning activities that are directly and causally related to student outcomes and systematize contextual facilitators for learning activities involving technology in higher education. The adoption of digital technologies in the proposed framework is crucial since digital technologies have been part of our



**Fig. 1** Conceptual framework model

daily lives, especially during COVID-19, in the teaching and learning process [12] and are widely integrated with the pedagogical process. Every higher education institution's consideration and capability to integrate digital technologies into classroom teaching have numerous benefits, but the institutions must address high expectations from shareholders [13]. There is a need to consider the diverse opportunities arising from digitalization to support the teaching and learning process of the stakeholders [14, 15]. The disruption of education in the Philippines during the pandemic highlights the vulnerability of every institution and its readiness to implement digital learning. As seen in the research, the crucial role of digital technologies for teaching and learning is the effective modality utilized by schools [15, 16] for students and teachers. It recognizes the point that technology can benefit student learning but, at the same time, can also be detrimental to the educational process [17]. The proposed framework for HEIs in the Philippines seeks to influence technology integration and assess the adoption of digitalization in the universities to further enhance students' learning process through a flexible learning framework.

### 3 Methodology

The study is basic research. Specifically, it is qualitative research as observed in philosophy and education. The study employs a social critical research strategy as an approach to the entire study. The data gathering tools and techniques employed are document and secondary data analysis. It uses qualitative data analysis (QDA) as the primary analysis tool. The advocacy lens utilized is educational praxis.

## 4 Discussions and Analysis

### 4.1 *Status*

The COVID-19 pandemic significantly challenges the Philippine education system. It compromises the entire system that relies on the face-to-face medium of instruction. Though across countries, face-to-face approach and style is still the most effective modality given principles of facilitating, coaching, and mentoring in teaching and learning the current situation makes it not efficient. Schools have to find ways to deliver quality instruction given the health crisis. The Philippines, as predominantly traditional in pedagogy, has to design an educational framework in the time of this educational crisis due to the COVID-19 pandemic. We need to present the status of our higher education system.

#### **Level of Readiness of the Philippine Education System**

Before the pandemic, the Philippine education system is on the right course. Gradually, our higher education institutions are gaining ground internationally. Many of our schools have already been part of the top schools in Southeast Asia, the whole of Asia, and the world. With the implementation of outcomes-based education coupled with inquiry-based (or problem-based) education, the Philippine curricula indeed prepare the young to be part of the nation's social capital, participate in broad spaces, and immerse in the national/international market economy. The way we articulate competencies composed of knowledge, skills, and behaviors to guarantee quality standards in higher education is remarkable. We have our differentiation as educational quality is given national and international standards.

During the pandemic, the Philippine system of education is heavily challenged. Given an almost traditional system, where teaching and learning occur in a physical classroom setting, the quality of education is compromised. The flexibility required for education, given the health crisis, is not immediately met. The whole system of education grappled for the best forms of education given health protocols and social distancing requirements. The public system of education fully recourse to modular distance learning. The logistics involved are not clear. The private schools found ways, given the pandemic, and employed flexible learning. The way it is done varies from one private school to the next, so this results in quality compromise for quality are not guaranteed. The whole system, in effect, leads to a digital divide and further discrimination in access and quality of education. Guidelines are not immediately prepared. Schools devised ways to adjust, cope, and adapt given the COVID-19 pandemic. Every school launched its flagship program for education at the time of the pandemic. Content, pedagogy, and technology are run, reviewed, and revised to suit various contexts. Given the health crisis, we had difficulty adapting to change since our educational system is OBE-focused but not flexible, creative, and innovative enough for a situation like this. It is still traditional in the framework. The alternative system or educational approach/ strategy for a crisis/pandemic situation is not in place. There was a need to review everything: curricula, syllabi, and modules

in content, pedagogy, and technology. Various delivery modes of learning systems must be incorporated as educational policy mandatory for all higher education institutions.

### **Level of Compliance of Philippine Higher Education Institutions**

The current curricula, syllabi, and modules for the degree programs are not appropriate for an alternative system in the case of a crisis or pandemic. Infrastructures have to be built in the institutions. The first thing to do is review the curricula, revise the syllabi, and modify/reconstruct modules. The outcomes-based education framework depends on the curriculum. If the curriculum is not fit for a pandemic context, then the quality of education is compromised. One crucial issue here is how we determine the essential competencies to be discussed in online learning and how all other competencies are devised and guaranteed in modular learning, given the presumption that blended learning means online education and modular learning. The curriculum then speaks of the competence of our higher education. Educational quality begins in the curriculum, is interpreted in the syllabus, and is implemented in the modules. The second crucial issue is how content in blended learning is delivered using pedagogical approaches using technology systems and applications. Every curriculum must show the dynamic interaction of content, pedagogy, and technology in online learning. The quality we create in implementing content, pedagogy, and technology defines the level of competence in higher education. In effect, curricula, syllabi, and modules are designed not only for the traditional style of teaching and learning. Infrastructures should be mandatory for all higher education institutions, especially the necessity for information and communication technology.

### **Level of Skills of Higher Education Teachers**

Since Philippine (higher) education is largely traditional in teaching and learning, teachers are not capacitated for technological enrichment to reckon content and pedagogy. In this time of the pandemic, education has to be technological. All teachers have to be technologically ready. It seems impossible now in the time of pandemic to teach without technology. Faculty development programs have to give priority to training in the use of educational technology. In higher education institutions, educational technology centers play a significant role in running online learning. Capacity-building programs for teachers to enrich their pedagogical styles and their innovativeness in using technology-based applications have to be the primary focus. Creativity and innovation in education are seen in the dynamic interplay of content, pedagogy, and technology.

### **Level of Practicability for Higher Education Students**

There are challenges Philippine higher education institutions are facing during this time of the pandemic. Learning becomes not workable given reasonable factors: strength of internet connectivity, quality of modules provided, the validity of the assessment, availability of learning devices such as smartphones, availability of budget for net data subscription, conduciveness of respective households for learning, et cetera. There should be standardized, normativized, and effective modules that



the government and higher education institutions produce. Requirements must be simple, reasonable, and workable given the students' time, resources, and capacity. Assessments should not be traditional and objective but should be authentic assessments. The digital divide should never be promoted directly or indirectly, and Internet connectivity must be public. The default system for all degree programs has to be the modules. The synchronous and/or asynchronous sessions should be made available to students but not mandatory; schools and teachers have to monitor every student's learning progress effectively.

### **Level of Adjustment of Higher Education Stakeholder (Parent, Community, LGU's, PO's)**

Parents are not equipped to assist their children learning, who are already tertiary students. The demand for a higher study budget, especially for the requirements, is difficult. Uncertainties of the community in the quality of education provided cannot be avoided. The nature of intervention of local government units and people's organizations cannot be undermined and overestimated. There should be new but rationalized guidelines for the fees of the schools, colleges, and universities. Even in private schools, students have to be subsidized by the government. All stakeholders should play respective roles to ensure quality education, guaranteeing educational access during this pandemic.

## **4.2 Challenges**

Given the status of the Philippine higher education system at the time of the COVID-19 pandemic, we have seen the need for reframing the higher education system, reformulating national and disciplinary learning outcomes for higher education, and revisiting curricula of programs, refocusing higher education to educational technology, and recasting involvement of stakeholders. These are the challenges the Philippine higher education needs to face with urgency: (1). There is a need to reframe the entire Philippine higher education system, (2). There is a need to reformulate learning outcomes, (3). There is a need to revisit the curricula of all degree programs, (4). There is a need to refocus higher education to educational technology, and (5). There is a need to recast the involvement of stakeholders in education.

These five challenges are identified to capture the action the country needs to take on education. They require a collective effort of various stakeholders and demand collaboration from government agencies, non-government organizations, local government units, people's organizations, higher education institutions, schools, parents, and students.

### 4.3 Principles

A framework on flexible learning is intended to be a framework that would exemplify strategies to realize the efficiency and effectiveness of flexible learning in consideration of the considered contexts of the Philippine higher education system, institutions, teachers, students, and stakeholders ultimately. It attempts to highlight flexible learning in the Philippine education system as a suitable response for schools, colleges, and universities during the COVID-19 pandemic. The core conception of the framework is educational praxis [5]. Educational praxis as a model of development education is a philosophy of education. It showcases the status of our higher education and the corresponding challenges it poses to the whole higher education system. It then extrapolates on the measures that need to be done: capacity-building of administrators and teachers, content mastery of various faculties, pedagogy-focused educational programming, considered the best practices in different settings, highly effective modules, and technology-enriched syllabi and modules and their implementation. These measures make educational praxis an educational philosophy of action. Such a philosophy of action articulates criteriology, efficiency, and normativity in the conception and application of flexible learning. There is a need to exemplify some principles for, such criteriology, efficiency, and normativity (Fig. 2).

#### The Educational Philosophy of Education in Community-Based and Context Relevant

Experiential learning is finding and creating meaning from experience, i.e., direct experience. It focuses on the learning process of and for the individual. It requires personal values such as self-initiative and self-evaluation. It is about creating an

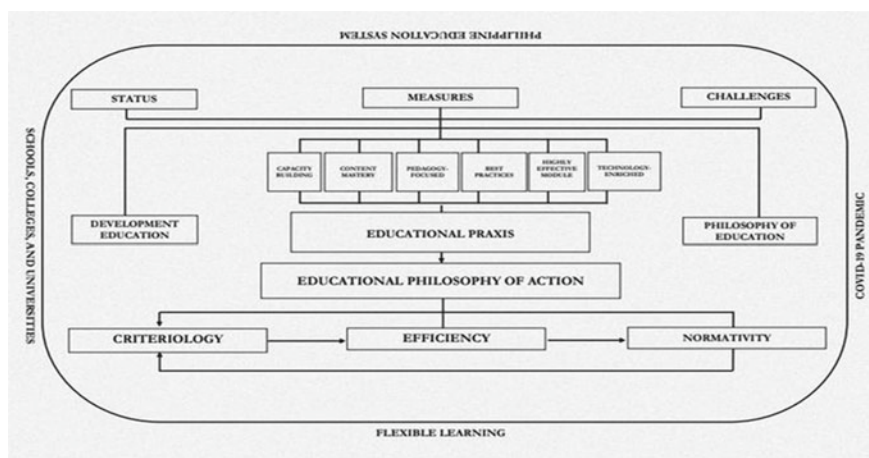


Fig. 2 Framework on flexible learning in Philippine higher education

experience where learning can be facilitated. Its dimensions are analysis, initiative, and immersion. Generally, experiential learning is learning through experience. Specifically, it centers on the elements of experiencing, reflecting, and applying. The COVID-19 Pandemic as a communal experience is an excellent source of experiential learning, given the various concepts and principles in various subjects, courses, and programs. Progressive education rests on the premise that humans are social beings who learn best in real-life activities with other people. Education is shared in this sense. The entire educative process of experiencing, sharing, processing, generalizing, and applying primarily occurs within the social context of a given dynamic human interaction. Progressive education emphasizes problem-solving and critical thinking, group work and development of social skills, collaborative and cooperative learning, education for social responsibility and democracy, integration of community service in the instruction, and experiential learning. This pandemic is a challenge to have social learning in the midst of social distancing. As a progressive education, it is called to be innovative and creative in social learning and social distancing.

### **The Education Philosophy of Action is a Form of Critical Pedagogy**

The discussion in Principle 1 is focused on access to relevant education. Principle 2 is centered on the quality of education. This pandemic, however, real, should not compromise educational quality. There is only a difference in modality, but the competencies remain the same. The proposed framework highlights critical pedagogy. This critical pedagogy bears a perspective that is collective and is based on consensus and community development. The so-called community-based and context-relevant education fulfill social transformation. Such education responds to the development needs still of the community. This pedagogy has to contribute to community development.

### **Educational Philosophy of Action is Anchored on Experiential Learning and Progressive Education**

Experiential learning is finding and creating meaning from experience, i.e., direct experience. It focuses on the learning process of and for the individual. It requires personal values such as self-initiative and self-evaluation. It is about creating an experience where learning can be facilitated. Its dimensions are analysis, initiative, and immersion. Generally, experiential learning is learning through experience. Specifically, it centers on the elements of experiencing, reflecting, and applying. The COVID-19 Pandemic as a communal experience is an excellent source of experiential learning, given the various concepts and principles in various subjects, courses, and programs. Progressive education rests on the premise that humans are social beings who learn best in real-life activities with other people. Education is shared in this sense. The entire educative process of experiencing, sharing, processing, generalizing, and applying primarily occurs within the social context of a given dynamic human interaction. Progressive education emphasizes problem-solving and critical thinking, group work and development of social skills, collaborative and cooperative learning, education for social responsibility and democracy, integration of community service in the instruction, and experiential learning. This pandemic is a challenge

to have social learning in the midst of social distancing. As a progressive education, it is called to be innovative and creative in social learning and social distancing.

### **The Education Philosophy of Action is a Measure of Quality Standards in Education**

Quality standards cannot be sacrificed. The pandemic is not an excuse to compromise knowledge and skills. We only have different modalities, but the goal remains the same. COVID-19 pandemic does not change the reality that we need graduates that are competitive, skilled, and ready as social capital for the national economy and social progress. The Philippine economy is in crisis. After the pandemic, the best way to recover is to have excellent human resources to fuel the economic flow. The values schools should promote for this purpose are relevance, academic atmosphere, institutional management, sustainability, adaptation, and efficiency. Productivity, especially of people, cannot be sacrificed. We need to have programs programmed to develop communities; that is, re-engineering education delivering and extending education to all.

### **The Educational Philosophy of Action is Political**

Education in the challenging times is still intended to establish and enhance democratic culture in a liberal democratic constitutional regime. It is fundamentally constructed for the transformation of communities. It is based on political principles of democracy, citizenship, and participation. It sees democracy as the most viable social and political system but tries to challenge that system to improve the system. The transformation can be realized when it is rested on fundamental rights and liberties, cooperative political virtues, social cooperation and shared responsibility, reflective equilibrium, public reason, public political culture, and objective political dialogue.

## **5 Conclusions**

Education, development education for that matter, is called to become idealist in principles and dynamism but at the same time must be a realist in considering socio-economic, socio-cultural, and socio-political contexts that we have. Nonetheless, it must not demean being pragmatist, for our situation calls for praxis. This redounds to the idea of holistic education. The idea of integrative wholeness in education speaks of considering the various aspects and approaches in education into a unified dynamic educational system. It is challenged to become constructivist in terms of framing and constituting aims, content, practice, and outcomes in development education. The ultimate guiding principle is that it should be an education that empowers and transforms. The educational philosophy of action in time of the COVID-19 pandemic encompasses the impact of development education on community, democracy, and citizenship. It has to be a critical pedagogy toward community development. It has

to be anchored on experiential learning and, at the time, progressive education. It has to continue to be a measure of quality standards. It has to strengthen our sense of democracy continually.

## References

1. United Nations Policy Brief: Education during COVID-19 and beyond (2020) Retrieved from [https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/08/sg\\_policy\\_brief\\_covid-19\\_and\\_education\\_august\\_2020.pdf](https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/08/sg_policy_brief_covid-19_and_education_august_2020.pdf)
2. Reimers FM (2020) Organization for Economic Cooperation and Development (2020) A framework to guide an education response to the COVID-19 pandemic of 2020. Retrieved from [https://globaled.gse.harvard.edu/files/geii/files/framework\\_guide\\_v2.pdf](https://globaled.gse.harvard.edu/files/geii/files/framework_guide_v2.pdf)
3. Cuaton G (2020) Philippines Higher Education Institutions in the time of COVID-19 Pandemic. *Revista Romaneasca Pentru Educatie Multidimensionala*, 12(1Sup2): 61–70. <https://doi.org/10.18662/rrem/12.1sup2/247>
4. DepEd Order 012 s (2020) Adoption of the basic education learning continuity plan for school year 2020–2021 in light of the covid-19 public health emergency. [https://www.deped.gov.ph/wp-content/uploads/2020/06/DO\\_s2020\\_012-1.pdf](https://www.deped.gov.ph/wp-content/uploads/2020/06/DO_s2020_012-1.pdf)
5. Sario A (2011) Educational praxis in development education. ISBN: 978-3-659-17281-6
6. Gingell J (2008) Philosophy of education: key concepts. <https://www.library/main/E15D22EF4D37067B9AB0A461FBBF0E247>
7. Toquero CM (2020) Challenges and opportunities for higher education amid the covid-19 pandemic: the Philippine context. [https://www.researchgate.net/publication/340680378\\_Challenges\\_and\\_Opportunities\\_for\\_Higher\\_Education\\_amid\\_the\\_COVID-19\\_Pandemic\\_The\\_Philippine\\_Context](https://www.researchgate.net/publication/340680378_Challenges_and_Opportunities_for_Higher_Education_amid_the_COVID-19_Pandemic_The_Philippine_Context)
8. International Association of Universities (2020) Regional/national perspectives on the impact of covid-19 on higher education. [https://www.iau-aiu.net/IMG/pdf/iau\\_covid-19\\_regional\\_perspectives\\_on\\_the\\_impact\\_of\\_covid-19\\_on\\_he\\_july\\_2020\\_.pdf](https://www.iau-aiu.net/IMG/pdf/iau_covid-19_regional_perspectives_on_the_impact_of_covid-19_on_he_july_2020_.pdf)
9. Anderson T (2008) Anderson's theory of online learning. Towards a theory of online learning. Retrieved from <http://www.aupress.ca/index.php/books/120146>. Accessed on 12 Aug 2021
10. Mishra P, Koehler MJ (2006) Technological pedagogical content knowledge: a framework for integrating technology in teacher's knowledge. *Teach Coll Rec* 108(6):1017–1054
11. Sailer M, Schultz-Pernice F, Fischer F (2021) Contextual facilitators for learning activities involving technology in higher education: the Cb-model. *Comput Hum Behav* 121:106794. ISSN: 0747-5632. <https://doi.org/10.1016/j.chb.2021.106794>. <https://www.sciencedirect.com/science/article/pii/S0747563221001175>
12. Seufert S, Guggemos J, Sailer M (2021) Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: the current situation and emerging trends. *Comput Hum Behav* 115:106552. <https://doi.org/10.1016/j.chb.2020.106552>
13. Schmid R, Petko D (2019) Does the use of educational technology in personalized learning environments correlate with self-reported digital skills and beliefs of secondary-school students? *Comput Educ* 136:75–86. ISSN: 0360-1315. <https://doi.org/10.1016/j.compedu.2019.03.006>. <https://www.sciencedirect.com/science/article/pii/S0360131519300648>
14. Castillo-Manzano JI, Castro-Nuño M, López-Valpuesta L, Sanz-Díaz MT, Yñiguez R (2016) Measuring the effect of ARS on academic performance: a global meta-analysis. *Comput Educ* 96:109–121. <https://doi.org/10.1016/j.compedu.2016.02.007>

15. Sailer M, Murböck J, Fischer F (2021) Digital learning in schools: what does it take beyond digital technology? *Teach Teach Educ* 103:103346. ISSN: 0742-051X. <https://doi.org/10.1016/j.tate.2021.103346>. <https://www.sciencedirect.com/science/article/pii/S0742051X21000706>
16. Seufert S, Guggemos J, Sailer M (2021) Technology-related knowledge, skills, and attitudes of pre- and in-service teachers: the current situation and emerging trends. *Comput Hum Behav* 115:106552. ISSN: 0747-5632. <https://doi.org/10.1016/j.chb.2020.106552>. <https://www.sciencedirect.com/science/article/pii/S0747563220303022>
17. Carstens KJ, Mallon JM, Bataineh M, Al-Bataineh A (2021) Effects of technology on student learning. *Turk Online J Educ Technol* 20(1)

# M-HEALTH System for Detecting COVID-19 in Chest X-Rays Using Deep Learning and Data Security Approaches



Johnny Delgado, Luis Clavijo, Carlos Soria, Juan Ortega,  
and Sebastian Quevedo

**Abstract** Advances in predicting different types of pathologies in medical images have been significant in the last decade, thanks to the performance and efficiency of models trained with deep learning approaches. In this context, the prediction of the COVID-19 disease in chest X-Rays has been no exception. However, the proposed models are not always put into production and those that do not consider data security a system requirement. In this work, we propose creating a mobile application for detecting COVID-19 disease in chest X-rays using deep learning and data security approaches. Our prediction model has a sensitivity of 92% and a specificity of 90%. Our application implements the OAuth 2.0 access delegation standard for system access authorization.

**Keywords** Deep learning · X-ray · Security · Software · COVID-19

## 1 Introduction

Automating the detection of COVID-19 disease in chest X-rays, which would at least match the effectiveness of practicing radiologists, could be beneficial for large-scale screening and health initiatives, considering that as of July 04, 2022, about 550,045,535 cases have been reported worldwide for this disease [1]. Advances made in deep learning in the last decade have enabled algorithms such as deep neural networks to match the performance of medical professionals in a wide variety of tasks such as cancer detection [2], pneumonia [3], and Alzheimer's [4] have enabled

---

J. Delgado · L. Clavijo · C. Soria · J. Ortega · S. Quevedo (✉)

Universidad Católica de Cuenca, Cuenca, Ecuador

e-mail: [asquevedos@ucacue.edu.ec](mailto:asquevedos@ucacue.edu.ec)

URL: <https://www.ucacue.edu.ec/>

S. Quevedo

Electrical and Computer Science Engineering Department, Escuela Superior Politécnica del Litoral—ESPOL University, Campus Gustavo Galindo, km. 30.5 Vía Perimetral, Guayaquil, Ecuador

studies in other branches involving medical imaging. In this context, COVID-19 detection has generated significant interest in the deep learning research community, and several solutions have been proposed [5–7].

To generate models to automate the detection of COVID-19 disease in chest radiographs, it is essential to have reliable datasets, given the sensitivity of this process. In this context, the scientific community has published many datasets that can be used for training models to identify the presence of COVID-19 disease in chest radiographs, the ones chosen for our work are the following:

- COVID-19 Radiography Database, a dataset resting on Kaggle servers [8, 9] consists of 3616 images labeled with COVID-19, and 10192 images labeled Normal
- BIMCV Full PADCHEST dataset [10] consists of 160861 images, with COVID-19 distributed in 52 zip files.
- COVID-19\_Pneumonia\_Normal\_Chest\_Xray\_PA\_Dataset “NOTA de pie 11” is a dataset resting on Kaggle servers and is organized in 3 folders (COVID, pneumonia, normal) containing posteroanterior (PA) chest X-ray images. A total of 6939 samples were used in the experiment, where 2313 samples were used for each case.
- COVID-19 + PNEUMONIA + NORMAL Chest X-Ray Image Dataset [11, 12], which is a medical image directory structure divided into three subfolders (COVID, NORMAL, PNEUMONIA) containing the chest X-ray (CXR) images.
- COVID: 1626 images, NORMAL: 1802 images, PNEUMONIA: 1800 images.

In this scenario, we can highlight that the training model that allows the detection of COVID-19 disease is feasible. However, we have reviewed the literature, and few studies have been able to bring these models to a production information system [13–15]. Along these lines, we have considered that a system of these characteristics must comply with information security parameters that allow the preservation of the confidentiality of documents containing health data as established by the regulations of confidential information of international health systems [16].

Considering these premises, we propose creating a Web-based information system that allows sending images of chest X-rays and performs an inference process to detect the presence of COVID-19 in chest X-rays using reliable security protocols. In addition, we have applied an OWAS zed attack proxy (ZAP) methodology “Nota de Pie 19”, which is an integrated penetration testing tool to find vulnerabilities in Web applications to test the effectiveness in the security of our system and thus ensure the confidentiality and security of the information.

The rest of the article is organized as follows: Sect. 2 related work, Sect. 3 materials and methods, Sect. 4. We discuss the results, and finally Sect. 5, we present our conclusions.



## 2 Related Work

For this research, we will rely on models that have been most successful in analyzing, classifying, and detecting objects in computer vision and information systems that have been put into production. In this context, experimental results on a set of available chest X-ray and CT data show that the DenseNet121 architecture has managed to have the best performance and highest classification accuracy among the main convolutional neural network architectures such as MobileNet, DenseNet, Xception, ResNet, InceptionV3, InceptionResNetV2, VGGNet, and NASNet [17].

Since COVID-19 disease was first detected, several models have been developed for its detection from chest X-rays, and so we consider proposals for prediction models with specific architectures based on CNN, but there are only a few in which information systems have been developed in production that has allowed users to access and use these trained models. In the following, we mention the most significant works:

- COVID-XR [13]: Open-access Web Management System that allows the use of a pre-trained VGG16 model to classify an image into COVID and non-COVID.
- COVID-19 Classifier [14]: It is a free Web service for fast COVID-19 detection in chest X-ray images using deep learning. Two deep learning models are presented, one for differentiating between X-ray and non-X-ray images based on the MobileNet architecture and another for detecting chest X-ray images with COVID-19 features using a dense block-based network and initialized from pre-trained ImageNet.
- COVID-19 Detection [15]: From a prediction model with CNN DenseNet architecture, a Web application was developed for general users to detect chest X-ray images as either COVID or normal. A GUI application was run for the COVID prediction framework. Medical personnel or the general public can examine and enter a chest X-ray image into the program.

All the investigated works implemented an inference system using CNN bedside architectures. However, they do not use as reference architecture the DenseNet family121, which, as we have already demonstrated, is state of the art in chest X-ray image classification.

Of the models analyzed, only one visualizes diagnostic results with Grad-CAM [18] based on chest radiographs. In addition, the information systems do not consider any security in system access and data processing. With these considerations, our work includes the following:

- Design and implement a prediction model employing a DenseNet121 architecture<sup>1</sup> that allows diagnosing the presence of COVID-19 by inference from chest X-ray images.

---

<sup>1</sup> <https://aws.amazon.com/es/lambda/>.

- Implement a Web information system using microservices that enable the process of inferencing a chest X-ray image to present prediction results and a Grad-CAM image [18].
- Security implementation using the OAuth2.0 framework for user authentication, authorization of access to the system, and protection of confidentiality in the processing and data flow with Spring security.
- The information system passes vulnerability control audits with the OWAS ZAP tool.

### 3 Materials and Methods

For the development of this proposal, it was necessary to execute four macro processes: The first corresponds to the creation of the prediction model; the second, the validation of the model; the third, the creation of the information system; and the fourth, the implementation of security and confidentiality of the data and results.

In creating the prediction model for COVID-19, it is necessary to collect information from chest X-ray images and, with these images, train the deep learning model whose algorithms were coded using the Python programming language. In collecting radiographic images of patients with COVID-19, the documentary analysis type data collection technique will be used to obtain data from primary sources and focused interviews on obtaining reliable data.

#### 3.1 Model

The creation of the inference model to detect COVID-19 disease consisted of 6 stages; namely: The first corresponds to the data preparation and development environment; the second is the initialization of the model hyperparameters and metrics; the third, the increase of data in the set and training to use them in the training phase; the fourth lies in continuing iterating in the previous steps until it is considered that the model learned to detect the disease. Subsequently, we proceed to evaluate the results obtained on the model performance, and if they are satisfactory, we enter the validation stage; otherwise, we must adjust the parameters.

**Data Preparation.** The datasets used are available on Kaggle, a subsidiary of Google LLC, an online community that allows users to find and publish datasets, and explore and create models in a Web-based data science environment.

BIMCV (El Banco de Imágenes Médicas de la Comunidad Valenciana) is an image bank conceived as a knowledge repository designed to provoke technological advances in medical imaging and provide technological coverage services to support D+I projects.

The source images are a total of 29,958 images, 13,506 with normal label and 16,452 with COVID label, a 90% division of the data is applied, and two datasets are obtained, one with images for training that has 12,155 with normal label and 14806 with COVID label; and the other with images for validation with 1351 with normal label and 1646 with COVID label.

All images we converted to grayscale since some of them had marks or even colored arrows introduced by the healthcare personnel. Next, we normalized the pixel value to the range  $[0, 1]$ . After this, we proceed to resize the radiographs so that they all have a size of  $224 \times 224$  pixels since neural networks require the inputs to have the same shape.

**Hyperparameters and Metrics.** Hyperparameters are adjustable parameters chosen to train a model and govern the training process itself. In the preparation of our model, we consider the following parameters in which the necessary adjustments are made to obtain the best prediction performance:

Hyperparameters: Adam, binary\_crossentropy.

Metrics: Accuracy, Precision, Recall, Sensitivity, Specificity, AUC-ROC.

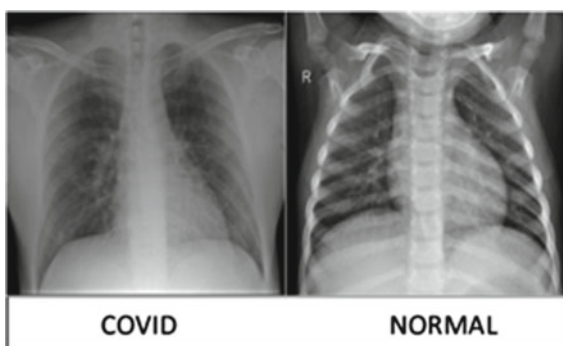
**Training Process.** The training and testing process of the architecture proposed for this work were built using Python and the machine learning framework Tensorflow and Keras.

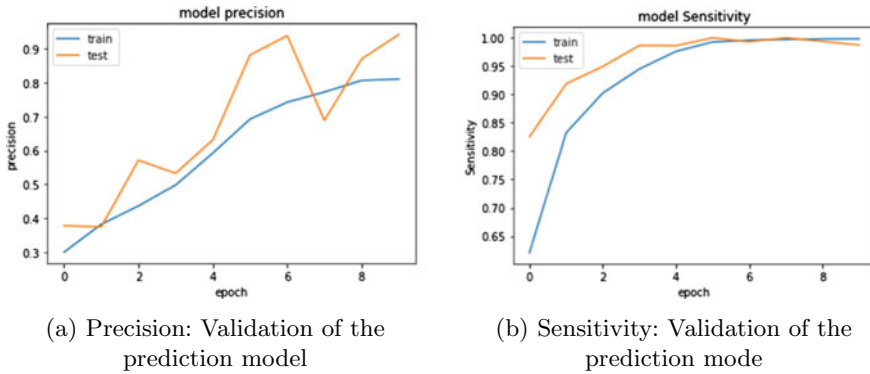
**Data Augmentation.** For data augmentation, we implement: random width shift (200, 200), height shift (0.5), random rotations ( $90^\circ$ ), brightness changes (0.2, 1.0), zoom (0.5, 1.0). Product of the data increase we obtain 26,961 images belonging to 2 classes and in validation 2997 images belonging to 2 classes.

**Fine Tuning.** We implement a fine-tuning on the DenseNet121 architecture by applying transfer learning with the pre-trained weights on the ImageNet dataset, and we modify the last layer of the model to do the binary classification with a sigmoid activation that encodes the probability that the analyzed image belongs to class 1, in this case (COVID-19). See Fig. 1.

**Validation.** The accuracy value or degree of success of the DETECTCOVID-19 training model gives us a percentage of 98.8% efficacy in the prediction. The metrics

**Fig. 1** COVID and normal chest X-ray





**Fig. 2** Precision an sensitivity validation

evaluated are precision, recall, sensitivity, and specificity. See Fig. 2. Later, we can verify the model's performance from new data using a test dataset that was not part of the training dataset. We will use datasets provided by Ecuadorian health entities to validate the model, using the methodology proposed by [18]. There is a board of expert raters examining the validation dataset. A simple majority decision (an image is classified as referable [COVID-19] if  $\geq 50\%$  of the experts rate it as referable). Raters are anonymized to the judgments of other raters.

## 4 System Architecture

For the deployment of the prediction system, represented in Fig. 3, the architecture presents us with three modules that are: first, the front end, second the back end, and the third refers to the prediction model, which interacts, allowing the sending of the images of chest X-ray, its respective processing and that in the end will return a result with a legend if the patient has COVID-19 or not, and an image based on Grad-CAM [19] product of the prediction made by the trained model.

### 4.1 The Front End

We created a welcome Web interface that allows user registration in the first instance with fields such as id, name, surname, institution, address, telephone, user email, and password, which are stored in a PostgreSQL database through the backend. Additionally, it is required to activate the account by the system administrator. After its activation, it is possible to enter securely using the OAuth2.0 authorization framework. The login interface requires the email and password to be validated in the backend and authorize access to the system's APIs through a token established by the authorization server. See Fig. 4.

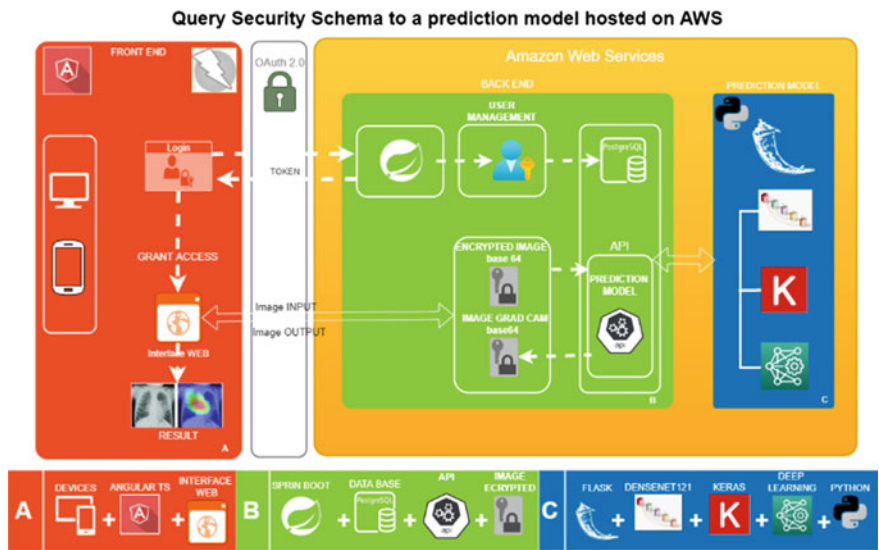


Fig. 3 System architecture

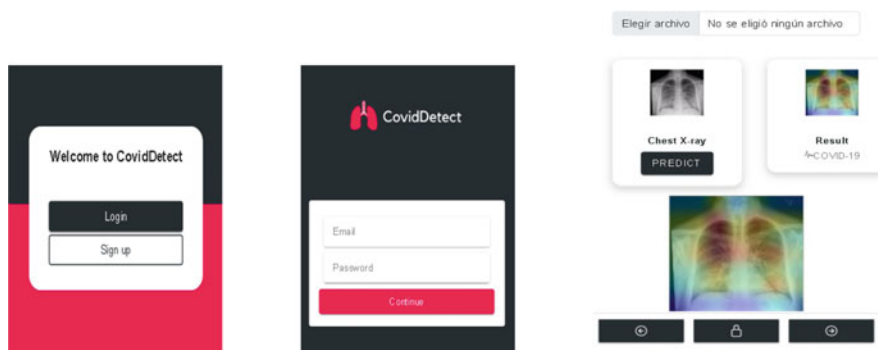


Fig. 4 GUI M-HEALTH system for detecting COVID-19

## 4.2 The Back End

An application based on microservices was implemented, which is created using a domain-based model software construction paradigm. The persistence layer implements an ORM object-relational mapping that considers the proposed system's data structure, a service layer to implement the business logic, and the controller layer where the REST APIs of the system are exposed. A security system based on the OAuth 2.0 security framework was implemented to access the exposed APIs. The front end connects with the APIs implemented so that the requirements set for this

**Fig. 5** Gradient-weighted  
class activation mapping  
final prediction



system is guaranteed. The security system allows entry to the system through a user authentication that will return an authentication token and then, from the front-end, carry out the inference process of the chest X-ray image that will be encoded in base64 in the second instance to be sent. To the prediction, APIs evaluate it and return the result that consists of the encrypted image. Finally, the prediction of the system allows for visualizing the results using the Grad-CAM gradient-weighted class activation visualization (see Fig. 5).

### ***4.3 Prediction Model***

We use Keras as a software library to create the inference model for the detection of COVID-19. We replicate a deep convolutional neural network of the DenseNet architecture family with 121 convolutional layers. Through a fine-tuning process, we adjust the network's last layer to program it to detect the presence of COVID-19. The model is displayed on the Web using the Flask framework that provides an API that receives an image as input and performs the inference process. The service is mounted in the Amazon AWS cloud and connects to the system's back end with SHA-256 encryption.

## 5 Security and Confidentiality

Securities were implemented using the Spring Security framework,<sup>2</sup> which is a framework to authenticate and control access to applications. In our system, we will apply it to control access to APIs that provide support for OAuth and OAuth2.

### 5.1 Spring Security Configuration

The security methods mainly verify the user's identity through authentication and the roles associated with it through authorization. The authorization is dependent on the authentication since it occurs after its process. Security in Spring manages three levels.

1. Filter chain bean
2. Password encryption bean
3. Security handler bean.

For the login process in the information system, Spring calls the user manager and performs a search for the user id in our database. In this case, it is the user email (Fig. 7), and to validate the password, it does a match as follows:

- The password entered is to be validated and encrypted with a Bcrypt algorithm, see Fig. 8.
- Subsequently, the passwords are compared with the one stored in the database.
- If the passwords match, the request is considered valid.

The effective way to establish security in storing passwords in the database is for the administrator to add a unique character or a secret word to the password entered by the user to be registered. After this, we would avoid vulnerability to brute force and dictionary attacks if encrypted and saved.

The UserDetailsService method fetches the complete user details from the database for reporting purposes authentication.

The BCryptPasswordEncoder method in Spring Security uses SHA-256 to encrypt the password. The SHA series is a hash algorithm but not an encryption one, so it can be decrypted. To apply security in the system in the Email fields of username and password, we perform encryption through the Bcrypt encryption algorithm (see Fig. 6).

---

<sup>2</sup> <https://spring.io/projects/spring-security>.

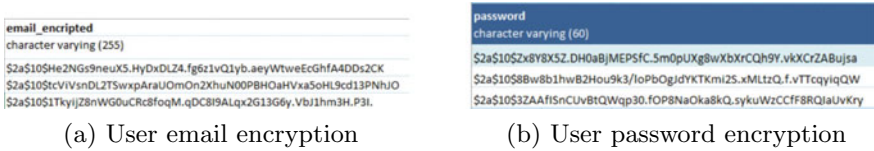


Fig. 6 User and password encryption

Table 1 Evaluación de calidad de un modelo de predicción de COVID-19 quality assessment of a COVID-19 prediction model

Method	Precision	Sensitivity	Specificity	ROC
Densenet121	94.89%	92.00%	90.00%	0.89

## 6 Results and Discussions

### 6.1 Model

The prediction model training was based on a DenseNet121 architecture for the diagnosis of COVID-19 disease. Using the confusion matrices, we can evaluate and know the measures of the prediction model and the ROC curve obtaining accuracy results shown in Table 1. For our model, we use the metrics: accuracy, sensitivity, and specificity, as well as the area under the ROC curve.

Given the data presented in Table 1, high percentages are shown in each of the parameters, which demonstrates a degree of effectiveness and reliability in the predictions, making this model viable for professional use.

### 6.2 Security and Confidentiality

Connection tests were carried out between the front and backend, and it has not been possible to establish communication because if the user and the key that validates it are not known, it is not possible to access the APIs. Several security methods were implemented both in the front and back end to strengthen vulnerabilities in the information system and thus allow access under authentication and authorization with OAuth2.0 and data encryption algorithms of registered users to prevent theft. Also, as API authentication is based on JWT, CSFR protection is not required. On the other hand, if the system is exposed to attacks such as session fixation, clickjacking, and forgery of requests between sites, the spring security configurations protect us from these attacks, which guarantees the integrity and confidentiality required of the data in the flow of information when using the services.



Alert type	Alert Tag OWASP	CWE ID	WASC ID
CSP: Wildcard Directive	OWASP_2021_A05 OWASP_2017_A06	693	15
Content Security Policy (CSP) Header Not Set	OWASP_2021_A05 OWASP_2017_A06	693	15
Cross-Domain Decouffiguration	OWASP_2021_A01 OWASP_2017_A05	264	14
Missing Anti-clickjacking Header	OWASP_2021_A05 WSTG-w47-CLNT-09 OWASP_2017_A06	1031	15
The server discloses information using an HTTP response header field(s) "X-Powered-By"	OWASP_2021_A01 WSTG-w42-INFO-08 OWASP_2017_A03	200	13
X-Content-Type-Options Header Missing (1)	OWASP_2021_A01 OWASP_2017_A06	693	15
Timestamp Disclosure - Unix	OWASP_2021_A01 OWASP_2017_A03	200	13
Disclosure of Information - Suspicious Comments	OWASP_2017_A03	200	13

(a) Alerts for risk and trust

Alert type	Solution
CSP: Wildcard Directive	Ensure that your web server, application server, load balancer, etc. is properly configured to set the Content Security Policy header.
Content Security Policy (CSP) Header Not Set	Ensure that your web server, application server, load balancer, etc. is configured to set the Content Security Policy header, to achieve optimal browser support. "Content Security Policy" for Chrome (2s, Prefers 23s and Safari 7s), "X-Content-Security-Policy" for Firefox 4.0s and Internet Explorer 10s, and "X-WebKit-CSP" for Chrome 10s and Safari 6s.
Cross-Domain Decouffiguration	Limit the set of addresses, for example, Configure the HTTP header "Access-Control-Allow-Origin" to a more restrictive set of domains, or completely remove all CORS headers, to allow the web browser to enforce the same source policy (SOP) in a more restrictive manner.
Missing Anti-clickjacking Header	Modern Web browsers support the Content Security Policy and X-Frame-Options HTTP headers. Ensure one of them is set on all web pages returned by your app/ops.
The server discloses information using an HTTP response header field(s) "X-Powered-By"	Make sure your web server, application server, load balancer, etc. is configured to suppress X-Powered-By headers.
X-Content-Type-Options Header Missing (1)	Ensure that the application's web server sets the Content Type header appropriately, and that it sets the X-Content-Type-Options header to "nosniff" for all web pages.
Timestamp Disclosure - Unix	Manually verify that timestamp data is not sensitive, and that data cannot be aggregated into exploitable patterns of disclosure.
Disclosure of Information - Suspicious Comments	Remove all comments that return information that could help an attacker and the any underlying issues they refer to.

(b) Alerts by site and risk

Fig. 7 Alerts

### 6.3 OWAS Zed Attack Proxy (ZAP)

Using the OWAS ZAP tool and making the necessary configurations so that it can act with a man-in-the-middle attack technique, an audit is carried out on the COVID-19 prediction information system, attacking the target IP, where the system for penetration tests resides virtually, for which the following levels of risk and trust are considered: Risk levels:

- Risk levels: High, medium, low, informative
- Confidence levels: Confirmed user, high, medium, low.

After scanning, the following results are reflected in Fig. 7.

### 6.4 Alerts

The attack on our information system has been successfully carried out. After analysis with the owas zap tool, in Fig. 9, we see the alert of the detected vulnerability, as well as its OWASP label,<sup>3</sup> the type of vulnerability CWE,<sup>4</sup> and the Threat Classification Reference Table (WASC).<sup>5</sup> See Fig. 8.

Finally, based on the alerts generated by the tool, solutions are proposed to mitigate the vulnerability. See Fig. 9.

<sup>3</sup> <https://owasp.org/Top10/>.

<sup>4</sup> <https://cwe.mitre.org/data/definitions/264.html>.

<sup>5</sup> [http://projects.webappsec.org/w/page/13246974/Threat\\_Classification\\_Reference\\_Grid](http://projects.webappsec.org/w/page/13246974/Threat_Classification_Reference_Grid).

Alert type	Alert Tag OWASP	CWE ID	WASC ID
CSP: Wildcard Directive	OWASP_2021_A05 OWASP_2017_A06	693	15
Content Security Policy (CSP) Header Not Set	OWASP_2021_A05 OWASP_2017_A06	693	15
Cross-Domain Deconfiguration	OWASP_2021_A01 OWASP_2017_A05	264	14
Missing Anti-clickjacking Header	OWASP_2021_A05 WSTG-v42-CLNT-09 OWASP_2017_A06	1021	15
The server discloses information using an HTTP response header field(s) ""X-Powered-By""	OWASP_2021_A01 WSTG-v42-INFO-08 OWASP_2017_A03	200	13
X-Content-Type-Options Header Missing (1)	WASP_2021_A05 OWASP_2017_A06	693	15
Timestamp Disclosure - Unix	OWASP_2021_A01 OWASP_2017_A03	200	13
Disclosure of Information - Suspicious Comments	OWASP_2021_A01 OWASP_2017_A03	200	13

Fig. 8 Type of alert, vulnerabilities and classification

Alert type	Solution
CSP: Wildcard Directive	Ensure that your web server, application server, load balancer, etc. is properly configured to set the Content-Security-Policy header.
Content Security Policy (CSP) Header Not Set	Ensure that your web server, application server, load balancer, etc. is configured to set the Content-Security-Policy header, to achieve optimal browser support: "Content-Security-Policy" for Chrome 25+, Firefox 23+ and Safari 7+, "X-Content-Security-Policy" for Firefox 4.0+ and Internet Explorer 10+, and "X-WebKit-CSP" for Chrome 14+ and Safari 6+.
Cross-Domain Deconfiguration	white-listed IP address, for example). Configure the HTTP header "Access-Control-Allow-Origin" to a more restrictive set of domains, or completely remove all CORS headers, to allow the web browser to reinforce the same source policy (SOP) in a more restrictive manner.
Missing Anti-clickjacking Header	Modern Web browsers support the Content-Security-Policy and X-Frame-Options HTTP headers. Ensure one of them is set on all web pages returned by your site/app.
The server discloses information using an HTTP response header field(s) ""X-Powered-By""	AMake sure your web server, application server, load balancer, etc. is configured to suppress X-Powered-By headers.
X-Content-Type-Options Header Missing (1)	Ensure that the application/web server sets the Content-Type header appropriately, and that it sets the X-Content-Type-Options header to 'nosniff' for all web pages.
Timestamp Disclosure - Unix	Manually confirm that timestamp data is not sensitive, and that data cannot be aggregated into exploitable patterns of disclosure.
Disclosure of Information - Suspicious Comments	Remove all comments that return information that could help an attacker and fix any underlying issues they refer to.

Fig. 9 Type of alert, vulnerabilities, and classification

## 7 Conclusions

We have performed prediction experiments for COVID-19 disease in this work using chest X-ray images. The proposed model provides 98.8% accuracy in detecting COVID-19 disease.

The implementation of security at access and authorization levels was carried out using the OAuth2.0 protocol that allows connecting to services using a token generated from encrypted data.

In this work, the COVID-19 prediction system in chest X-rays was secured using encryption methods. In addition, the information system was audited with the OWASP ZAP methodology, with encouraging results since it presents four risk alerts of type medium and 3 of a low type, which place the information system as safe and much more after making the corrections in the programming that are presented in the audit report, for that reason and under complying with the objective of this article. We can rely on the security and confidentiality of the data entered into this information system, processed, and generated from it.

## References

1. Coronavirus (COVID-19). <https://news.google.com/home?hl=en-IN&gl=IN&ccid=IN:en>
2. Riquelme D, Akhloufi MA (2020) Deep learning for lung cancer nodules detection and classification in CT scans. *Ai* 1:28–67
3. Jaiswal AK et al (2019) Identifying pneumonia in chest X-rays: a deep learning approach. *Measurement* 145:511–518
4. Ebrahimighahnavieh MA, Luo S, Chiong R (2020) Deep learning to detect Alzheimer's disease from neuroimaging: a systematic literature review. *Comput Methods Progr Biomed* 187:105242
5. Islam MM, Karray F, Alhaji R, Zeng J (2021) A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19). *IEEE Access* 9:30551–30572
6. Oh Y, Park S, Ye JC (2020) Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans Med Imaging* 39:2688–2700
7. Jamshidi M et al (2020) Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *IEEE Access* 8:109581–109595
8. Chowdhury ME et al (2020) Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* 8:132665–132676
9. Rahman T et al (2021) Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput Biol Med* 132:104319
10. Vayá MDLI et al (2020) Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint* [arXiv:2006.01174](https://arxiv.org/abs/2006.01174)
11. Kumar S et al (2022) LiteCovidNet: a lightweight deep neural network model for detection of COVID-19 using X-ray images. *Int J Imaging Syst Technol*
12. Shastri S et al (2022) CheXImageNet: a novel architecture for accurate classification of Covid-19 with chest X-ray digital images using deep convolutional neural networks. *Health Technol* 12:193–204
13. Orozco CI, Xamena E, Martínez CA, Rodríguez DA (2021) Covid-xr: a web management platform for coronavirus detection on x-ray chest images. *IEEE Latin Am Trans* 19:1033–1040
14. Castro JDB et al (2020) A free web service for fast COVID-19 classification of chest X-ray images. *arXiv preprint* [arXiv:2009.01657](https://arxiv.org/abs/2009.01657)
15. Meem AT, Khan MM, Masud M, Aljahdali S (2022) Prediction of covid-19 based on chest X-ray images using deep learning with CNN. *Comput Syst Sci Eng* 1223–1240
16. Sadan B (2001) Patient data confidentiality and patient rights. *Int J Med Inform* 62:41–49
17. Kassania SH, Kassanib PH, Wesolowskic MJ, Schneidera KA, Detersa R (2021) Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach. *Biocybern Biomed Eng* 41:867–879

18. Gulshan V et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316:2402–2410
19. Selvaraju RR et al (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*, pp 618–626

# Interpolated Solutions of Abel Integral Equations Using Barycentric Lagrange Double Interpolation



E. S. Shoukralla and B. M. Ahmed

**Abstract** We provide Interpolant solutions to the Apple integral equations that emerge in climate, atmosphere, heat transfer, superfluid, astrophysics, solid mechanics, scattering theory, spectroscopy, stereology, elasticity theory, and plasma physics, and other fields. We developed adequate formulas for the optimal distribution of kernel nodes to address the kernel's singularity, ensuring that the kernel does not reach infinity when one of the two variables approaches the other. Four matrices represent the data function, whereas five matrices represent the kernel. We achieved two formulas for the matrix–vector single interpolated solution, the first based on interpolated the data function while the second based on interpolated the kernel only. The matrix–vector single interpolated solution has two formulas: the first is based on interpolating the data function and the kernel, while the second is based on interpolating only the kernel. The first formula simply involves the calculation of two matrices: the elements of the first matrix are correspond to the functional values of the data function, and the elements of the second matrix correspond to the functional values of the kernel at the two sets of nodes that are associated with the kernel's variables. When compared to the solutions provided by other approaches, were found to be convergent to the exact solutions with a minimum CPU time and high accuracy, demonstrating the novelty and simplicity of the suggested method.

**Keywords** Interpolation · Abel integral equations · Weakly singular kernels · Atmosphere · Ecology · Super fluids

---

E. S. Shoukralla (✉)

Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt  
e-mail: [shoukralla@el-eng.menofia.edu.eg](mailto:shoukralla@el-eng.menofia.edu.eg)

B. M. Ahmed

Faculty of Engineering and Technology, FUE in Egypt, Cairo, Egypt  
e-mail: [Basma.magdy@fue.edu.eg](mailto:Basma.magdy@fue.edu.eg)

# 1 Introduction

Many scientific applications use Abel integral equations, including heat transfer, superfluidity, crystal formation, laser spectroscopy, planetary and cometary atmospheres, and others [1, 2]. Many articles have been published to solve these types of integral equations, such as [3–6]. Using the Taylor expansion of the unknown function, Li et al. [3] solved the Abel integral problem by changing it into an algebraic system of linear equations. The required solution is then obtained by solving the resulting system using Cramer’s rule. Zhang et al. [4] employed the multistep collocation method to solve first kind Volterra integral equations, where the multistep collocation method’s convergence condition is evaluated, and the corresponding convergence order is displayed for specified values. The Abel integral equations were solved by combining the homotopy analysis method with the Laplace transform method by Noeiaghdam et al. [5]. They proved that their combined method is convergent. Jahanshahi et al. [6] presented a method for numerically resolving Abel integral equations of the first kind based on fractional integral approximations and Caputo derivatives. Shoukralla [7] offered an interpolated solution to the weakly singular Volterra integral equation of the second kind. Shoukralla [8–10] proposes a number of rules for isolating kernel singularities that are based on the kernel’s optimal node distribution. Shoukralla et al. [11–15] proposed a number of methods for solving weakly singular First kind Fredholm integral equations, in which the unknown functions are singular near and at the integration domain. These techniques that produce high-accuracy solutions free of logarithmic singularities cannot be applied to Abel integral equations due to variations in kernels. The main goal of this research paper is to find interpolating solutions for the singular Apple integral equations by applying an advanced stable form of the traditional barycentric Lagrange formula [16–18] to interpolating the data function and the kernel twice in a row for the two variables of the kernel using a matrix–vector product, where one matrix is the monomial basis functions matrix. Shoukralla et al. [19–23] used this advanced formula to find numerical solutions to non-singular Volterra integral equations of the second kind, and the results were extremely accurate. The situation is different for Apple’s equation because of the kernel’s singularity, which inspired us to develop an optimal rule to overcome the kernel’s singularity and achieve high-accuracy matrix–vector interpolated solutions free of singularity. Furthermore, we investigated convergence of the solution’s mean and the maximum norm of error estimation in this research article, and we offered two theories in this regard. The proposed method begins by interpolating the data function through four matrices to obtain the matrix–vector single interpolated data function. Then we devise a strategy for selecting two sets of nodes for interpolation for each of the kernel’s two variables, ensuring total isolation of the kernel’s singularity. The kernel is interpolated for both of these groups of nodes, yielding the matrix–vector double interpolated kernel via five matrices, the square matrix is one of them, of the kernel’s functional values for these two sets of nodes. We can obtain the interpolated solution in matrix–vector form by employing the Laplace formula to solve the Abel integral equation [26]. We can demonstrate the

extraordinary precision of the results produced using lower-degree interpolation by solving two instances supported by tables and figures. The method's main advantage is its simplicity, as it only requires two matrices to generate adequate convergent numerical results.

## 2 Matrix–Vector Barycentric Double Interpolation Method

Consider the linear Abel integral equation

$$f(x) = \int_0^x \kappa(x, t)u(t)dt; \quad 0 < x \leq \mu < \infty \quad (1)$$

where  $f(t)$  is a known smooth function,  $u(t)$  is the unknown function to be determined, and the kernel  $\kappa(x, t)$  is given by  $\kappa(x, t) = \frac{1}{(x-t)^\alpha}$ ;  $0 < \alpha < 1$ . The Abel operator  $A$  is defined by  $Au = \int_0^x \kappa(x, t)u(t)dt$  and acting in  $L_2[\alpha, \mu]$ . Moreover, it is assumed that  $\|\kappa(x, t)\|_2 \leq \delta < \infty$  for some real number  $\delta$ . Suppose  $u(t) \in C^{n+1}[0, x]$  is provided as a tabulated function in the form  $u(x_i) = u_i$ ;  $i = \overline{0, n}$  where  $x_i = ih$ ;  $h = \frac{\mu}{n}$  and let  $u_n(x)$  be the  $n$ th degree matrix–vector barycentric interpolant polynomial which interpolates  $u(x)$  at the  $(n+1)$  equidistance distinct nodes  $\{x_i\}_{i=0}^n \subset [0, \mu]$  such that  $u_n(x_i) = u_i \forall i = \overline{0, n}$ . Furthermore, the interpolant unknown function  $u_n(x)$  then

$$u_n(x) = X(x)C^T W U \quad (2)$$

Here  $W = \text{diag}[w_i]$ ;  $i = \overline{0, n}$  is a diagonal square matrix whose entries are provided by  $w_i = (-1)^i \binom{n}{i}$ ;  $i = \overline{0, n}$ ,  $X(x) = [x^i]_{i=0}^n$  is  $1 \times (n+1)$  row monomial basis matrix, and  $U = [u_i]_{i=0}^n$  is  $(n+1) \times 1$  column unknown coefficients matrix, and  $C = [c_{ij}]_{i,j=0}^n$  is a square known Maclaurin coefficients matrix of the barycentric functions  $\Psi_i(x)$  such that  $c_{ij} = \frac{\Psi_i^{(j)}(0)}{j!} \forall i, j = \overline{0, n}$ , where  $\Psi_i(x) = \frac{\vartheta_i(x)}{\tau(x)}$ ;  $\tau(x) = \sum_{i=0}^n w_i \vartheta_i(x)$ ;  $\vartheta_i(x) = \frac{1}{x-x_i}$ . Similarity the data function  $f(x)$  can be interpolated to produce the matrix–vector as,

$$f_n(x) = X(x)C^T W F \quad (3)$$

where  $F = [f_i]_{i=0}^n$  is  $(n+1) \times 1$  column matrix of the data functional values such that  $f_i = f(x_i)$ ;  $i = \overline{0, n}$ . By the same way, the kernel  $\kappa(x, t)$  will be interpolated with respect to the variables  $x$  and  $t$  respectively to obtain the matrix–vector double interpolant kernel  $\kappa_{n,n}(x, t)$ . The strategy is choosing  $x_i, t_i$  in such a manner that the denominator of the kernel  $\kappa(x, t)$  never becomes zero and never contains complex

numbers as an inevitable result of always being  $x$ , greater than  $t$ . Hence, we put

$$\tilde{x}_i = \frac{\mu}{2} + i h_1; \quad h_1 = \frac{\mu}{2n}; \quad \tilde{t}_i = i \times h_2; \quad h_2 = \frac{\mu}{2(n + \Delta)}; \quad \Delta > 0; \quad i = \overline{0, n} \quad (4)$$

Based on the two sets of nodes  $\{\tilde{x}_i\}_{i=0}^n, \{\tilde{t}_i\}_{i=0}^n$ , we find the two corresponding square known Maclaurin coefficients matrices  $C_1 = [c_{ij}^1]_{i,j=0}^n$  and  $C_2 = [c_{ij}^2]_{i,j=0}^n$  respectively of the barycentric functions  $N_i(x)$  and  $M_i(x)$  such that  $c_{ij}^1 = \frac{N_i^{(j)}(0)}{j!} \forall i, j = \overline{0, n}$ , where  $N_i(x) = \frac{\rho_i(x)}{v(x)}; v(x) = \sum_{i=0}^n w_i \rho_i(x); \rho_i(x) = \frac{1}{x - \tilde{x}_i}$  and  $c_{ij}^2 = \frac{M_i^{(j)}(0)}{j!} \forall i, j = \overline{0, n}$ , where  $M_i(x) = \frac{\tilde{\rho}_i(x)}{\tilde{v}(x)}; \tilde{v}(x) = \sum_{j=0}^n w_j \tilde{\rho}_j(x); \tilde{\rho}_j(x) = \frac{1}{x - \tilde{t}_j}$ . Hence, we can get the matrix–vector double interpolated kernel

$$\kappa_{n,n}(x, t) = X(x) C_1^T K C_2 X^T(t); \quad K = [w_{ij} \kappa_{ij}]_{i,j=0}^n; \quad w_{ij} = w_i \times w_j, \quad \kappa_{ij} = \kappa(\tilde{x}_i, \tilde{t}_j) \quad (5)$$

For  $\alpha = \frac{1}{2}$ , the solution of (1) is then given by [26]

$$u_n(x) = \frac{1}{\pi} \frac{d}{dx} \int_0^x \tilde{\kappa}(x, t) f(t); \quad \tilde{\kappa} = \frac{1}{\sqrt{x-t}} \quad (6)$$

Thus, we get

$$u_n(x) = \frac{1}{\pi} \frac{d}{dx} \left[ X(x) C_1^T \tilde{K} C_2 \Phi(x) C^T W F \right]; \quad \Phi(x) = \int_0^x X^T(t) X(t) dt \quad (7)$$

We can get another formula for the interpolate solution  $u_n(x)$  without interpolate the data function  $f(t)$ . In this case, we get

$$u_n(x) = \frac{1}{\pi} \frac{d}{dx} \left[ X(x) C_1^T \tilde{K} C_2 N(x) \right]; \quad N(x) = \int_0^x X^T(t) f(t) dt \quad (8)$$



### 3 Convergence in the Mean and Error Analysis

The convergence in the mean of the unknown interpolant function illustrated by formula (7) to the exact solution is investigated in this section. We also looked at error norm estimation and proved two theorems in the process.

**Theorem 3.1** Assume that  $\max_{x,t \in [0,\mu]} |k(x,t)| = \varepsilon\%$  and  $\int_0^\mu \int_0^\mu |k(x,t)|^2 dx dt \leq M < \infty$  where  $\varepsilon$  is a positive real number and assume that  $f(x)$  and  $u(x)$  belong to  $L_2(\alpha, \mu)$  with  $\max_{x \in [0,\mu]} f(x) = N$  and  $\max_{x \in [0,\mu]} f(x) = L$ ;  $N, L$  are positive real numbers. Then  $\lim_{n \rightarrow \infty} \|u(x) - u_n(x)\|_2 = 0$ .

**Proof** For convergence in the mean of  $u_n(x)$ , we put Eq. (2) in the polynomial

$$\text{form } u_n(x) = \sum_{i=0}^n u_i x^i = X(x)U; X(x) = [1 \ x \ \dots \ x^n], U = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_n \end{bmatrix}, \text{ where}$$

$u_i$  are the unknown coefficients to be determined. Let the exact solution takes the form  $u(x) = \sum_{j=0}^n a_j x^j = \tilde{A} \tilde{X}^T(x)$ ;  $\tilde{A} = [a_0 \ a_0 \ \dots \ a_n]$ , where  $a_i$  are known coefficients then,

$$-2 \int_0^\mu |u(x)u_n(x)| dx = -2 \int_0^\mu \left| \sum_{j=0}^n u_j x^j \sum_{i=0}^n a_i x^i \right| dx = -2 \int_0^\mu |\tilde{A} \tilde{X}(x)U| dx = -2 |\tilde{A} \Phi(\mu)U| \quad (9)$$

$$\tilde{X}(x) = \tilde{X}^T(x) \tilde{X}(x); \quad \Phi(\mu) = \int_0^\mu \tilde{X}(x) dx = \left[ \frac{\mu^{i+j}}{i+j+1} \right]_{i,j=0}^n \quad (10)$$

Moreover, we have

$$\int_\alpha^\mu |u_n(x)|^2 dx = \int_0^\mu |U^T \tilde{X}(x)U| dx = |U^T \Phi(\mu)U|, \int_\alpha^\mu |u(x)|^2 dx = \int_0^\mu |\tilde{A} \tilde{X}(x) \tilde{A}^T| dx = |\tilde{A} \Phi(\mu) \tilde{A}^T| \quad (11)$$

Letting  $n \rightarrow \infty$  means that  $i, j \rightarrow \infty$  and hence  $\Phi(\mu) \rightarrow 0$ . Thus, it is proved that

$$\lim_{n \rightarrow \infty} \|u(x) - u_n(x)\|_2 = 0 \quad (12)$$

**Theorem 3.2** Rewrite Eq. (1) in the shape  $f = Au$ , where the operator  $A$  is defined by  $Au = \int_0^\mu \kappa(x,t)u(t)dt$  is the Abel operator. Sample the error norm of the interpolation by  $\mathfrak{R}_n(x)$  such that  $\mathfrak{R}_n(x) = \|Au - Au_n\|_2$  where  $\|\cdot\|_2$  denotes the Euclidean norm in  $\mathbb{R}^2$  and  $Au_n = f_n$ . Then  $\mathfrak{R}_n(x) = \|Au - Au_n\|_2 = 0$ .

**Proof** We have  $\|Au - Au_n\|_2 \leq \|f - f_n\|_2$ .

$\max_{x,t \in [0,\mu]} |\kappa(x,t)| = \delta$ , and  $\int_0^\mu \int_0^\mu \int_0^\mu |\kappa(x,t)|^2 dx dt \leq M < \infty$  where  $\lambda_1, \lambda_2$  are positive real numbers. Let  $f(x) = \sum_{i=0}^n f_i x^i = X(x)F$ ;  $F = [f(x_0) f(x_1) \dots f(x_n)]$  and  $f_n(x) = X(x)C^T W F$ , then we find that

$$\int_0^\mu |f(x)|^2 dx = |F^T \Phi(\mu) F| \leq \gamma_1; \int_0^\mu |f_n(x)|^2 dx = |F^T W C \Phi(\mu) C^T W F| \leq \gamma_2 \quad (13)$$

where  $\gamma_1, \gamma_2$  positive real numbers. By the inequality

$$\int_0^\mu |f(x) f_n(x)| dx \leq \frac{1}{2} \left( \int_0^\mu |f(x)|^2 + \int_0^\mu |f_n(x)|^2 \right) dx = \frac{1}{2} (\gamma_1 + \gamma_2) \quad (14)$$

we find that  $\|f - f_n\|_2 = 0$ . Moreover, we have

$$\|Au - Au_n\| = \left[ \int_0^t \left| \int_0^x \kappa(x,t) u(t) dt - \int_0^x \kappa_{n,n}(x,t) u_n(t) dt \right|^2 dx \right]^{\frac{1}{2}} \quad (15)$$

By Cauchy–Bunyakovski inequality [24, 25], we get

$$\begin{aligned} \int_0^t \left| \int_0^x \kappa(x,t) u(t) dt \right|^2 dx &\leq \int_\alpha^t \|\kappa(x,t)\|_2^2 \|u(t)\|_2^2 dx = \|u(t)\|_2^2 \left[ \int_0^t \left( \int_0^x |\kappa(x,t)|^2 dt \right) dx \right] = \frac{(\delta t)^2}{2} \|u(t)\|_2^2 \quad (16) \\ \int_0^t \left| \int_0^x \kappa_{n,n}(x,t) u_n(t) dt \right|^2 dx &\leq \int_\alpha^t \|\kappa_{n,n}(x,t)\|_2^2 \|u_n(t)\|_2^2 dx = \|u(t)\|_2^2 \left[ \int_0^t \left( \int_0^x |\kappa_{n,n}(x,t)|^2 dt \right) dx \right] = H(t) \|u_n(t)\|_2^2 \quad (17) \end{aligned}$$

Here, we have

$$\int_0^x |\kappa_{n,n}(x,t)|^2 dt = \int_0^x \left| X(x) Z \tilde{X}(t) Z^T X^T(x) \right| dt = X(x) Z \Phi(x) Z^T X^T(x) \quad (18)$$

where

$$|\kappa_{n,n}(x,t)|^2 = \left| X(x) Z X^T(t) X(t) Z^T X^T(x) \right| = \left| X(x) Z \tilde{X}(t) Z^T X^T(x) \right|; Z = C_1^T \tilde{K} C_2 \quad (19)$$

$$\int_0^t \left( \int_0^x |\kappa_{n,n}(x,t)|^2 dt \right) dx = H(t) \quad (20)$$

Furthermore,

$$-2 \int_0^t \left| \int_0^x \kappa(x, t) u(t) dt \times \int_0^x \kappa_{n,n}(x, t) u_n(t) dt \right| dx = -\frac{(\delta t)^2}{2} \|u(t)\|_2^2 - H(t) \|u_n(t)\|_2^2 \quad (21)$$

By substituting from (16), (17), and (21) into (15), we get  $\|Au - Au_n\| = 0$ . Thus, we proved that  $\mathfrak{R}_n(x) = 0$ .

## 4 Computational Results

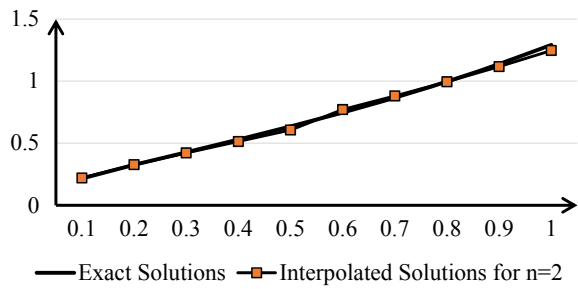
We solved two instances to demonstrate the method's outstanding performance. We utilized MATLAB Version R2019s for the calculations. On the basis of the node distributions provided in formula (4), formula (7) was used to solve instances 1, while formula (8) was utilized to solve the second example. In all cases, we used the lowest interpolation degree  $n = 2$  to find the interpolated solutions for  $x = 0.1 : 0.1 : 1.0$ .

**Example 1** Given  $e^x - 1 = \int_0^x \frac{u(t)}{\sqrt{x-t}} dt$  with exact solution  $u(x) = \frac{e^x}{\sqrt{\pi}} \operatorname{erf}(\sqrt{x})$  [6]. We use formula (4) for  $x = 0.1 : 0.1 : 0.5$  with  $\Delta = 3$ . For  $x = 0.6 : 0.1 : 1.0$ , we put  $\Delta = 2$  and then applying formula (7) to find the matrix–vector barycentric interpolated solutions  $u_2(x)$  for  $x = 0.1 : 0.1 : 0.5$ . The absolute errors are expressed by  $R_2(x)$ , where  $R_2(x) = |u(x) - u_2(x)|$ , whereas the relative errors test are denoted by  $\mathfrak{R}_n(x)$  where  $\mathfrak{R}_2(x) = \frac{1}{2} (10)^{-5} |u_2(x)|$ . See Table 1. Figure 1 depicts the graphs of the exact solutions and the interpolated solutions for  $x = 0.1 : 0.1 : 1.0$ . When compared to the solutions in [4, 6], the obtained interpolated solutions strongly converge to the exact solution and have good accuracy.

**Table 1** Results of the solution of Example 1 for  $n = 2$

$x$	$u(x)$	$u_2(x)$	$R_2(x)$	$\mathfrak{R}_2(x)$	CPU time (s)
0.100	0.21529	0.21941	0.0041161	$1.097e^{-6}$	12.080
0.200	0.32588	0.32674	0.00085543	$1.6337e^{-6}$	10.487
0.300	0.42757	0.42174	0.0058257	$2.1087e^{-6}$	10.359
0.400	0.52933	0.51365	0.015678	$2.5683e^{-6}$	12.465
0.500	0.63503	0.60624	0.028797	$3.0312e^{-6}$	11.473
0.600	0.74704	0.7716	0.024556	$3.858e^{-6}$	10.955
0.700	0.86719	0.88055	0.013358	$4.0066e^{-6}$	12.978
0.800	0.99709	0.99538	0.0017054	$4.9769e^{-6}$	13.814
0.900	1.1383	1.1173	0.021019	$5.5864e^{-6}$	11.259
1.000	1.2924	1.2473	0.045047	$6.2367 e^{-6}$	11.816

**Fig. 1** Graphs of the exact solutions  $u(x)$  and the interpolated solutions  $u_2(x)$

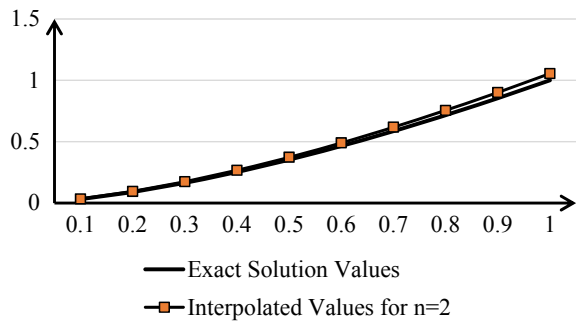


**Example 2** Given  $\frac{3\pi}{8}x^2 = \int_0^x \frac{u(t)}{\sqrt{x-t}}dt$  with exact solution  $u(x) = x^{\frac{3}{2}}$  [26]. Using formula (4) with  $\Delta = 1$ , and  $n = 2$  for  $\mu = 0.1 : 0.1 : 1.0$ , we find the matrix–vector barycentric interpolated solutions  $u_2(x)$  for  $x = 0.1 : 0.1 : 1.0$  by employing formula (8). The obtained results are shown in Table 2. The generated interpolated solutions using lower-degree interpolation exhibit good accuracy and strongly converge to the exact solution. See Table 2 and Fig. 2 for more information.

**Table 2** Results of the solution of Example 2 for  $n = 2$

$x$	$u(x)$	$u_2(x)$	$R_2(x)$	$\Re_2(x)$	CPU time (s)
0.1000	0.031623	0.033365	0.0017426	$1.6683e^{-7}$	15.707
0.2000	0.089443	0.094371	0.0049287	$4.7186e^{-7}$	15.951
0.3000	0.16432	0.17337	0.0090546	$8.6686e^{-7}$	17.519
0.4000	0.25298	0.26692	0.01394	$1.3346e^{-6}$	17.560
0.5000	0.35355	0.37304	0.019482	$1.8652e^{-6}$	17.169
0.6000	0.46476	0.49037	0.02561	$2.4518e^{-6}$	17.530
0.7000	0.58566	0.61793	0.032273	$3.0897e^{-6}$	17.612
0.8000	0.71554	0.75497	0.03943	$3.7749e^{-6}$	17.651
0.9000	0.85381	0.90086	0.047049	$4.5043e^{-6}$	17.349
1.0000	1.0	1.0551	0.055105	$5.2755e^{-6}$	16.118

**Fig. 2** Graphs of the exact solutions  $u(x)$  and the interpolated solutions  $u_2(x)$  at  $x = 0.1 : 0.1 : 1.0$



## 5 Conclusion

The present study is a simple and straightforward approach to get interpolated solutions of Abel integral equations. To interpolate both the given data function and the kernel, we apply new matrix–vector barycentric single and double interpolation formulas. We can totally eliminate the kernel singularity when one variable approaches the second using the rule, we created for node distributions of kernel variables. When compared to the exact ones and those found by other referenced methods, the matrix–vector barycentric Lagrange interpolated solutions, which are achieved with the lowest interpolation degrees and minimum CPU time, are exceptionally accurate, unlike other methods that employ high approximation degrees. This demonstrates the proposed method's originality and superior results. In future work, we will simplify the calculations, decrease CPU time, and calculate the numerical solutions when  $n$  more than two to reduce errors and make the numerical solution more strongly converge to exact solutions.

## References

1. Wu J, Zhou Y, Hang C (2020) A singularity free and derivative free approach for Abel integral equation in analyzing the laser-induced breakdown spectroscopy. *Spectrochim Acta Part B Atomic Spectrosc* 167:105791
2. Hubert B, Munhoven G, Moulane Y, Hutsemekers D, Manfroid J, Opitom C, Jehin E (2022) Analytic and numerical methods for the Abel transform of exponential functions for planetary and cometary atmospheres. *Icarus* 371:114654
3. Huang L, Huang Y, Li XF (2008) Approximate solution of Abel integral equation. *Comput Math Appl* 56(7):1748–1757
4. Zhang T, Liang H (2018) Multistep collocation approximations to solutions of first-kind Volterra integral equations. *Appl Numer Math* 130:171–183
5. Noeiaghdam S, Zarei E, Kelishami HB (2016) Homotopy analysis transform method for solving Abel's integral equations of the first kind. *Ain Shams Eng J* 7(1):483–495
6. Jahanshahi S, Babolian E, Torres DF, Vahidi A (2015) Solving Abel integral equations of first kind via fractional calculus. *J King Saud Univ-Sci* 27(2):161–167
7. Shoukralla ES (2021) Interpolation method for solving weakly singular integral equations of the second kind. *Appl Comput Math* 10(3):76–85
8. Shoukralla ES (2021) Interpolation method for evaluating weakly singular kernels. *J Math Comput Sci* 11(6):7487–7510
9. Shoukralla ES, Ahmed BM, Sayed M, Saeed A (2022) Interpolation method for solving Volterra integral equations with weakly singular kernel using an advanced barycentric Lagrange formula. *Ain Shams Eng J* 13(5):101743
10. Shoukralla ES, Ahmed BM, Saeed A, Sayed M (2023) The interpolation-Vandermonde method for numerical solutions of weakly singular Volterra integral equations of the second kind. In: *Proceedings of seventh international congress on information and communication technology*. Springer, Singapore, pp 607–614
11. Shoukralla E (2020) A numerical method for solving Fredholm integral equations of the first kind with logarithmic kernels and singular unknown functions. *Int J Appl Comput Math* 6(6):1–14

12. Shoukralla ES (2021) Application of Chebyshev polynomials of the second kind to the numerical solution of weakly singular Fredholm integral equations of the first kind. *IAENG Int J Appl Math* 51:8
13. Shoukralla ES, Markos MA (2018) The economized monic Chebyshev polynomials for solving weakly singular Fredholm integral equations of the first kind. *Asian-Eur J Math* 12(1):2050030-1–2050030-10
14. Shoukralla ES, Markos MA (2020) Numerical solution of a certain class of singular Fredholm integral equations of the first kind via the Vandermonde matrix. *Int J Math Models Methods Appl Sci* 14:48–53
15. Shoukralla ES, Kamel M, Markos MA (2018) A new computational method for solving weakly singular Fredholm integral equations of the first kind. In: 13th international conference on computer engineering and systems (ICCES), pp 202–207
16. Berrut JP, Trefethen LN (2004) Barycentric Lagrange interpolation. *SIAM Rev* 46(3):501–517
17. Higham NJ (2004) The numerical stability of barycentric Lagrange interpolation. *IMA J Numer Anal* 24(4):547–556
18. Gander W (2005) Change of basis in polynomial interpolation. *Num Linear Algebra Appl* 12:769–778
19. Shoukralla ES, Ahmed BM (2022) Barycentric Lagrange interpolation matrix–vector form polynomial for solving Volterra integral equations of the second kind. In: *Proceedings of sixth international congress on information and communication technology*. Springer, Singapore, pp 151–161
20. Shoukralla ES, Elgohary H, Ahmed BM (2020) Barycentric Lagrange interpolation for solving Volterra integral equations of the second kind. *J Phys Conf Ser* 1447(1):012002. IOP Publishing
21. Shoukralla ES, Ahmed BM (2020) Numerical solutions of Volterra integral equations of the second kind using Lagrange interpolation via the vandermonde matrix. *J Phys Conf Ser* 1447(1):012003. IOP Publishing
22. Shoukralla ES, Ahmed BM (2019) Multi-techniques method for solving Volterra integral equations of the second kind. In: 2019 14th international conference on computer engineering and systems (ICCES), pp 209–213
23. Shoukralla ES, Ahmed BM (2020) The Barycentric Lagrange interpolation via Maclaurin polynomials for solving the second kind Volterra integral equations. In: 2020 15th international conference on computer engineering and systems, pp 1–6
24. Cvetkovski Z (2012) *Inequalities: theorems, techniques and selected problems*. Springer Science & Business Media
25. Botelho FS (2018) *Real analysis and applications*. Springer International Publishing AG, Part of Springer Nature
26. Wazwaz AM (2015) *A first course in integral equations—solutions manual*, 2nd edn. World Scientific Publishing Co. Pte. Ltd

# Procurement of the Future: Investing Today in the Technologies of Tomorrow



Elizabeth Koumpan and Anna W. Topol

**Abstract** Procurement is at inflection point. Digitization and advanced technology innovation enforce procurement evolution to resolve existing challenges which include lack of procurement data visibility, sharing, insights and consensus, supply chain disruptions, delayed time to value with monolithic IT platforms, and complex labor, tax, privacy, environmental, and invoicing laws. It is expected that AI technology will spread widely throughout society in the future, promoting competitiveness and establishing new industry structures. Investing for a sustainable future is driving major client buying behaviors and long-range corporate strategies. Our study focused on evaluating current trends, and projecting toward the vision of the next generation of procurement and its impact on technologies and solutions deployed in the future.

**Keywords** Digitization · Ecosystems · Innovation · Connected industries · Human-centric · AI · Edge computing · Intelligent workflows · Value creation · Intelligent systems · Procurement · Supply chain · Sustainability · IoT

## 1 Introduction

*Procurement's* last major shift occurred during the '80–'90s [1]. Over time, the function's role has been elevated from transactional management to strategic sourcing. Now, backed by digital tools that automate downstream procurement activities, the focus of procurement will be on collaborating with a network of external partners to create new, innovative business models [2]. Today's changes are driven by adoption of advanced technologies, which include artificial intelligence (AI), internet

---

E. Koumpan (✉)

IBM Consulting, 3600 Steeles Ave East, Markham, ON L3R 9Z7, Canada

e-mail: [ekoumpan@ca.ibm.com](mailto:ekoumpan@ca.ibm.com)

A. W. Topol

IBM Research—Watson, 1101 Kitchawan Road, Yorktown Heights, New York 10598, USA

e-mail: [atopol@us.ibm.com](mailto:atopol@us.ibm.com)

of things (IoT), intelligent workflows (IW), blockchain cloud, edge computing, and Web3.0. These innovations focus on insightful outcomes, enhanced decision-making, and next generation of customer experience enablement's. This digital procurement era requires contract management and supply chain optimization, 360° view on the business process, with sustainability, security, privacy and advanced information technology as foundational building blocks.

COVID-19 has shown the need to build just-in-time supply chain solutions to meet emergency needs. The 'new normal' will focus on capabilities to effectively adapt supply chains to new needs with new demand sensing capabilities. *For procurement—that means investing into new supplier relationships, enabling new ecosystem-based collaboration, and data sharing to provide value with social and individual benefits.*

We are moving into new era of Society 5.0, the fifth industrial revolution centered around interconnected 'systems', with cyber/physical/human interconnected engagement enabled by advanced technologies, bringing AI, IoT, robotics, blockchains, big data and other game-changing innovations into daily value chains and production processes, to provide more resilient and sustainable goods and services [3]. Human-centric ecosystems are new market generators that enable multi-entity common interest networks to form and generate mutual value for the participants. Benefits of value driven ecosystems include a deeper understanding of human engagement behavior through richer data.

The global economy is shifting from physical to digital currency [4] with digital wallet simplifying the experience of the digital marketplace and becoming more reliant on the digital financial system. Moving into digital space—everything including payments, need to leverage spend analytics.

In current environment, it is difficult to manage complexity of transactions between sellers and buyers, complicated by rapidly changing business interactions, supply chains, emergence of B2B payments and complex regulatory landscape. Governments also have issues with managing payments and invoices related to corporate taxes that leads us into implementing digital *invoicing (e-invoicing) technology*.

In this 'data-driven action' world, ensuring digitization, become a strategic imperative, we first put data into context to provide meaning; next, to understand it in relationship to other data and events to gain knowledge; and finally, to add judgment and action to achieve the full potential of value realization.

## 2 Procurement Market Insights

Our research shows procurement trends supporting (1) more thoughtful sourcing and **sustainability**, including supplier management and tracing; (2) progress toward **autonomous procurement**, including automation; (3) conversational AI and a focus on technologies that enable better decision-making, including data, advanced AI-infused analytics and dashboards; and (4) and a flexible, composable platforms that allow for **best-in-breed technologies and services**.





Fig. 1 IBM market research outcome

The procurement function must evolve beyond its transactional focus on purchasing and embrace innovation and transformation. As indicated in Fig. 1 IBM market research outcome, based on our research from analyst and thought leader materials [5] future vision for procurement includes

- Predictive, autonomous, and fully digitized processes from source to pay
- Responsible, traceable, and risk-managed sourcing
- Nearly instantaneous ordering and payment, with faster delivery.

More specifically, our new procurement vision, is implemented with intelligent workflow, allowing flexible plug and play architecture to integrate with other systems. Our study indicates that to achieve new level in procurement optimization, organizations need to embrace continuous business transformation thru information sharing, connected digital ecosystems, and new business models.

Such transformation can be achieved thru:

- *Tasks automation to eliminate manual data entry across many disconnected systems across supplier management, contract compliance, and payables by creating highly maintainable yet loosely coupled services, using RPA, machine learning, and natural language processing for capturing, recording, and interpreting structured and unstructured data.*
- *Enabling intelligent workflow, ensuring collaboration to effectively select the right sources of supply, providing a high data visibility on the supplier and the buyer side of the transaction.*
- *Generating prescriptive insights, using artificial intelligence (AI) and machine learning to analyze massive amounts of data to analyze the market, make price predictions and product recommendations.*
- *Providing personalized self-service experience with social, mobile, and conversational procurement applications, based on the user role, suppliers and commodities managed, time of year, etc.*
- *Emphasis on ecosystem, creating prebuilt integrations with ERP, Finance, SCM, PLM, and other adjacent systems to streamline the data flow between systems.*
- *Microservices, to provide necessary flexibility in dynamic market landscape.*

### 3 Next Level of Procurement Revolution

‘The crisis pushed procurement beyond its traditional hesitancy to adopt new tools. In the future, success will depend, in large part, on broader implementation and adoption of modern enabling tools, including newer capabilities such as artificial intelligence (AI) and smart automation. While procurement teams have been experimenting with these technologies. The adoption of AI-enabled and cognitive tools is expected to remain constrained’ [6, 7].

Modernizing legacy technology platforms is critical, but to achieve transformational change, these projects must be treated *as broader business change projects* that have an impact on roles, responsibilities/skills, rather than only involving technology transformation. ‘In preparation to next normal, procurement must focus on delivering sustainable savings, streamlining processes, managing risks proactively, enabling organizational agility and driving business outcome through better decision-making’ [8].

Our vision for a reimagined procurement is organized around 5 strategic themes:

- 1. Re-invent the procurement processes for the digital era through real-time automated and frictionless processes
- 2. Make procurement ethics-driven and risk-managed
- 3. Take a third parties value add procurement approach
- 4. Establish an agile operating model
- 5. Empower the workforce of the future and create advanced user experiences.

This vision is enabled thru cloud native event-driven architecture delivering an ecosystem built on connected microservices. Figure 2 key components of the next gen procurement deployments showcases how the 5 strategic themes help to realize the vision of the procurement of the future.

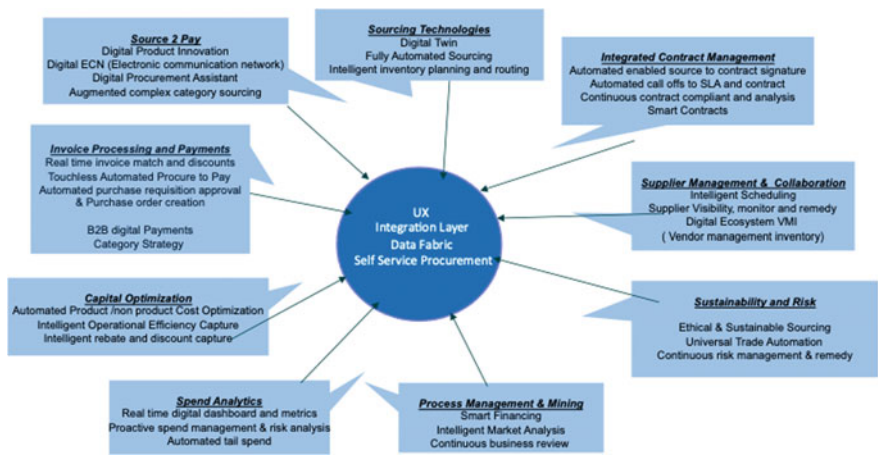


Fig. 2 Key components of the next gen procurement deployments

### 3.1 *Resolving Procurement Challenges Thru Data Sharing*

On a daily basis, procurement teams are tasked with reviewing contracts, responding to supplier inquiries, and retrieving order, product, and inventory information. Finance teams are executing hundreds of thousands of receivable and payable transactions, needing to source information about invoices, payment batches, and bank statements, among other information. This data often lives across multiple systems and still require manual effort to access it.

Data is crucial for procurement teams because, without the data, they cannot track the spending on goods and services and manage supplier and vendor relationships. Data-driven decision-making is required to ensure the buyer acquires goods and services at the best possible price when aspects such as quality, quantity, time, and location are compared. Data is vital when developing AI-based solutions and this ***data must be trusted***. Trusted data includes aspects of appropriately curated data sources (such as non-biased data) as well as appropriate data governance (including standards, provenance, and proper use).

We will need to get a better visibility of pricing and cost drivers, and then using this information in sourcing events and supplier negotiations to analyze pricing trends, market sentiment to guide buyers for specific items to purchase at specific time:

### 3.2 *Utilizing AI*

Utilizing AI technology will allow us to manage contracts from different parties (vendors, partners, and customers), covering the terms and conditions and deadlines, ensuring the vendor relationships are efficient and profitable. There is also a need to automate tasks, like areas where customers are still performing significant data entry or are leaving the procurement system to complete a task, such as supplier management, contract compliance, and payments. Table 1 procurement AI use cases highlights key AI-enabled solutions supporting procurement of the future.

For example, strategic sourcing will leverage NLP to get a better visibility of pricing and cost drivers, and then using this information in sourcing events and supplier negotiations—analyze pricing trends, market sentiment to guide buyers for specific items to purchase at specific time. For spend analytics, we will proactively identify savings opportunities, manage risks, using ML algorithms to classify procurement spend into categories and sub-categories, identify incoming orders from the business.

Another challenging area is to efficiently procure and manage complex services procurement spend, as it also must deal with temporary workforce/staff augmentation management. Early alert indicators on what workforce to engage, predicting event's impact (like COVID-19) to adjust staffing strategies and ***take more proactive approach to ensure that there is available workforce***—will be essential in the new world we live now.

**Table 1** Procurement AI use cases

AI-based e-sourcing	A next-generation e-sourcing application that adds AI capabilities to increase event automation and provide embedded decision support. There are two primary use cases for AI-based sourcing: <b>event management</b> (i.e., pause based on responses and automating feedback to suppliers), and <b>decision support</b> (i.e., identifying the best time of year and day to run a sourcing event) [9]
B2B network intelligence	Offers macrolevel insights such as identification of suppliers with increasing or declining sales, and of buying organizations that award contracts only to incumbent suppliers. It can extend to pricing trends for specific spending categories. Buyers can gain access to detailed insights related to pricing, supplier behavior, process metrics and industry trends, driving better sourcing and procurement decisions. Sourcing and procurement leaders can use such insights to improve supplier enrichment by enabling timely interventions that improve supplier performance
Supplier risk management	Using machine learning, we can provide early indicators about events to help implement contingency plans and avoid supply chain disruption. Blockchain smart contracts can replace existing transactional tracking mechanisms and execution systems with AI system integration. Organizations need to continuously update their boards/investors on the progress of sustainability initiatives, seeking suppliers who make <b>sustainability a core value</b> with the focus on direct spending with small and diverse suppliers
Smart report	Using AI/ML analyze large amounts of different data sets to generate the alerts, providing buyers with recommendations to procure specific commodities (like metals/platinum). We watch for the price of such product, analyze the events that may cause /drive the price change, and produce the recommendation, explaining the situation, suggesting why there is a right time now to buy such product, what is the best amount and price, etc. We can also add a future option of an Avatar having a conversation with the Buyers, explaining the recommendations, rationalizing benefits provided and answering any questions
Automated procurement contract operations	Gain better visibility of contracts and related operations, eliminate paper revisions and approvals, drive contract compliance to improve procurement efficiencies, ensures end-to-end company master information management and governance while providing best of breed integration for quotations, requisitions, purchase orders and subsequent fulfillment. NLP allows contract managers to extract key information from existing contracts, such as start dates or payment terms, in order to identify terms and conditions that might carry a risk. Some AI solutions can even auto-determine whether the language in third-party documents is compliant with a company's contracting policy
AI for procure to pay	Automate workflows <b>to request, procure, receive, and pay for goods and services across an enterprise</b> . Extended P2P functionality includes supplier registration, employee expenses, services and contingent labor procurement <b>and payments</b> , evolved with AI helping to the point when predicting and suggesting purchase orders for frequent, low-value indirect spend is possible

(continued)

**Table 1** (continued)

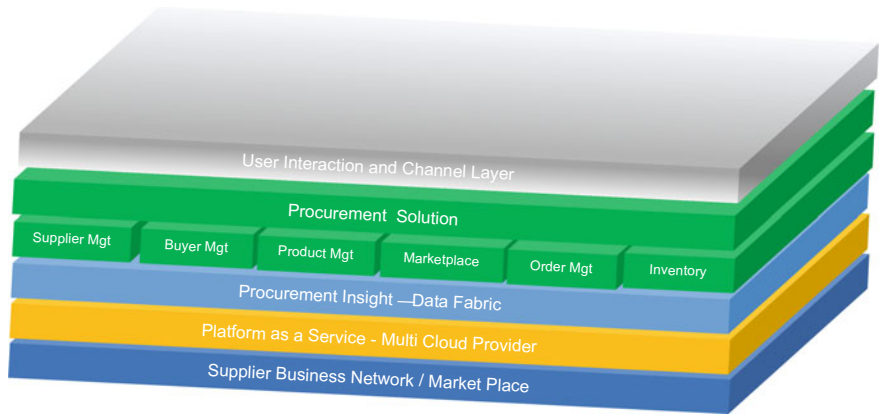
E-invoicing	Full automation of the process from the time of purchase requisition to purchase order submission and delivery, with on-line purchase and electronic invoicing, provide better customer experience. But what if we go even further, leveraging real-time unified payments, for the user to quickly identify a need and simultaneously make the purchase? How can we automatically reconcile payments to suppliers, providing visibility into payment status? <b>A payment is no longer a payment, but an integrated offering [10]</b> of processing money movement. The global economy is shifting from physical to digital currency [11] and the demand for frictionless payment grows <b>(linking physical and digital world; digital wallet disruption; digital bartering—exchange of services in lieu of funds)</b>
-------------	---

3.3 Architecting the Procurement of the Future

From an architecture point of view (see Fig. 3—High-level view of the procurement solution), we can see the procurement solution to package together a set of business services to address the procurement domain (green layer).

The business services will be microservice based, cloud native, event-driven and will be grouped into domains for adopting clear separation of concern principles and unique ubiquitous language inside each domain practices for more information. The following domains (green layers in Fig. 3) are considered:

- Procurement: the major domain of interest, supporting others to address how to buy product to support self-service procurement.
- Supplier: focuses on managing supplier entities and its metadata, supplier onboarding process, brand registry, storefront definition link supplier’s product



**Fig. 3** High-level view of the procurement solution (IBM)

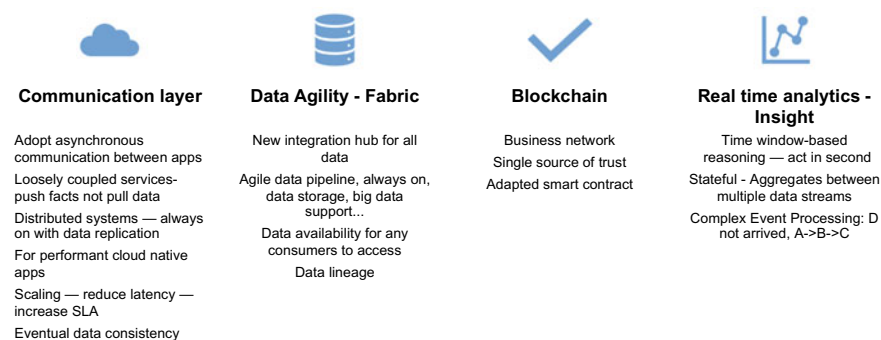
catalog, custom pricing by customer profile, quantity discount, and offers dashboard to manage the virtual store to optimize the supplier's selling strategy.

- Buyer: keeps information about buyers, buyer profile, historical shopping cards, historical product recommendation with hit rate.
- Product: product catalog and product definition to serve procurement and marketplace domain.
- Inventory: instance of the catalog and set of services to search for product availability in the marketplace.
- Finance: supports payment and financial contract management, and cryptocurrency for the market.
- Order: supports the management of an order from creation to fulfillment.
- Marketplace: as an entity, manage specific industry focused suppliers and B2B contract between suppliers, fulfillment, extends reach cross-geographies.

**Those domains support business logic to serve the self-service procurement vision.** The procurement solution will package together those services to expose reusable APIs and business events that can be used by user channels like web App, mobile, chat bots, and virtual reality devices. The light blue layer is one of the core characteristics of this architecture as it offers continuous visibility of the data in motion between all the microservices within each domain and cross domains. Figure 4 architectural blocks of the next gen procurement deployments highlights key building elements supporting this high-level architecture.

Adopting an event-driven architecture helps to address the following important capabilities:

1. Data pipeline with support for real-time event streaming to data lake, with continuous visibility of the data in motion.
2. Deliver asynchronous communication between components so they can be more responsive, resilient, and scalable.
3. Support real-time analytics on data streams, bringing business insight at the time of event occurrence.

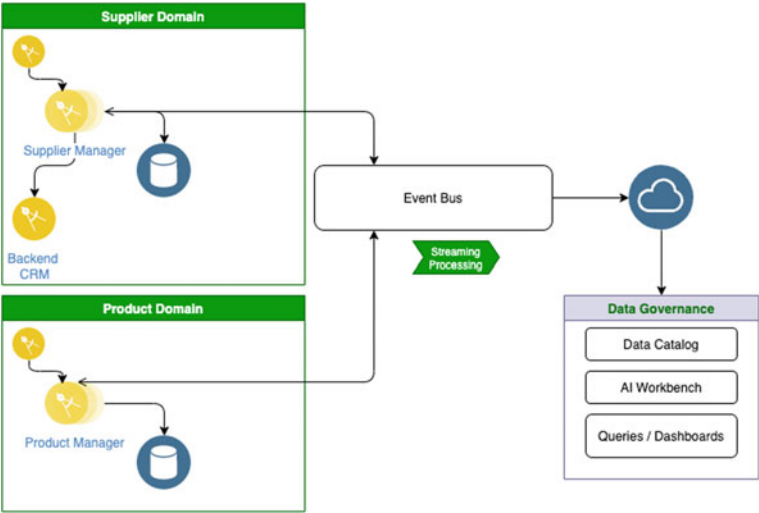


**Fig. 4** Architectural blocks of the next gen procurement deployments

*Today, there is a new way to think about data fabric, by* seeing that data is created as continuous streams of events, which can be processed in real time, and serve as the foundation for stateful stream processing application: the analytics processing moves closer to the data. In the proposed data fabric, the collect and infuse levels are running close to the same ingestion platform and moves the predictive scoring (infusion of the operationalized AI) close to the data, instead of moving the data everywhere.

The adoption of event-driven architecture, in the procurement, opens for new opportunity to leverage powerful data ingestion layer that is keeping data transformation and enrichment, as source of trust and then be able to support real-time analytics and data streaming processing to get business insight and act on the data.

Figure 5 data fabric high-level diagram below illustrates some of the procurement domains supported by different event-driven, reactive, microservices, and how the architecture supports modern data pipeline and offer flexible data fabric for better data consumption and processing. The approach is to bring computation closer to the data, for doing ad-hoc processing, with scaling to zero when possible so the overall deployment uses less resources, cost less money and has less carbon footprint.



**Fig. 5** Data fabric high-level view

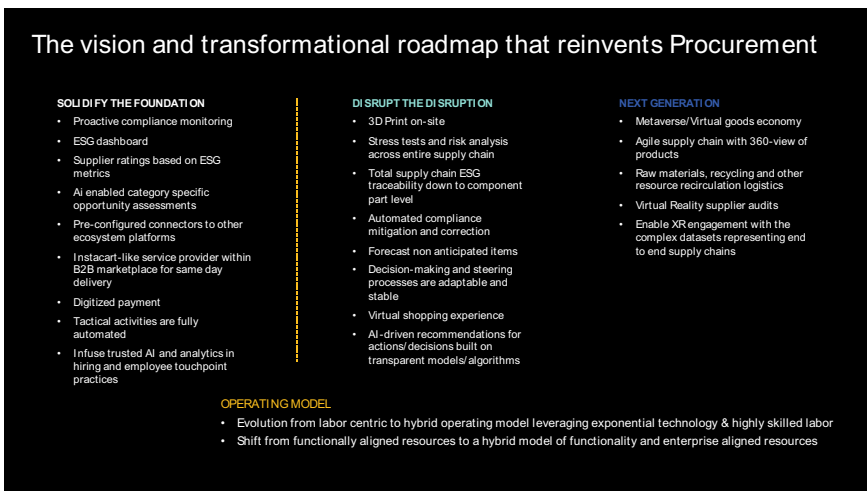
## 4 Next Steps

The outcome of our study points to procurement of the future enabling new human-centric ecosystems through collaboration with supplier partners, and the creation of value alongside social and individual good. Our vision for a reinvented procurement focuses on symbiotic relationships between people and technology with:

- Value-add approach to procurement services and technologies, delivered through composable platforms
- Predictive, autonomous, fully digitized, value-focused processes
- Unique, user-centric experiences
- Responsible, traceable, risk-managed sourcing
- Near-instantaneous ordering, delivery, and digital B2B payments
- Sustainability: Environmental, social, and governance (ESG).

As shown in Fig. 6 technology and capability roadmap to enable the procurement of the future, we believe that new levels in procurement transformation will be realized by using advanced technology and composable event-driven architecture to reduce manual effort through:

- Intelligent workflows, realized by reinforcement learning.
- Automated tasks across linked supplier management, contract compliance, and payables systems.
- Quality comparison using supplier reputational and sustainability index.
- Ecosystems prebuilt integrations with our partners to streamline the data flow between systems.



**Fig. 6** Technology and capability roadmap to enable the procurement of the future (this study)



- Personalized experience and intelligent decision management with high data visibility on the supplier and buyer side of the transaction.
- Event-driven microservices to form large applications dynamically.

**Acknowledgements** We are very grateful to Abbygale Brause, Raquel Katigbak, Vivek Visvanath, Justin McBryan, Matt Seul, Jermoe Boyer, Brenda Haddock, Didier Denove, Yada Zhu, Conor Quarry, Chester Karwatowski, Derek Harrison, Péter Hrabovszki, Michelle Lam, and Pawan Chowadhary, for their many constructive comments on the draft of the Academy of Technology study. We appreciate the support from subject matter experts and contributors including Michael Haydock, Jiandong Yin, Laura Beth, Jason Mudd, Moray Reid, Ankush Bhatia, Vadim Sheinin, Chandra Nrayanaswami, the IBM Academy of Technology, and IBM leaders Bob Murphy, Matthew Bounds, and Kyle Brown who work every day to raise the importance and awareness of the topics and technologies highlighted in this document.

## References

1. The history of procurement: past, present and future. <https://www.sourcesuite.com/procurement-learning/purchasing-articles/history-of-procurement-past-present-future.jsp>
2. Kearney Procurement: making digital transformation work for you. <https://www.de.kearney.com/procurement/article/?a=procurement-riding-the-transformative-digital-wave>
3. Koumpan E, Topol AW (2021) Promoting economic development and solving societal issues within connected industries ecosystems in society 5.0. In: Advances in artificial intelligence, software and systems engineering, pp 174–183
4. The economist—The shift from paper to virtual cash. <https://www.economist.com/finance-and-economics/2020/07/23/a-shift-from-paper-to-virtual-cash-will-empower-central-banks>. Milken Institute—The promise and peril of digital currency. <https://www.milkenreview.org/articles/the-promise-and-peril-of-digital-currency-in-a-global-economy>.
5. Analytical Research. a. IDC Future of Technology Sourcing and Procurement: <https://ibm.box.com/s/dj6ano9jtm731mlhorksowqym7pofq6x>. b. Procurement Leaders CPO Planning Guide 2022: <https://ibm.box.com/s/ckhamkix7x7hwp05hkb1x49amppm0lq>. c. IDC Market Analysis Perspective: Worldwide Digital Business Operations and Analytics Services, 2021: <https://www.idc.com/getdoc.jsp?containerId=US47082921>. d. Everest Group—Elevating procurement’s role in the next normal through digital enablement: <https://ibm.ent.box.com/file/859424488935/https://www2.everestgrp.com/reportaction/EGR-2020-22-R-4149/Marketing>. e. Procurement Leaders Ovation 2021: <https://ibm.box.com/s/xg0dk57rm4isbcium1oyzqwtk3yklwy1>
6. The Hackett Group. <https://www.thehackettgroup.com/blog/2021-procurement-key-issues/>
7. Rethinking operations in the next normal. <https://www.mckinsey.com/business-functions/operations/our-insights/rethinking-operations-in-the-next-normal>
8. Everest Group—Elevating procurement’s role in the next normal through digital enablement—PO SOTM 2021. <https://www2.everestgrp.com>
9. Harvard Business Review. <https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like>
10. Stax—What is an integrated payment system. <https://fattmerchant.com/blog/what-is-an-integrated-payment-system-and-how-it-helps-your-business/>
11. Economist. <https://www.economist.com/finance-and-economics/2020/07/23/a-shift-from-paper-to-virtual-cash-will-empower-central-banks>

# Development and Validation of a Health Information System to Improve Prenatal Controls in Guatemala



Ignacio Prieto-Egido , Aitor Garrido Madrigal,  
and Cristina Barrena García

**Abstract** Reducing maternal mortality, one of the targets in the sustainable development goals, continues to be a challenge for many low- and middle-income countries. Initiatives that seek to address this problem often face difficulties in recording and managing information on antenatal check-ups. Information on controls is critical to be able to monitor pregnancy adequately and to be able to evaluate the results of interventions. This article details the process of developing and validating an information system used in a project to improve prenatal controls in rural areas of Guatemala. The development of the system began with a needs analysis in which local institutions participated, and OpenMRS, a free software tool, was selected for implementation. The proposed information system became operational in September 2014 and, by December 2016, had already registered more than 10,000 attendances. The cases reported increased progressively in that period while the percentage of completed forms improved, reaching close to 100% at the end of the analysis period. The tool also allowed 99.31% of cases to be reviewed by specialists. The analysis results show that the system was adopted by health personnel, although some aspects of improvement that should be considered in future versions have been identified.

**Keywords** Health information system · Maternal mortality · Electronic clinical forms · Open free software · Prenatal controls

## 1 Introduction

Reducing maternal mortality is one of the health challenges highlighted by the United Nations in the sustainable development goals (SDGs). The highest maternal mortality ratios occur in developing countries, especially in rural areas [1]. In the case of

---

I. Prieto-Egido (✉) · A. Garrido Madrigal  
Universidad Rey Juan Carlos, Madrid, Spain  
e-mail: [ignacio.prieto@urjc.es](mailto:ignacio.prieto@urjc.es)

C. Barrena García  
Fundación EHAS, Madrid, Spain

Guatemala, the trend of the maternal mortality ratio has been reduced in recent years, being 2019 the year with the lowest mortality, presenting 101 maternal deaths per 100,000 live births [2]. However, although the trend is positive, the goal set in the SDGs is still far: reducing the number of maternal deaths to 70 per 100 live births [3].

Given the needs outlined above, the EHAS Foundation and the TulaSalud association have developed the Healthy Pregnancy project [4] to address the problem of maternal mortality with an innovative solution which adapts to the unfavourable conditions of rural communities. The main idea is to provide equipment and training to itinerant health brigades. The health staff is equipped with a portable ultrasound scanner powered by a folding solar panel and a clinical analysis system using rapid test strips [5]. The two tests guarantee quality prenatal care for pregnant women in rural areas.

Reliable health information is critical to evaluate initiatives like Healthy Pregnancy, but it is lacking in rural areas of Guatemala and other developing countries [6]. Therefore, the Healthy Pregnancy project was required to systematise information collection. During the project's first pilot, which took place between 2012 and 2014, data was collected through an online platform aimed at deferred teleconsultations. This pilot was implemented with three health brigades. However, when expanding the project to more than 25 health brigades in 2014, some specific requirements for the health information system were identified.

Specifically, the health information system should be able to manage forms with text data and images collected by the portable ultrasound scanner to have the complete information of each prenatal control throughout pregnancy. The system should also allow an offline mode since there are no telecommunication services (neither mobile nor fixed) in the intervention communities. In addition, it should generate indicators enabling monitoring of the project's development and evaluating its impact. In addition, the data from the consultations should be reviewed by specialists in gynaecology remotely and on a delayed basis with the dual purpose of confirming that the ultrasound scans were performed with diagnostic quality and providing feedback to the health brigades. Finally, the platform must be available in Spanish, the common language for health personnel who participate in Healthy Pregnancy, although there are different local languages in each region of Guatemala.

The platform used so far belonged to a company that was not interested in making the adaptations required by the project, and a new solution was needed. This article details the process followed to develop the required information system, which was implemented in 2014 and has been used until 2021. It shows the results that were achieved to obtain learnings that can be used in future deployments of this type.

## 2 Previous Work

In recent years, various works have tried to improve the collection of medical information in rural areas of developing countries [7, 8]. The Ebola epidemic in West Africa from 2014 to 2016, with more than 28,000 infected and 11,000 deaths, posed a challenge to health systems in three of the world's most impoverished countries [9]. Existing health facilities in the affected areas did not have the capacity, staffing and infection control capabilities needed to cope with this outbreak [10]. Therefore, an electronic medical record (EHR) based on OpenMRS was created to collect patient data. These data were urgently needed for care, surveillance, and research [11]. The result was OpenMRS-Ebola [12], an open-source Ebola EHR system that was deployed in 2015 in Sierra Leone. To increase the deployment speed, the software was built as a browser-based web application that requires constant access to the server. The final solution has two versions: a desktop application based on a browser and an Android app. Each of these applications has a different interface, although it allows access to the same data and software infrastructure. Both versions offer the same functionalities and work on the same server database. From January to March 2015, nearly 100 doctors were trained to use the system. In total, 112 out of 456 patients in Sierra Leone were registered in OpenMRS-Ebola [12], with data recorded at least until 2017.

Another system is the HIV-EMR [13], created in 2005 to support patient care in Rwinkwavu, Rwanda, and based on OpenMRS. The system was adapted to match local conditions, including clinic details and Rwandan reporting requirements. More than 1500 HIV patients taking antiretroviral drugs were registered until 2019. This implementation was successfully deployed in less than a year. The system offered a reporting module for nationwide reporting, clinical support, administrative summaries, and research. Also, forms and reports are created to support related treatment programmes, allowing their export to CSV format.

Finally, in 2009, an OpenMRS-based system was implemented in Nigeria in 2009 to develop an electronic information system to support maternal and child health. In addition, OpenMRS could be quickly customised to support not only patient-based data management, but also aggregated data at the clinical, district and state levels [14].

All these projects have served as an example for developing and implementing other OpenMRS-based systems in different countries [12]. Some shared features of these systems are the use of customised forms and remote access from a web browser or an Android App using an internet connection. However, the designs in these projects did not meet all the requirements of the Healthy Pregnancy initiative mentioned above: to operate without the need for a continuous internet connection, incorporate ultrasound images into the forms, and be translated into Spanish.

### 3 Objective

The main objective of this work is to propose and validate an information system that covers the needs of a prenatal care programme in the rural context of Guatemala. This health information system is intended to improve the information collection and management in the Healthy Pregnancy project, facilitating the sharing of information, monitoring health indicators and evaluating the initiative.

### 4 Methodology

To achieve the objective, we started with an analysis of the project's needs and then selected a platform to meet those needs. Next, we tested the platform by deploying it in a real-world scenario and analysing the results to confirm that the established objectives had been met.

#### *4.1 Analysis of the Need of the Project*

Meetings were held with two key stakeholders in the intervention area, the Alta Verapaz Department of Guatemala, to identify the project's needs. The two stakeholders are the Ministry of Public Health and Social Assistance of Guatemala (MSPAS), represented by the Alta Verapaz Health Area, and the TulaSalud Association. The Ministry of Health, through the Directorate of The Area of Alta Verapaz, manages the public health system to provide comprehensive health care to the population. It is the institution responsible for guaranteeing free access to quality prenatal controls for pregnant women in Guatemala. On the other hand, TulaSalud is a non-governmental organisation from Guatemala that works with the support of the Canadian Tula Foundation and cooperates with the Ministry of Public Health and Social Assistance (MSPAS). TulaSalud contributes to reducing maternal and infant mortality through tools for the remote training of health staff and e-health services in rural areas of the country. In addition, specialists in gynaecology from the Alcorcón Foundation University Hospital in Spain also supported this process.

In the meetings held, we sought:

1. Define the critical information to appropriately follow-up on pregnant women in the context of Guatemala: information fields, type, restrictions, etc. This information depends on the context because specific tests or treatments are unavailable in the rural context.
2. Identify information flows: who records the information, when, who needs to access it, and when.
3. Structure the information in forms as close as possible to those already known by health personnel.

## ***4.2 Select the Tool/Platform/Technology to be Used to Develop the Solution***

When the information system's needs were precise, a review and comparison of tools that allow responding to these needs were conducted. The comparison was based on the following criteria:

- If they were free software tools that avoid generating dependencies with specific companies.
- The available information and documents for both users and developers.
- Whether there were recent versions and how often they were updated.
- If the technology used was widespread and had a vast community of users and developers.
- If the tool worked with web technologies, it could be used through the web browser of different devices.
- That allowed adapting the forms according to the project's needs.
- That allowed incorporating of new modules with the functionalities required by the project.

## ***4.3 Process of Adapting the Selected Tool to the Project***

1. Collect in paper forms what information was required during the different phases of the project and what indicators were needed based on that information.
2. Propose a mechanism to identify patients who do not have an identity document.
3. Generate a first version of the forms with the selected tool. Contrast it with project partners and health authorities and iterate until we had a version that met all parties' expectations.
4. Study how to configure the tool to provide the rest of the required functionalities.

## ***4.4 Project Evaluation Method***

The system was deployed in the Healthy Pregnancy project in two departments of Guatemala. Data collected between September 2014 and December 2016 was analysed to evaluate the information system. Specifically, the evolution of the number of records in that period, the age range of pregnant women registered in the system, the percentage of completed fields and the number of cases reviewed by specialists were analysed.

## 5 Results

### 5.1 *Analysis of the Needs*

A suitable form for controlling pregnant women in rural communities should include the fields consulted by gynaecologists in that specific context. According to the systematisation of the consultation process carried out by the WHO [15] (World Health Organisation), we can divide medical visits into three types:

- Initial visit: includes the patient's history and medical examination. At this visit, personal, clinical and obstetric history should be evaluated to determine factors that may affect the development of pregnancy. Personal data includes information such as marital status or economic resources. The clinical history focuses on investigating the specific diseases and pathologies suffered by the patient. The obstetric history refers to the medical data of previous states of gestation, ranging from the number of pregnancies to the characteristics and complications in each. After the patient's history has been registered, the health staff will proceed with a medical exam that includes: vital signs, physical examination, blood tests (anaemia and test for sexually transmitted diseases) and urine (bacteriuria and proteinuria). The exam will also include an ultrasound scan.
- Follow-up visits: the following visits will be based on the same medical exam described at the first visit (without collecting the patient's history again), adapted to the weeks of pregnancy and the previous medical and ultrasound data.
- Final visit: this type of visit is the postpartum medical consultation. It includes a physical examination, collecting birth data such as the place of delivery, the state of the newborn and the mother, and whether there was a need for surgical intervention.

Although these are the recommendations of the WHO, it was necessary to know how they are implemented in the prenatal care system in Guatemala and the management of the information they carried out. That is why meetings were held with the heads of the Alta Verapaz Health Area and with the NGO TulaSalud.

The protocol established by the Ministry of Health was to fill in the information about prenatal check-ups in paper forms. These forms are stored in the rural health facility, and only a summary of the number of prenatal controls provided is shared with the higher levels. In addition, these forms only collect information related to basic tests usually available in rural health facilities, which do not include issues such as blood tests or ultrasounds. For this reason, when implementing the Healthy Pregnancy project, it was necessary to have some additional fields on HIV, syphilis, hepatitis B and urine testing. Also, we incorporated fields for the measurements made in the ultrasounds that allow estimating the weight of the fetus and the weeks of gestation. It also included the option of saving the images of the ultrasounds so that specialists in gynaecology could later review the whole case.

Although some additional fields were introduced, the resulting form was very similar to the paper form and focused on issues aimed at detecting obstetric risks

that the public health system can manage in Guatemala. Some health issues, such as malformations, are not included in this form, nor are they the subject of study by the project because they are excessively complex for the rural staff or the health system cannot solve them.

Along with the prenatal care form, a review form was defined that allows gynaecology specialists to assess the quality of ultrasounds and patient management. These forms will enable the identification of the most common difficulties of the staff involved in the project and offer them reinforcement sessions to ensure that they provide good-quality prenatal care. Section 5.3 shows how the information in the forms was structured when implementing them in the project information system.

## 5.2 *Selecting the Tool*

To select the most appropriate tool to implement the project information system, we began by reviewing those available at the time of implementation (the year 2012). The review focused on tools based on web technologies and the rest of the project requirements already mentioned. Among all the existing SIS, we proceed to list those ones that met the above requirements:

- OpenVista [16]: It is the open-source version of Vista, a Health Information System implemented in more than 1500 health facilities worldwide. It is a multi-platform tool that seeks to increase clinical efficiency and adapt to the prevailing health system.
- DHIS2 [17]: The acronym corresponds to District Health Information System based on web technology. It provides the means for reporting and analysis based on the data collected. It is a generic tool, very little preconfigured, with an open metadata model and a flexible user interface that allows the user to design the contents of the specific information of the system without programming. The tool is developed with free and open-source Java frameworks. At the time the project was launched, it did not have a version translated into Spanish, nor did it allow the inclusion of images in the forms. Nor did it allow offline work.
- OpenEMR [18]: It is a web-based tool for administering medical practise, storing medical records, medical prescriptions and billing. It is one of the free software Electronic Medical Records Management applications (medical records) with more use in both developed and developing countries.
- OpenMRS [19]: It is a web-based electronic medical record system aimed at improving health care in environments lacking resources. It is a multi-funded application implemented in 2004, and its functionalities are continuously extended with new modules. It is maintained by a community of developers and is open source.



- MedClipse [20]: The MedClipse application offers an electronic medical record based on open source. Complemented by agenda management, physicians and data management

Among all the SIS mentioned, it was considered that OpenMRS was the one that best suited the needs of the project. The main advantage of OpenMRS was that it offered a synchronisation module that allowed working without an internet connection.

### 5.3 *Process of Adapting the Selected Tool to the Project*

Once the information that the forms should contain had been identified and a tool defined to implement the system, the following development steps were taken:

1. **Propose a mechanism to register patients who do not have an official identification document.** Many patients did not have an identification document, and many common names made it difficult to register patients by name. The system generates a unique code for all these patients without an identification number using the patient's name, surname, date of birth and community.
2. **Generate a first version of the forms with the selected tool.** The first version of the forms was tested with the health staff. Pregnant women came from communities with little access to medical resources, so they had little verified information about their clinical and obstetric history. That is why the health personnel considered it appropriate to simplify the background fields and focus on the ones that most impact treatment or subsequent tests. Table 1 shows the background data included in the final form. The form of Table 1 includes some validation rules to check if the information is coherent. For example, the sum of live births and stillborn children should equal the number of births.

On the other hand, prenatal care consultations provide the necessary tests to ensure quality prenatal controls considering the existing technologies for ultrasound, blood or urinary tests in rural areas with difficult access. With the background already filled, the rest of the first and the other visits before delivery had a standard format to simplify the data collection process. The data to fill in were those in the following Tables 2, 3, 4 and 5.

In the "Ultrasound" field, the data collected through the portable ultrasound scanner were shown and were used to calculate foetal age. The ultrasound images were also included in the form to facilitate the case review by having the complete information in one place. In the first trimester, only one ultrasound image is stored, the CLR. In the following trimesters, the images are: the length of the femur, the cranioencephalic diameter, the cranial perimeter, the abdominal perimeter, the placenta image and the maximum vertical column index.

Data about the delivery is collected in a final form (Table 6) to have this information for future pregnancies and to allow the analysis of the project results.

**Table 1** The different fields collected in the electronic form and corresponding to the medical history of the pregnant woman

Medical history	
Previous pathologies	<ul style="list-style-type: none"> <li>• Healthy</li> <li>• Diabetes</li> <li>• HTA</li> <li>• Asthma</li> <li>• Smoker</li> <li>• Epilepsy</li> <li>• Hypothyroidism</li> <li>• Hyperthyroidism</li> <li>• HIV</li> <li>• Hepatitis B</li> <li>• Tuberculosis</li> <li>• Malnutrition</li> <li>• Renal Lithiasis</li> <li>• Cardiopathy</li> <li>• Cancer</li> <li>• Previous abdominal surgery (not caesarean section)</li> </ul>
Current gestation number	(number)
Abortions	(number)
Birth	(number)
Previous caesarean section	S/N
Live births	(number)
Stillborn children	(number)
Date of last menstruation	(date)
Estimated birth date	(Data to be calculated)

**Table 2** Data collected related to the current state of the pregnant woman

Prenatal care		
General data		
Blood pressure (in mm of Hg)	Systolic blood pressure	<ul style="list-style-type: none"> <li>• &lt; 130</li> <li>• 131–139</li> <li>• 140–159</li> <li>• &gt; 160</li> </ul>
	Diastolic blood pressure	<ul style="list-style-type: none"> <li>• &lt; 80</li> <li>• 81–89</li> <li>• 90–109</li> <li>• &gt; 110</li> </ul>
Weight	(number)	
Height	(number)	
Body mass index	(Calculated based on weight and height)	

**Table 3** Data collected related to the ultrasound results

Prenatal care	
Echography	
Number of fetuses	<ul style="list-style-type: none"> <li>• Unique</li> <li>• Twin</li> <li>• Triple</li> </ul>
Fetal heartbeat	Present/Absent
Position	<ul style="list-style-type: none"> <li>• Cephalic</li> <li>• Breech</li> <li>• Transverse</li> <li>• Not applicable</li> </ul>
Fetal biometrics	<ul style="list-style-type: none"> <li>• CRL (mm) [1st trimester]</li> <li>• DBP (mm) [2nd and 3rd trimester]</li> <li>• CC (mm) [2nd and 3rd trimester]</li> <li>• CA (mm) [2nd and 3rd trimester]</li> <li>• LF (mm)</li> </ul>
Fetal age	Hadlock formula
Placenta	<ul style="list-style-type: none"> <li>• Fundal</li> <li>• Previous</li> <li>• Posterior</li> <li>• Low insertion</li> <li>• Previous suspicion</li> </ul>
Amniotic fluid	(number)

Finally, specialists can review each case, including ultrasound images, and recommend actions such as modifying treatment or repeating ultrasounds. They also evaluate the ultrasounds within 5 levels: diagnostic quality, almost all acceptable ultrasounds, one acceptable ultrasound, necessary to repeat the ultrasound examination or not assessable. They can also assess whether the decisions made by the nursing staff (give crazy treatment, refer and transfer urgently ...) are consistent with the tests and ultrasound results.

## 6 Adaptation to be Able to Use the Offline System

Once the forms were implemented, we worked on the system to comply with the rest of the identified requirements. An essential adaptation was to modify a synchronism module that allows collecting and recording data offline. Information synchronisation is necessary when multiple devices use the same platform, each with its local database. If you want to share information between devices, that means copying the data from one database to another. Then you need to use some synchronisation mechanism to control the information that needs to be copied and thus avoid duplicates. OpenMRS has an optional synchronisation module [Sync OpenMRS Wiki Module <https://wiki.openmrs.org/display/docs/Sync+Module>]. This module allows

**Table 4** Data collected related to the blood tests results

Prenatal care	
Blood tests	
HBV screening (Hepatitis B)	<ul style="list-style-type: none"> <li>• Negative</li> <li>• Positive without confirmation</li> <li>• Confirmed positive</li> <li>• It was not done</li> </ul>
Syphilis screening	<ul style="list-style-type: none"> <li>• Negative</li> <li>• Positive without confirmation</li> <li>• Confirmed positive</li> <li>• It was not done</li> </ul>
HIV screening	<ul style="list-style-type: none"> <li>• Negative</li> <li>• Positive without confirmation</li> <li>• Confirmed positive</li> <li>• It was not done</li> </ul>
Hemoglobin	(number)
Glucose	(number)
Urine sample	<ul style="list-style-type: none"> <li>• It was not done</li> <li>• Negative</li> <li>• Leukocytes: Present/Absent</li> <li>• Nitrites: Positive/Negative</li> <li>• Proteins: 1+/2+/3+/Negative</li> </ul>
Diagnosis	<ul style="list-style-type: none"> <li>• Within normality</li> <li>• Abnormal ultrasound finding</li> <li>• Positive screening that requires confirmation</li> <li>• ITU</li> <li>• Anaemia</li> <li>• Other (to be filled)</li> </ul>
Treatment	<ul style="list-style-type: none"> <li>• None</li> <li>• Antibiotic therapy</li> <li>• Ferrotherapy</li> </ul>

**Table 5** Data collected in case of referral

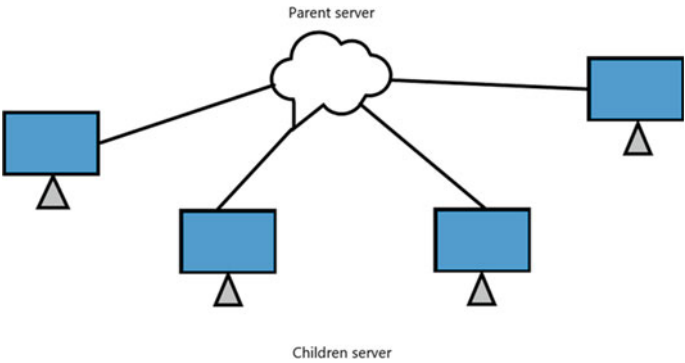
Prenatal care	
Referral	
Referral	<ul style="list-style-type: none"> <li>• No</li> <li>• Yes: Health Center</li> <li>• Yes: District Hospital</li> <li>• Yes: National Hospital</li> </ul>
Urgent transfer	<ul style="list-style-type: none"> <li>• No</li> <li>• Yes: Health Center</li> <li>• Yes: District Hospital</li> <li>• Yes: National Hospital</li> </ul>

**Table 6** Data collected related to the labour

Birth	
Delivery route	<ul style="list-style-type: none"><li>• Vaginal</li><li>• Caesarean section</li></ul>
Newborn status	<ul style="list-style-type: none"><li>• Alive</li><li>• Dead</li><li>• Abortion</li></ul>
Place of birth	<ul style="list-style-type: none"><li>• Domicile</li><li>• Health Center</li><li>• District Hospital</li><li>• National Hospital</li></ul>
Person who attended the birth	<ul style="list-style-type: none"><li>• Midwife</li><li>• Facilitator</li><li>• Doctor</li><li>• Nurse</li><li>• Familiar</li><li>• Single</li><li>• Unknown</li></ul>
Feedback	

you to synchronise the databases of different devices, choosing what type of information you want to synchronise. Thanks to this module, the complete OpenMRS system can be installed on several laptops so that you can work offline on each computer. When there is an internet connection, the information on each laptop is synchronised with a central server, which we call Parent. By contrast, the laptops you work with online are called Children. We can see this architecture in the following Fig. 1.

The limitation of the OpenMRS synchronism module is that it did not allow synchronising images stored in the database. To solve this problem, it was necessary to modify this module. With the modifications, the images can now be synchronised like the rest of the data.



**Fig. 1** This illustration shows us how the different child servers connect to the central server or parent

## 7 Definition of Roles and Permissions

In any information system, the configuration of user roles is critical since it allows us to limit the actions they can perform on the platform. OpenMRS allows defining differentiated roles for each group of users within the platform using privileges. For the Healthy Pregnancy project, three types of roles were defined:

- **Specialist:** The specialists will be the gynaecologists who will supervise the medical meetings and the information collected in these meetings. These gynaecologists will oversee reviewing the data collected by the providers and detecting possible errors or bad practises during the consultations. The interaction of this group with the platform will be minimal since its primary function will be to supervise, and specialists will have access to patients and medical meetings. Still, they will not be able to change the information collected by providers.
- **Provider:** Providers will be nurses who will travel to communities in areas far from health centres to consult pregnant women. They will oversee consultations with pregnant women, including ultrasounds and rapid tests. They will be able to record the information in the system when they perform the care or enter it later if they collect the data on paper.
- **Administrator:** The administrators will be the technicians developing and maintaining the platform. These users will have access to the entire platform and will be able to modify the access permissions for each role.

## 8 Encounter Alerts Module

The alert generation module allows specialists and suppliers to exchange information. Thanks to this module, gynaecologists in charge of reviewing cases can contact them to send comments on the cases reviewed.

This module was introduced in OpenMRS to fill a gap in communication between suppliers and specialists since, to perform their function well, specialists needed a continuous communication channel with suppliers. To implement this channel, a drop-down was added at the top of the forms and a text field at the bottom. The specialist will fill in the text field with comments, indicating the necessary corrections in the form and will use the drop-down to communicate to the nurse the result of the review, which can take three values:

- **Completed and pending review:** The specialist marks this option on the form when he has reviewed the meeting and requires the nurse to perform some tasks.
- **Completed and pending urgent review:** When the form filled out by the provider has serious errors that can directly affect the patient and require immediate action
- **Filled in and confirmed:** The specialist marks this option to confirm that the form does not contain any errors.

8.1 Results of Using the Tool

In this section, the data recorded with the Healthy Pregnancy SIS are analysed to study the tool’s degree of adoption. The years analysed go from September 2014 to December 2016, the first years of the health information system implementation. To begin with, in the following graph, we can visualise the number of patients registered throughout the years 2014, 2015 and 2016 (Fig. 2).

At the end of 2016, we found 10,108 patients registered in the system over all years. As you can see, the number of registered patients has increased considerably over the years of study.

To check if the records kept in the project’s SIS are complete, the fields considered necessary by the gynaecologists are analysed. The following tables (Tables 7 and 8) show the percentage of filling in various fields.

The filling percentage is very high and has increased over the years, indicating that the health personnel accepted the platform. Some fields, such as the estimated lifetime, were always filled. Instead, the field with the most gaps was the Hepatitis

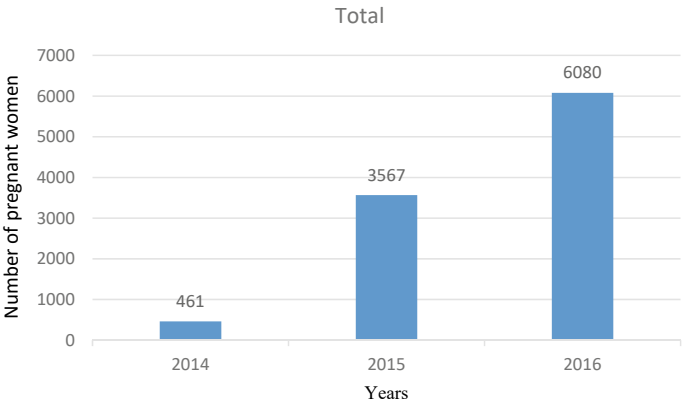


Fig. 2 Number of pregnant women registered in the system from 2014 to 2016

**Table 7** This table shows some tests performed on the pregnant woman and their corresponding filling by year: Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Body Mass Index (BMI) and Estimated Lifetime of the fetus (obtained from the ultrasounds)

Year	Total	SBP	DBP	BMI	Estimated life-time
2014	461	357	356	384	461
2015	3567	3496	3496	3409	3567
2016	6080	6024	6023	6046	6080
Total	10,108	9877	9875	9839	10,108
% filled	100.00	97.71	97.69	97.34	100.00

**Table 8** This table shows the results of blood tests performed on the pregnant woman and their corresponding filling by year: Hb (Hemoglobin), HIV (AIDS), Syphilis, Glucose and HBV (Hepatitis B)

Year	Total	Hb	HIV	Syphilis	Glucose	HBV
2014	461	340	437	441	338	441
2015	3567	3214	3522	3521	2863	3526
2016	6080	5546	6067	6063	5586	6065
Total	10,108	9100	10,026	10,025	8787	10,032
% filled	100.00	90.03	99.19	99.18	86.93	99.25

B (HB) test. The only data recorded below the average were glucose and HB tests, which may be due to the lack of punctual strips to do such tests.

In addition to registering cases, the SIS was to be used to enable specialists to review cases. Table 9 shows that most of the patients registered in the system have ultrasounds evaluated by specialists, except for 69 women.

Finally, having a systematised and reliable registry of prenatal controls has allowed the promoters of the project to publish two scientific articles [4, 22] and a doctoral thesis [23]. These articles analyse from a clinical point of view the data collected and identify which are the most frequent problems during pregnancy in the intervention areas: non-cephalic presentation was found in 14.87% of the pregnant women attended from week 32 onwards; 20 patients were referred for non-evolutive gestation; an 11.08% prevalence of anaemia was detected; and urine infections were diagnosed in 16.43% of the cases. With these data, it was shown that the project contributes to identifying obstetric risks and a possible trend in the reduction of maternal mortality was identified. The following section will discuss some of the keys that led to this result.

**Table 9** Table that collects the different evaluations by the specialists: A—Ultrasounds with diagnostic quality; B—An ultrasound acceptable; C—Almost all acceptable ultrasounds; D—It is necessary to repeat ultrasound examination; E—Not valuable

Year	A	B	C	D	E	Total
2014	122	34	281	18	6	461
2015	1743	365	1284	145	30	3567
2016	2876	520	2532	119	33	6080
Total	4741	919	4097	282	69	10,108



## 9 Discussion

The results show that the health information system has been used intensively for the follow-up of pregnant women in the Healthy Pregnancy project. The success of the proposed solution can be related to the following aspects:

- The system was designed in coordination with the main actors involved in the project, which brought together an essential experience in using health information systems in the rural environment.
- The selected technology made it possible to implement the forms in a format and language like the prenatal control forms used in rural areas on paper, facilitating their use by health personnel.
- The system could be used without the need for internet access, which is vital in areas of low connectivity to facilitate entry and prevent health personnel from being frustrated by connectivity problems.
- Using the SIS was part of an innovative initiative that improved care for pregnant women through equipment and training for rural health personnel. This was an incentive for rural staff to acquire new knowledge. In addition, one of the system's functionalities was to offer health personnel feedback on the prenatal controls they performed. In this way, health personnel improved their knowledge and skills about performing ultrasounds and prenatal controls, obtaining a direct benefit from using the system and, consequently, greater motivation to use it.

Another noteworthy aspect is that gynaecologists reviewed 99.31% of the cases. This data shows that it is feasible to carry out remote and deferred supervision. This functionality can also be oriented to the quality control of care and records, seeking to identify improvement aspects to strengthen the knowledge and practises of health personnel who perform prenatal controls.

On the other hand, the implemented system had some shortcomings that could be improved with currently available tools:

- Reports on the care provided had to be generated manually and needed updates every year to incorporate data from a new period. It would be desirable to automate the data analysis and use some tool that shows graphs, tables and/or maps with the different indicators. This would facilitate the use of the data by personnel with agility and less training.
- The system was designed to work on laptops, which was the equipment used with the probes at that time and is still used in some cases. However, it would be desirable for the system to work on Android devices since, currently, there are ultrasound probes that work with tablets or mobile phones with that operating system.
- The ultrasound images had to be uploaded to the platform one at a time and manually, and sometimes the health personnel erred in selecting the image that corresponded to each field. For example, they uploaded a picture of a femur when they should have uploaded one of the biparietal diameter.

- Local public institutions demand that the project information system be interconnected with other platforms used in the public health system. Currently, health staff can extract tables with the information that other platforms need, but that process is manual.
- The interface of the OpenMRS version is outdated and less intuitive than current versions of that tool or others currently available.

## 10 Conclusions

This article analyses the design, implementation, and subsequent adoption of a free software tool, specifically OpenMRS, as a health information system for a project to improve prenatal controls in rural areas.

Key to the design process was the participation of local institutions, including end-users of the system. It was also critical to have a highly configurable tool that allowed specific modules to be adapted to the project's needs, as with the image transfer module. All this was done when free software information systems still offered limited image-handling tools.

The information system began to be used in September 2014 and is currently operational. However, the data analysed focuses on the period from September 2014 to December 2016, with more than 10,000 registered attentions. The analysis of the recorded data offered in the results chapter shows the progressive acceptance of the tool in that period. At the end of the period, many fields show a record close to 100%. On the other hand, thanks to the information system, gynaecologists could review 99.31% of the cases, which was another of the system's objectives. These results show that the tool has had a positive acceptance and use. In addition, the data obtained have made it possible to publish two scientific articles and a doctoral thesis.

In analysing the case, some limitations of the tool have also been identified. These limitations can be addressed by implementing the system with new versions of OpenMRS or with alternatives, such as DHIS2, that have evolved enormously since the Healthy Pregnancy project was implemented. Work is currently underway on a review of the tools available today to implement and operationalise a new version of the health information system of the Healthy Pregnancy project.

**Acknowledgements** We acknowledge the support of the Health Directorates of the Alta Verapaz and San Marcos Departments. This work would not have been possible without the collaboration of the Directorate Representatives and the staff nurses at the public health facilities. The work carried out by the personnel of the EHAS Foundation, the Tulasalud Association, and the Tula Foundation was also crucial. We also acknowledge the support of the Spanish Agency for International Development Cooperation (AECID) with the project 2021/PRYC/000638.

## References

1. Home | Sustainable Development (n.d.). Retrieved 27 June 2022 from <https://sdgs.un.org/>
2. Aroche S (2021) Situación Epidemiológica de muerte materna de enero a septiembre de 2021
3. United Nations (2015) The Millennium Development Goals Report. United Nations, 72. 978-92-1-101320-7
4. Crispín Milart PH, Diaz Molina CA, Prieto-Egido I, Martínez-Fernández A (2016) Use of a portable system with ultrasound and blood tests to improve prenatal controls in rural Guatemala. *Reprod Health* 13(1):1–8
5. Prieto-Egido I, Simó-Reigadas J, Liñán-Benítez L, García-Giganto V, Martínez-Fernández A (2014) Telemedicine networks of EHAS Foundation in Latin America. *Frontiers Public Health* 2:1–9. <https://doi.org/10.3389/fpubh.2014.00188>
6. Hoxha K, Hung YW, Irwin BR, Grépin KA (2020) Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: a systematic review. *Health Inform Manage J*. <https://doi.org/10.1177/1833358320928729>
7. Dehnavieh R, Haghdoost A, Khosravi A et al (2018) The district health information system (DHIS2): a literature review and meta-synthesis of its strengths and operational challenges based on the experiences of 11 countries. *Health Inform Manage J* 48(2):62–75
8. Thomas J, Barry MA (2018) PRISM: performance of routine information system management series. MEASURE Evaluation, Chapel Hill
9. WHO (2016) Ebola situation report, June 2016. World Health Organization, June, 1–2. <http://apps.who.int/ebola/ebola-situation-reports%0A> [https://apps.who.int/iris/bitstream/handle/10665/208883/ebolasitrep\\_10Jun2016\\_eng.pdf;jsessionid=F18A12FCE559B4AE97FC43BC1907961?sequence=1](https://apps.who.int/iris/bitstream/handle/10665/208883/ebolasitrep_10Jun2016_eng.pdf;jsessionid=F18A12FCE559B4AE97FC43BC1907961?sequence=1)
10. Moon S, Sridhar D, Pate MA, Jha AK, Clinton C, Delaunay S, Edwin V, Fallah M, Fidler DP, Garrett L, Goosby E, Gostin LO, Heymann DL, Lee K, Leung GM, Morrison JS, Saavedra J, Tanner M, Leigh JA, Hawkins B, Woskie LA, Piot P (2015) Will Ebola change the game? Ten essential reforms before the next pandemic. The report of the Harvard-LSHTM Independent Panel on the Global Response to Ebola. *Lancet* (London, England) 386(10009):2204. [https://doi.org/10.1016/S0140-6736\(15\)00946-0](https://doi.org/10.1016/S0140-6736(15)00946-0)
11. Abir M, Mostashari F, Atwal P, Lurie N (2012) Electronic health records critical in the aftermath of disasters. *Prehosp Disaster Med* 27(6):620–622. <https://doi.org/10.1017/S1049023X12001409>
12. Oza S, Jazayeri D, Teich JM, Ball E, Nankubuge PA, Rwebembera J, Wing K, Sesay AA, Kanter AS, Ramos GD, Walton D, Cummings R, Checchi F, Fraser HS (2017) Development and deployment of the OpenMRS-Ebola electronic health record system for an Ebola Treatment Center in Sierra Leone. *J Med Internet Res* 19(8). <https://doi.org/10.2196/JMIR.7881>
13. Allen C, Jazayeri D, Miranda J, Biondich PG, Mamlin BW, Wolfe BA, Seebregts C, Lesh N, Tierney WM, Fraser HSF (2007) Experience in implementing the OpenMRS medical record system to support HIV treatment in Rwanda. *Stud Health Technol Inform* 129(Pt 1):382–386. <https://europepmc.org/article/med/17911744>
14. Thompson A, Castle E, Lubeck P, Makarfi PS (2010) Experience implementing OpenMRS to support maternal and reproductive health in Northern Nigeria. *Stud Health Technol Inform* 160(PART 1):332–336. <https://doi.org/10.3233/978-1-60750-588-4-332>
15. OMS (2010) Grupo de Investigación del Estudio de Control Prenatal de la OMS. [http://whqlib.doc.who.int/hq/2001/WHO\\_RHR\\_01.30\\_spa.pdf](http://whqlib.doc.who.int/hq/2001/WHO_RHR_01.30_spa.pdf)
16. Hospital EHR/Inpatient EHR (2019, febrero 9) Medsphere. <http://www.medsphere.com/open-vista>
17. Home (2018, noviembre 26) DHIS2. <https://dhis2.org/>
18. OpenEMR. <http://www.open-emr.org/>
19. OpenMRS. Contenidos generales. <http://openmrs.org/>
20. MedCLipse. <http://www.medclipse.ch/>

21. Dongarwar D, Salihi HM (2019) Influence of sexual and reproductive health literacy on single and recurrent adolescent pregnancy in Latin America. *J Pediatr Adolesc Gynecol* 32(5):506–513
22. Crispín Milart PH, Prieto-Egido I, Díaz Molina CA, Martínez-Fernández A (2019) Detection of high-risk pregnancies in low-resource settings: a case study in Guatemala. *Reprod Health* 16(1):1–8
23. Milart PC (2018) Ecografía portátil y cribado de sangre y orina para el control de gestantes en zonas rurales de países en desarrollo: estudio de caso en Guatemala. Doctoral dissertation, Universidad Rey Juan Carlos

# Adapting Atmospheric Chemistry Components for Efficient GPU Accelerators



Christian Guzman Ruiz, Matthew Dawson, Mario C. Acosta, Oriol Jorba, Eduardo Cesar Galobardes, Carlos Pérez García-Pando, and Kim Serradell

**Abstract** Atmospheric models demand a lot of computational power, and solving the chemical processes is one of its most computationally intensive components. This work shows how to improve the computational performance of the Multiscale Online Nonhydrostatic Atmosphere Chemistry (MONARCH), a chemical weather prediction system developed by the Barcelona Supercomputing Center. The model implements the new flexible external package chemistry across multiple phases (CAMP) for the solving of gas- and aerosol-phase chemical processes that allows multiple chemical processes to be solved simultaneously as a single system. We introduce a novel strategy to simultaneously solve multiple instances of a chemical mechanism, represented in the model as grid cells, obtaining a speedup up to  $9\times$  using thousands of cells. In addition, we present a GPU strategy for the most time-consuming function of CAMP. The GPU version achieves up to  $1.2\times$  speedup compared to CPU. Also, we optimize the memory access in the GPU to increase its speedup up to  $1.7\times$ .

**Keywords** Chemistry · Parallelism and concurrency · Performance

## 1 Introduction

Atmospheric models can be defined as a mathematical representation of dynamical, physical, chemical, and radiative processes in the atmosphere [9]. They provide valuable information on the nature of real-world phenomena and systems, with many applications in science and engineering. However, they are often associated with large computational costs because of their complexity [6].

---

C. G. Ruiz (✉) · M. C. Acosta · O. Jorba · C. P. García-Pando · K. Serradell  
Barcelona Supercomputing Center, Barcelona, Spain  
e-mail: [christian.guzman@bsc.es](mailto:christian.guzman@bsc.es)

M. Dawson  
National Center for Atmospheric Research (NCAR), Boulder, CO, USA

E. C. Galobardes  
Universitat Autònoma de Barcelona, Bellaterra, Spain

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_11](https://doi.org/10.1007/978-981-99-3091-3_11)

Due to the high computational cost, these models often divide their load into multiple parallel processes through domain decomposition [12]. This technique divides a region into smaller regions, which for now on will be referred to as cells. The model assigns collections of cells to independent threads for parallel solving of the many physical and chemical processes in the atmosphere.

To make this assignment, atmospheric models make use of parallel programming interfaces like MPI, OpenMP, OpenACC, and CUDA. MPI is the most used tool to distribute work across independent nodes in supercomputers. In addition, a model can use MPI to compute multiple cells in each computer node and then use another parallel approach to further divide the load across individual CPU or GPU threads. Studies using the CUDA language have reported high speedups from parallelizing a demanding component of atmospheric models, the chemical kinetics module. For example, a study using the EMAC Earth system model developed a CUDA version of the kinetic preprocessor library (KPP) reporting a speedup of up to  $20.4\times$  against a single-thread execution [3]. Another research of simple chemical kinetics processes developed two different solver methods designed specifically for CUDA execution, the Runge-Kutta-Cash-Karp (RKCK) and Runge-Kutta-Chebyshev (RKC) [10]. This study achieved a speedup of up to  $59\times$  compared with a single-thread execution. The large difference in speedup between these studies highlights the importance of developing new methods specifically focused on parallel GPU execution and the impact of translating classic CPU-oriented methods to GPU. However, the GPU-specific methods are harder to adapt to atmospheric models and are often only tested for specific types of chemical equations. In contrast, CPU-based solvers, like KPP-GPU, are already prepared to run in atmospheric models with the same types of chemical equations currently solved by purely CPU-based code.

The performance difference among these methods is primarily derived from their different parallelization methods. CPU-based solvers divide the load by domain decomposition, where each GPU-thread solves an individual small system. The efficiency of this approach compared with CPU-only execution has been demonstrated multiple times [8]. However, thousands of domain grid cells are required to achieve a significant speedup. An alternative approach applied by the GPU-focused methods is to parallelize explicitly for chemistry equations. This approach allows for a greater degree of parallelization as each grid cell has multiple chemical reactions to solve. Also, these methods typically apply solving algorithms that are specifically designed to execute more steps in parallel. These two characteristics of GPU-focused methods result in better performance results compared with translation-based approaches.

In this work, we present results from solving simultaneously multiple cells in a single-thread execution. Also, we tested a GPU implementation following this strategy on the most time-consuming function of the chemical module. This approach is a combination of the CPU-based and GPU-specific methods, as we still use a CPU-based solver but use specific GPU techniques for one solver function to achieve a high degree of parallelization.

The implementations proposed are tested in the module chemistry across multiple phases (CAMP). CAMP is developed to treat gas and aerosol chemical reactions in

a single system, thus simplifying optimization and introduction of new multi-phase chemistry [7]. It is integrated into the multiscale online nonhydrostatic atmosphere chemistry (MONARCH) model [4].

The remainder of this document is organized into sections according to these objectives. In Sect. 2, we provide a brief description of MONARCH and CAMP, plus presenting the most time-consuming function of CAMP. In Sect. 3, we present the GPU implementation of this function, an optimization to reduce GPU accesses, and the Multi-cells implementation for the whole CAMP module. In Sect. 4, we define the hardware and software environment used. Section 5 shows the result of the implementations presented. Finally, Sect. 6 concludes the work and overviews possible future work.

## 2 Background

MONARCH couples an online meteorological driver with gas and aerosol continuity equations to solve atmospheric chemistry processes in detail. The model is designed to account for feedbacks among gases, aerosol particles, and meteorology. This work focuses on its chemical components, which can consume up to 80% of the model execution time. From the chemistry solvers available in MONARCH, we choose to work with the most and promising option, the framework CAMP.

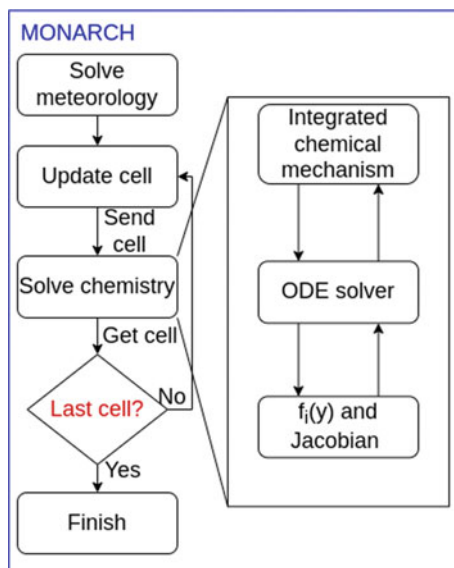
CAMP is a novel framework permitting run-time configuration of chemical mechanisms for mixed gas- and aerosol-phase chemical systems (including gas- and aerosol-phase reactions and mass transfer), available at Github [2]. It also allows an abstract non-fixed representation of aerosols that can be configured at run-time, describing the life cycle of mineral dust, sea-salt, black carbon, organic matter (both primary and secondary), sulfate, nitrate, and ammonium aerosols. It computes a greater selection of types of chemical processes than the other MONARCH options. Thus, applying our implementations into CAMP affects more part of the chemistry time.

The chemical reactions in CAMP can include both integer parameters (e.g., array indices, stoichiometric coefficients, ionic charge, etc.) and floating-point parameters (e.g., conversion factors, rate parameters, etc.). The set of chemical species concentrations ( $y$ ) is named the *state* array, and the set of partial derivatives of these species with respect to time ( $f$ ) is named the *deriv* array.

After the data is read, CAMP predicts future concentrations using the external ODE solver CVODE [5]. CVODE solves the time-dependent equation ( $y' = f(t, y)$ ) using the CAMP-provided set of derivatives ( $f(y)$ ) stored in the *deriv* array. CVODE also uses a Jacobian matrix provided by CAMP. From the matrix structure options that CVODE offers, we choose the SPARSE structure [11] to store the Jacobian, as this is a good choice for Jacobian structures with few non-zero elements, as is the case for many chemical mechanisms.

Either derivative and Jacobian functions have very similar input and output, following the same structure. The only difference is the structure where we store the

**Fig. 1** MONARCH overall flow diagram with CAMP as chemistry solver

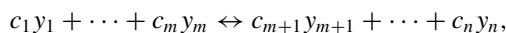


data (an array for the derivative and a sparse matrix for the Jacobian) and some extra linear operations. So, we only need to analyze one of them since we can extrapolate the optimization ideas and techniques.

Inside MONARCH, CAMP is required to solve chemistry multiple times. One time for each MONARCH time-step and cell. A cell represents a volume of the atmosphere, the collection of which composes a 3-dimensional grid that represents the whole atmosphere. The number of cells depends on the user-selected MONARCH configuration. MONARCH typically computes a large number of cells, over a large surface area with high precision. Each cell has its own *state*, which in terms of chemical processes is independent of other cell *state* values. In Fig. 1, we summarize the flow described in a diagram.

The CAMP functions executed during the solving take a considerable execution time. We configured an experiment with a CB05 chemical mechanism to measure this impact. The experiment results show that CVODE occupies 70% of the total execution time, and derivative and Jacobian around 30%. Despite being small functions compared to the whole ODE solver, the derivative and Jacobian have a relevant impact on general performance, being the derivative generally more time-expensive than Jacobian. So, in a similar way as selecting the chemistry component of MONARCH, we choose to work around the derivative for analyzing our GPU implementation and search for a relevant reduction of the model execution time.

In general, chemistry models try to predict future concentrations of a set of chemical species by solving a set of ordinary differential equations that represent the reactions that compose a chemical mechanism. Reactions take the general form:





where species  $y_i$  is a participant in the reaction with stoichiometric coefficient  $c_i$ . The rate of change for each participating species  $y_i$  with respect to reaction  $j$  is given by

$$\left(\frac{dy_i}{dt}\right)_j = \begin{cases} -c_i r_j(\mathbf{y}, T, P, \dots) & \text{for } i \leq m \\ c_i r_j(\mathbf{y}, T, P, \dots) & \text{for } m < i \leq n \end{cases},$$

where the rate  $r_j$  of reaction  $j$  is an often complex function of the entire model state (including species concentrations  $\mathbf{y}$ , environmental conditions, such as temperature,  $T$ , and pressure,  $P$ , physical aerosol properties, such as surface area density and number concentration, etc.). The overall rate of change for each species  $y_i$  at any given time is, thus,

$$f_i \equiv \frac{dy_i}{dt} = \sum_j \left(\frac{dy_i}{dt}\right)_j,$$

where  $\mathbf{f}$  is referred to as the derivative of the system throughout this document.

Then, in the derivative function, we multiply the rate constants saved on the reaction parameters array with the corresponding concentrations on the *state* array, filling the next concentration array (*deriv*). This operation is done for each reaction, adding all the results obtained from the reactions in the corresponding place of the *deriv* array. So, we can say that each reaction adds a contribution to the *state* concentrations, increasing or decreasing the value.

### 3 Implementations

The multi-cells implementation groups the different input data from each cell into a single data structure to be computed. The MONARCH workflow described in Fig. 1 is updated to Fig. 2a. The cells loop disappear inside the solving internal functions, avoiding the process of updating the input data from cells and re-initializing the ODE solver. As an example, the derivative equation is updated as following:

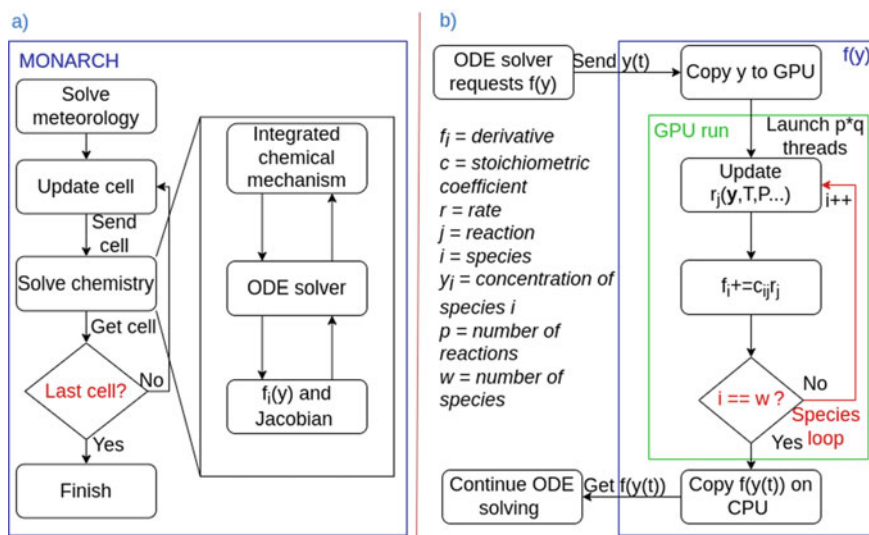
$$f_i \equiv \frac{dy_{ik}}{dt} = \sum_j \left(\frac{dy_{ik}}{dt}\right)_j$$

where  $y_{ik}$  refers to the species  $y_i$  from cell  $k$ .

Our GPU strategy is the parallelization of each reaction data packet. In Fig. 2b, we can see the resultant GPU-based derivative flow diagram.

We compute the sum of contributions to  $\mathbf{f}$  by using the CUDA operation *atomicAdd*. This function avoids a possible thread overlapping on updating the same variable. This interference can be produced by reactions with common species between them.

Reaction data is allocated on global memory at the initialization of the program. To send and receive from the GPU the rest of the data (*state* array), we check first the



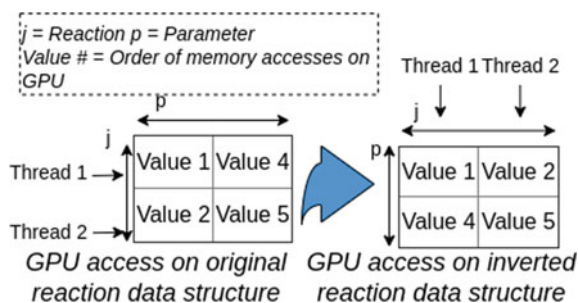
**Fig. 2** On the left (a): comparison of original and multi-cells overall workflows from the MONARCH point of view. On the right (b): derivative workflow diagram for GPU execution.

size of this array. If it contains few data variables, then *state* is passed as a function parameter, taking advantage of the constant memory. Otherwise, the data is copied through a direct transfer to the global memory.

The number of GPU threads initialized is equal to the number of reactions. Another relevant GPU parameter, the number of blocks per threads, is configured to the maximum available for the GPU used (1024 threads/block). Lower configurations of threads/blocks do not show a performance improvement for our tests. Due to the possibility of using a GPU with less capacity in future, we add a run-time checking of GPU hardware specifications to ensure the correct execution of the program regardless of the GPU used (for example, avoid demanding more threads than the GPU limit).

In the still CPU-based implementation, all the reaction data packets are initially stored consecutively in memory. Then, the parallelization by reactions results in each thread accessing no-consecutive values of the reaction data structure. We reordered this structure to follow a sequential reading of the data in the GPU. The first reaction parameters accessed are stored consecutively in the reaction data structure, and so on. Figure 3 illustrates the changes in the data packet structure, simulating the structure as a matrix where initially the rows are the data packets and columns the parameter values.

**Fig. 3** Data structure inversion for GPU derivative. “Value” numbers represent the GPU memory arrangement and access order, “ $j$ ” the number of reactions, and “ $p$ ” the number of parameters



## 4 Test Environment

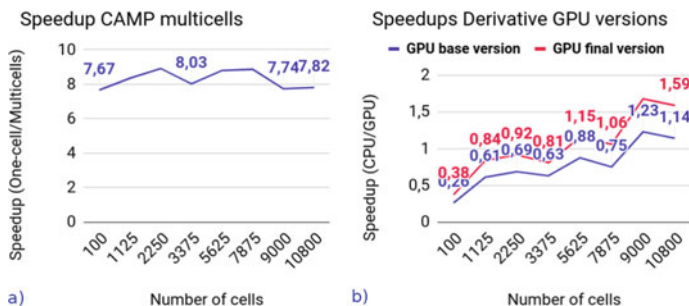
During the work, all the tests and executions were performed in the CTE-POWER cluster provided by the Barcelona Supercomputing Center (BSC) [1]. We used the compilers GCC version 6.4.0. and NVCC version 9.1, an IBM Power9 8335-GTH @ 2.4GHz and a GPU NVIDIA V100 (Volta) with 16GB HBM2.

We work around a basic chemical mechanism of 3 species, where species  $A$  generates  $B$  and  $C$  through 2 Arrhenius reactions.  $A$  is initialized at 1.0, while  $B$  and  $C$  are set to zero. Each cell has a small offset of 0.1 on the initial concentrations to generate different results. For example, at the first concentration value, we sum a 0.1 offset value, at the second 0.2, and so on till multi-cells species. The rest of variables like temperature, pressure, or reaction data parameters are initialized to the same values for all the cells.

## 5 Results

In Fig. 4a, we can see how multi-cells speed up CAMP a factor of  $8\times$  for multiple numbers of cells. Most of this speedups are produced by the reduction of solving iterations. In the one-cell case, the number of iterations scales linearly with the number of cells factor, while in the multi-cells case, the number of iterations is independent from the number of cells computed, keeping almost the same number of iterations for the number of cells. For example, one-cell takes around  $6e^6$  iterations to solve 10,000 cells (an average of 600 iterations per cell), while multi-cells take around 700 to solve all cells, with independence of the number of cells.

The GPU implementation speeds up the derivative function for a high number of cells. In Fig. 4b, we can see how for 10,000 cells the GPU version achieves  $1.2\times$  speedup. On the other hand, a lower number of cells slows down the function, shown as a speedup below  $1\times$  for less than 10,000 cells. We can also see that the optimization on memory access improves the overall speedup by a factor of  $1.3\times$  approximately for all numbers of cells.



**Fig. 4** On the left (a): CAMP speeds up using multi-cells optimization in front of the original one-cell version. On the right (b): speedup of base and final single-GPU versions compared to single-thread CPU versions. Final version applies the optimization on GPU memory access into the base version

We also compare the final GPU version against a CPU case parallelized with MPI, emulating the parallelization used in MONARCH. The number of MPI processes are configured to follow the proportion of GPUs used for GPU available. So, we use 40 MPI threads from the 160 available, like the GPU experiments presented use 1 GPU from the 4 available. We obtain that the GPU execution is  $3\times$  times slower than the MPI, but only because the time of data movements between CPU and GPU is taking near 90% of the GPU execution time. The GPU computation time is  $3.5\times$  times faster than the MPI time (0.04 s in GPU and 0.14 s for MPI). This data movement is produced by updating the species concentrations on each call to derivative. We can conclude that the GPU derivative function has a small computation load for data movement produced (reaction data, concentration values, etc.).

## 6 Conclusions

In this paper, we focused on improving the performance of CAMP for an execution in an atmospheric model environment like MONARCH. MONARCH simulations perform one CAMP simulation for each grid cell of the geographic simulation region for each MPI thread. These cells have no inter-dependencies during the chemistry solving; thus, they have potential to be parallelized by the GPU. However, the classical MONARCH implementation calls the CAMP solving process for each grid cell. In each cell iteration, the CAMP solving library (CVODE) needs to reinitialize its internal solving variables. Furthermore, to implement a GPU implementation over the cells, it would be necessary to translate the complete solving code to GPU format, which can be an exhaustive work. The first implementation presented in this study aims to solve these issues. This strategy is relatively novel in the atmospheric community and can be used as an example to speed up the model. We refer in the paper to this implementation using the name of multi-cells.

The multi-cells strategy groups the data for each cell into a single structure to be solved. The cells loop from MONARCH is moved into the internal solving functions of CAMP. The results show a considerable reduction of the calls to the derivative function. The solving module uses approximately the same number of iterations to solve all the cells than to solve a single cell. With respect to the improvement in execution time, the multi-cells implementation achieves near  $8\times$  speedup for all the cells tested, up to  $9\times$  speedup.

Next, we developed a CUDA version of the derivative function by parallelizing its reaction loop among GPU threads. The new version obtains near  $1.2\times$  speedup for 10,000 cells approximately. For a lower number of cells, the CPU version has better performance than GPU. The third implementation reorders the reaction data structure to improve its access in the GPU derivative version, increasing the GPU speedup by a factor of  $1.3\times$  for all the cells tested.

Finally, we inspect the time execution consumed on moving data between GPU and CPU. For 10,800 cells, this time on data movement takes 90% of the total time execution. Comparing the results with a 40 MPI process execution, the computation time for the GPU version is  $3.5\times$  faster. Thus, future work will focus on reducing GPU data movement by translating more CPU functions to the GPU, for example, the Jacobian or functions from the ODE solving and overlapping some CPU and GPU work. This should increase the computation performed on the GPUs and reduce data movement by transferring data only at the start and the end of the solving, reducing data movement during solver iterations. This can be done by parallelizing the next solver functions executed after or before the derivative calculation until all the solver would be executed in GPU. We also expect to explore load balancing the CPU and GPU using overlapping and asynchronous communication, since currently, the CPU is not performing any work during GPU execution. Lastly, we expect to evaluate the GPU-based chemistry solving in MONARCH, checking the impact for a variety of atmospheric experiments with an MPI implementation alongside the GPU-CUDA chemistry.

**Acknowledgements** This work was partially supported by funding from the grant BROWNING project (RTI2018-099894-BI00) funded by MCIN/AEI/10.13039/501100011033, the CAROL project (MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21), the Generalitat de Catalunya GenCat-DIUIE (GRR) (2017-SGR-313) and the AXA Research Fund through the AXA Chair on Sand and Dust Storms established at BSC. This work has also received funding from “Future of Computing Center, a Barcelona Supercomputing Center and IBM initiative (2020)”. Matthew Dawson has received funding from the European Union’s Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. 747048. This paper expresses the opinions of the authors and not necessarily those of the funding commissions. BSC co-authors acknowledge the computer resources at CTE-POWER, the technical support provided by the Barcelona Supercomputing Center, and the support from Partnership for Advanced Computing in Europe (PRACE) and Red Española de Supercomputación (RES).

## References

1. Support Knowledge Center @ BSC-CNS
2. CAMP (Nov 2021)
3. Alvanos M, Christoudias T (2017) GPU-accelerated atmospheric chemical kinetics in the ECHAM/MESSy (EMAC) Earth system model (version 2.52). *Geosci Model Dev* 10(10):3679–3693
4. Badia A, Jorba O, Voulgarakis A, Dabdub D, Pérez García-Pando C, Hilboll A, Gonçalves M, Janjic Z (2017) Description and evaluation of the multiscale online nonhydrostatic atmosphere chemistry model (NMMB-MONARCH) version 1.0: gas-phase chemistry at global scale. *Geosci Model Dev* 10(2):609–638
5. Hindmarsh AC, Brown PN, Grant KE, Lee SL, Serban R, Shumaker DE, Woodward C (2004) SUNDIALS: suite of nonlinear and differential/algebraic equation solvers. *ACM Trans Math Softw (TOMS)* 31:363–396
6. Bennett ND, Croke BWF, Jakeman A, Newham LTH, Norton JP (2010) Performance evaluation of environmental models
7. Dawson ML, Guzman C, Curtis JH, Acosta M, Zhu S, Dabdub D, Conley A, West M, Riemer N, Jorba O (2021) Chemistry Across Multiple Phases (CAMP) version 1.0: an integrated multi-phase chemistry model. [arXiv:2111.07436](https://arxiv.org/abs/2111.07436) [physics]
8. Haidar A, Abdelfattah A, Zounon M, Tomov S, Dongarra J (2017) A guide for achieving high performance with very small matrices on GPU: a case study of batched LU and Cholesky factorizations. *IEEE Trans Parallel Distrib Syst* PP:1–1
9. Jacobson MZ (2005) *Fundamentals of atmospheric modeling*, 2nd edn. Cambridge University Press, Cambridge
10. Niemeyer KE, Sung CJ (2014) Accelerating moderately stiff chemical kinetics in reactive-flow simulations using GPUs. *J Comput Phys* 256:854–871 [arXiv: 1309.2710](https://arxiv.org/abs/1309.2710)
11. Skeel RD (1986) Construction of variable-stepsizes multistep formulas
12. Tintó O, Acosta M, Castrillo M, Cortés A, Sanchez A, Serradell K, Doblas-Reyes FJ (2017) Optimizing domain decomposition in an ocean model: the case of NEMO. *Procedia Comput Sci* 108:776–785

# Frequency Interleaved DAC System Design: Fundamental Problems and Compensation Methods



Nagito Ishida, Koji Asami, Shogo Katayama, Anna Kuwana,  
and Haruo Kobayashi

**Abstract** This paper describes wideband digital-to-analog conversion circuits (DACs) using frequency interleaving architecture. Along with broadband communication standards, wideband DACs are required for wideband communication device measurement and test equipment. First, we explain the basic configuration and operation of the frequency interleaved DAC (FI-DAC) architecture. There, the digital input signal is divided into multiple bands. Then they are demodulated and provided to several sub-DACs, whose analog outputs are modulated and synthesized to the wideband analog output signal. Note that image components are generated by sub-DACs and modulations, which have to be removed by subband synthesis analog filters. The other fundamental problems of this architecture are signal attenuation by zero-th order hold of each sub-DAC output, phase nonlinearity characteristic of synthesis analog filters, group delay differences among subband channels, and phase discontinuity between adjacent subband channels. We examine compensation methods for these problems, and their effectiveness is confirmed with MATLAB simulation.

**Keywords** DAC · Frequency interleaving · Wideband signal generation · Modulation · Broadband

---

N. Ishida (✉) · S. Katayama · A. Kuwana · H. Kobayashi  
Gunma University, 1-5-1 Tenjin-Cho Kiryu, Gunma 376-8515, Japan  
e-mail: [t190d011@gunma-u-ac.jp](mailto:t190d011@gunma-u-ac.jp)

S. Katayama  
e-mail: [t15304906@gunma-u-ac.jp](mailto:t15304906@gunma-u-ac.jp)

A. Kuwana  
e-mail: [kuwana.anna@gunma-u-ac.jp](mailto:kuwana.anna@gunma-u-ac.jp)

H. Kobayashi  
e-mail: [koba@gunma-u-ac.jp](mailto:koba@gunma-u-ac.jp)

K. Asami  
Advantest Corporation, Tokyo, Japan  
e-mail: [koji.asami@advantest.com](mailto:koji.asami@advantest.com)

# 1 Introduction

Communication standards have been changing from 4 to 5G, and furthermore, 6G is being considered; they are becoming increasingly wideband. Therefore, wideband measuring instruments are required to measure and test these devices. As one of the methods for realizing a wideband ADC, interleaved ADCs using multiple channels such as time-interleaved ADC (TI-ADC) [1] and frequency-interleaved ADC (FI-ADC) [2–4] have been investigated for a long time. Recently, research has begun to apply the interleaving techniques to DACs, and similarly, there are two interleaving methods: time-interleaved DACs (TI-DACs) and frequency-interleaved DACs (FI-DACs).

The concept of the TI-DAC is to combine the output signals of multiple DACs whose output timings are shifted each other, and which are combined in the time domain [5, 6], whereas that of the FI-DAC is to combine the output signals of multiple DACs whose covering bandwidths are different each other in output, but sampling timings are the same [7, 8]. The TI-DAC has a limitation to increase the number of channels; analog switches for their output multiplexing are difficult to realize in case of the wideband DAC [9]. Therefore, it is difficult to realize a wideband TI-DAC by zero-th order hold. On the other hand, since the FI-DAC is not restricted by the sub-DAC zero-th order hold output characteristic, the number of channels can be increased for the wideband signal generation. There the removal of Nyquist image by interleaving has been also investigated [10, 11].

The FI-DAC is a novel concept proposed in recent decades, and several approaches and compensation methods have been proposed as follows: (i) The FI-DAC has the interdependencies among the bandwidth, the sample rate, the number of samples and the frequencies of the local oscillators (LOs). In order to balance the interdependent system parameters, a mathematical optimization approach is introduced in the form of two mixed-integer nonlinear optimization programs (MINLPs) [8]. (ii) A closed algebraic expression for the distribution of the data samples among the DACs is shown in order to realize guard bands suppressing DAC aliases and unused mixer side bands [12]. (iii) As their compensation method, a multiple-input multiple-output (MIMO) digital signal processing (DSP) algorithm to avoid the crosstalk between the frequency bands has been devised [13]. There DSP for frequency domain equalization utilizes a repetitive data sequence. (iv) A compensation method using a MIMO equalizer and a backpropagation algorithm by adapting its coefficients has been developed [14, 15].

In this paper, we investigate compensation for FI-DAC fundamental problems using digital and analog filters and adjusting the initial phases of the carriers. The organization of this paper is as follows. First, we explain the basic configuration and operation of the FI-DAC architecture. We show image component generation by modulation and DAC output zero-th order hold. Next, we explain signal attenuation by the DAC and phase characteristic of digital and analog filters. Then we describe their compensation methods using digital signal processing [16, 17].



**Table 1** Fundamental problems and their compensations for the FI-DAC system

Problems	Compensation methods
Signal attenuation by DAC zero-th order hold output	Applying an inverse sinc filter
Phase-nonlinearity characteristic of smoothing and synthesis analog filters	Applying all-pass filters
Differences in group delay among subband channels	Changing the sampling timing
Phase discontinuity between adjacent subband channels	Adjusting the initial phase of the carrier signal in digital signal processing

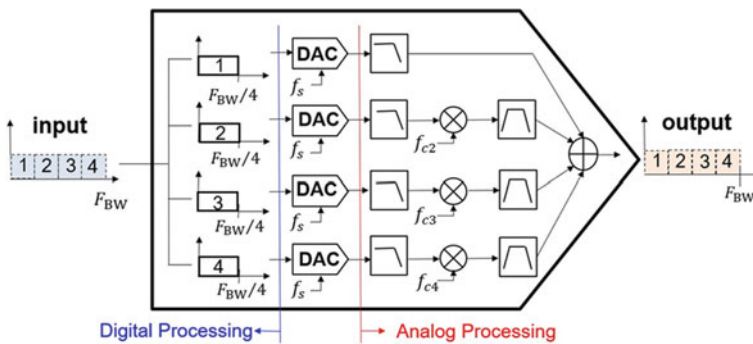
They are summarized in Table 1. Finally, MATLAB simulation results of these compensations are shown for their verification.

## 2 Principle and Structure of FI-DAC Architecture

### 2.1 FI-DAC Architecture

Figure 1 shows the investigated structure of the 4-channel FI-DAC. First, the input signal band is divided into four subband channels by digital signal processing. Second, each corresponding sub-DAC provides the output for each band. Third, these output signal bands are shifted back to the targeted signal band using analog modulation. Finally, the outputs of all subband channels are combined.

Here,  $F_{BW}$  denotes the input signal band,  $f_s$  does the sampling frequency of sub-DACs,  $k$  does the number of channels, and  $f_{ck}$  does the carrier frequency of the  $k$ -th subband channel. In this structure, we have the following relationship between the sampling frequency and the output signal band:

**Fig. 1** Structure of the investigated 4-channel FI-DAC architecture

$$\frac{f_s}{2} \geq \frac{F_{BW}}{4} \quad (1)$$

In case of M-subband channel, we have the following:

$$\frac{f_s}{2} \geq \frac{F_{BW}}{M} \quad (2)$$

## 2.2 Generation of Image Components

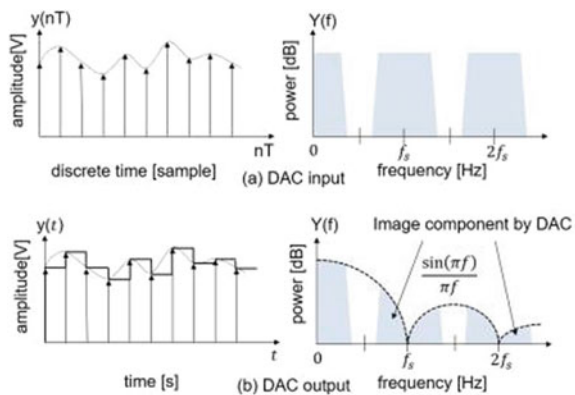
### Image Components by DAC Zero-th Order Hold Output

Figure 2 shows input and output signal waveforms and power spectrum of a zero-th order hold DAC. In Fig. 4, (a) shows the DAC input, while (b) does the DAC output. Image components are generated in the input, and their spectrum are mirrored with respect to  $f_s/2$ . Besides, the signal spectrum power has the attenuation tendency for higher frequency in output. Image components have to be removed by the following synthesis analog low-pass filter as shown in Fig. 3. This filter is called as smoothing analog filter.

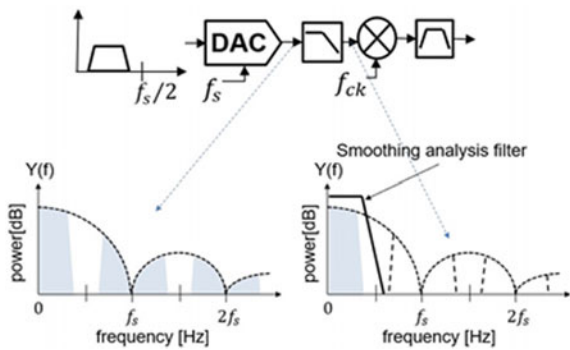
### Image Components Generated by Modulation

Figure 4 shows the process of generating image components by the amplitude modulation (AM) on the frequency axis. In Fig. 4, (a) shows the signal spectra, and (b) shows carrier spectra, while (c) shows the modulated signal spectra. The signal spectrum is shifted to the carrier frequency by the AM, and the image components appear in the lower side of the carrier frequency, as shown in Fig. 4c. In FI-DAC, they can be removed by the analog bandpass filter followed by the mixer as shown in Fig. 5.

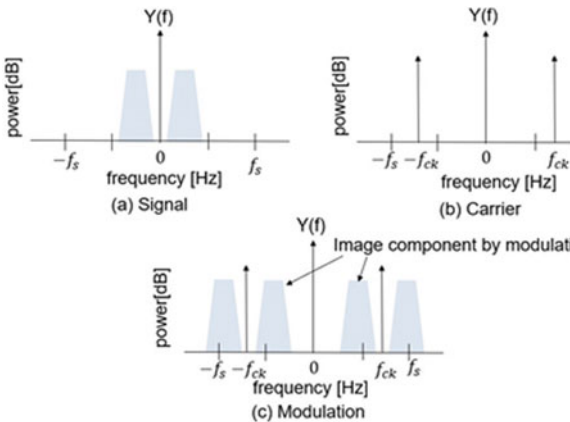
**Fig. 2** DAC output with zero-th order hold



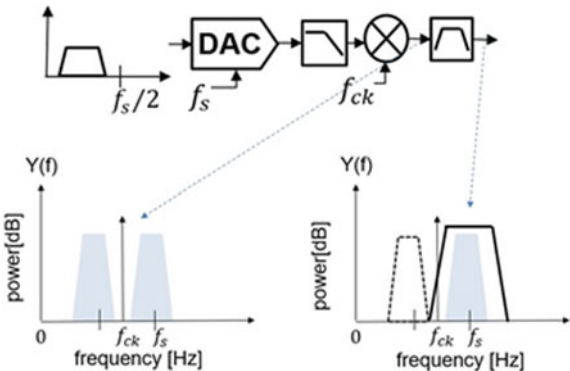
**Fig. 3** Removal of image components caused by DAC zero-th order output using smoothing analog filter

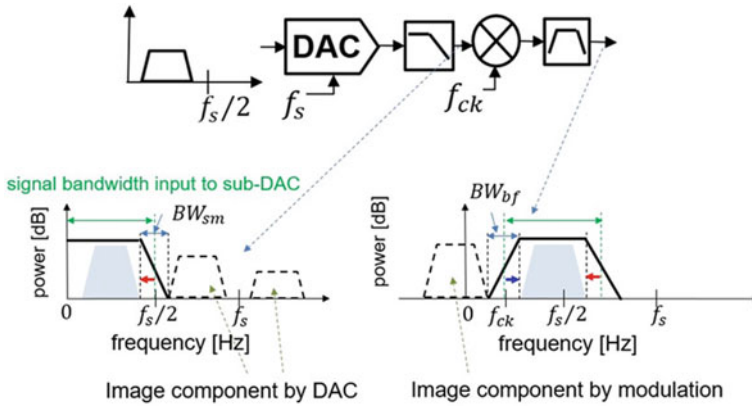


**Fig. 4** Image components generated by modulation



**Fig. 5** Removal of image components caused by modulation using the analog band-pass filter





**Fig. 6** Restriction of signal bandwidth input to sub-DAC by DAC zero-th order output and modulation

### Restriction of Signal Bandwidth Input to Sub-DAC

The input signal bandwidth to sub-DAC is between 0 and  $f_s/2$ . Since the image components shown in Figs. 2 and 4 degrade the overall FI-DAC output signal quality, they have to be removed by synthesis analog filters. Since the filter has the transition band between the signal component and the image component, the frequency interval corresponding to the transition bandwidth has to be secured. For this reason, the input band becomes narrower than that between 0 and  $f_s/2$ , as shown in Fig. 6. There  $BW_{sm}$  is the transition bandwidth of the smoothing analog lowpass filter and  $BW_{bf}$  is that of the synthesis analog bandpass filter. The image components are generated by the DAC, and they are mirrored with respect to  $f_s/2$ , so the signal bandwidth becomes narrower; it yields from  $f_s/2$  to  $BW_{sm}/2$ . The image components are generated by modulation, and they are mirrored with respect to  $f_{ck}$ , and the signal bandwidth becomes narrow; it is from 0 to  $BW_{bf}/2$ . Accordingly, the signal bandwidth is expressed as follows:

$$\frac{BW_{bf}}{2} \leq \text{signal bandwidth input to sub - DAC} \leq \frac{f_s}{2} - \frac{BW_{sm}}{2} \quad (3)$$

## 3 Subband Analysis Digital Signal Processing

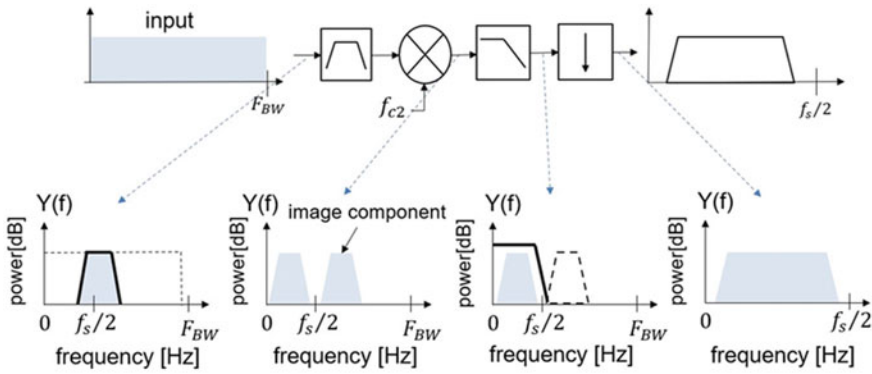
### Overall Block Diagram of Subband Analysis Digital Signal Processing

Figure 7 shows digital signal processing for the sub-DAC input. In Fig. 7, (a) shows the block diagram from the input to the sub-DAC input, while (b) shows the power spectrum of each processing. The input signal is split in frequency domain using analysis digital filters surrounded by the red frame in Fig. 7a, and the frequency

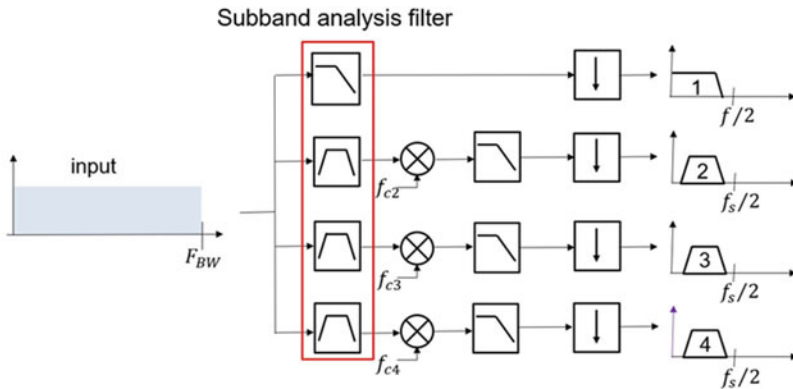
shift is performed for the bands to be within  $0 \sim f_s/2$ . However, the first-subband channel is already in the band  $0 \sim f_s/2$ , so the frequency shift is not required. The image components by frequency shift are generated at subband channels 2, 3 and 4, and they have to be removed by analysis digital filters. Finally, the sampling rate is converted from the sampling frequency of the input to the sampling frequency of the sub-DAC.

### Design of Subband Analysis Digital Filters

Figure 8 shows the design specifications of the subband analysis digital filters used in Fig. 7a. We see that each analysis digital filter has a transition band. Therefore, by overlapping the transition bands between adjacent filters, they are designed so that signals in the bands inside the dotted lines in Fig. 8 are not attenuated at the

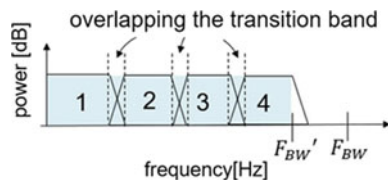


(a) Block diagram of subband analysis digital signal processing

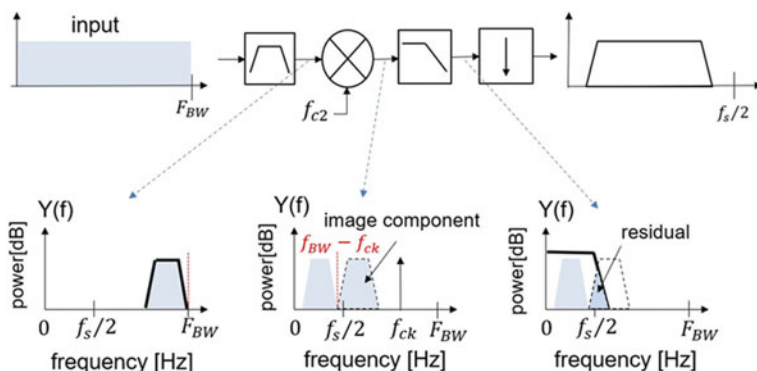


(b) Subband channel analysis digital signal processing

Fig. 7 Subband analysis digital filter processing



**Fig. 8** Design of subband analysis digital filters



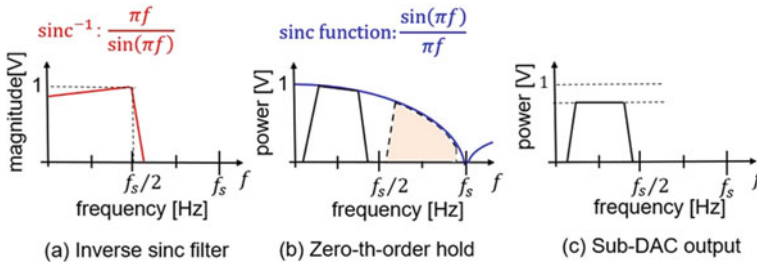
**Fig. 9** Restriction of subband analysis digital filter

final output. The signal band must satisfy Eq. (3). Figure 9 shows the image components generated in case of the signal in the vicinity of  $F_{BW}$ . When the signal is frequency-shifted, the split band is moved from the original frequency by  $f_{ck}$ . An image component is generated close together, so it cannot be completely removed by the following filter; therefore, the overall FI-DAC output quality decreases. The overall DAC output band should be a little narrower than  $F_{BW}$ , which is defined as  $F'_{BW}$ .

## 4 Fundamental Problems of FI-DAC Architecture in Principle and Their Compensation

### 4.1 Signal Attenuation by DAC Zero-th Order Hold Output and Its Compensation Method

As mentioned in Sect. 2.2, the DAC output has zero-th order hold characteristic. So, the gain of the sub-DAC should be compensated to be flat over the signal band by using a pre-digital filter having the inverse gain characteristic of the sinc filter as shown in Fig. 10a. In this way, even if the filter coefficients are normalized to the



**Fig. 10** Compensation of signal attenuation caused by DAC zero-th order hold output

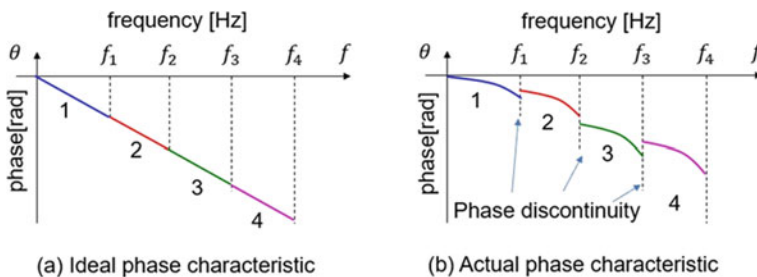
largest value, the dynamic range of the DAC output is reduced. Then, the amplitude of the output is reduced from the signal bandwidth input to sub-DAC. This loss can be compensated by the FI-DAC output amplifier.

## 4.2 Problems of Phase Characteristic and Their Compensation

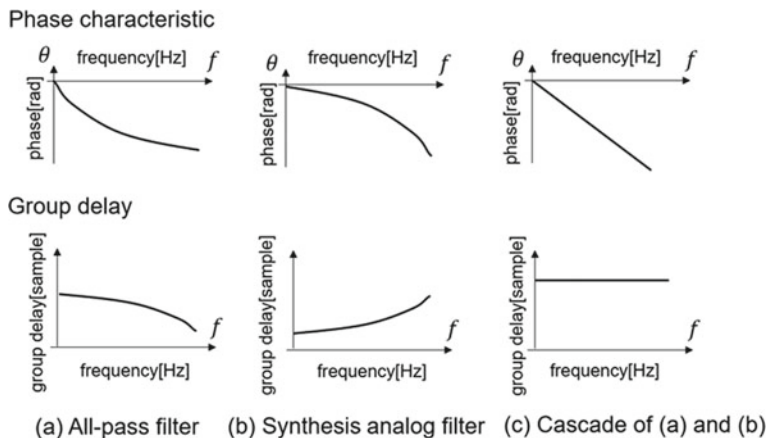
Figure 11a shows the ideal overall DAC output phase characteristic which is completely linear. However, the actual phase characteristic for the 4-channel FI-DAC tends to be degraded, as shown in Fig. 11b, which leads to the output waveform distortion. The causes of such phase characteristic degradation are as follows:

1. Phase-nonlinearity characteristic of synthesis analog filters
2. Differences in group delay among subband channels
3. Phase discontinuity between adjacent subband channels.

Now we describe their compensation methods.



**Fig. 11** Overall DAC output phase characteristic



**Fig. 12** Compensation for phase-nonlinearity characteristic of smoothing and synthesis analog filters

### Phase-nonlinearity Characteristic of Smoothing and Synthesis Analog Filters and their Compensation

The phase-nonlinearity of the smoothing and synthesis analog filter can be compensated using the all-pass filter. In Fig. 12, (a) shows phase characteristic of the all-pass filter for the compensation and (b) shows phase characteristic of an analog filter, while (c) shows phase characteristic of cascade of these filters. Due to the phase-nonlinearity characteristic of the smoothing and synthesis analog filters, the group delay varies depending on the frequency. Therefore, an all-pass filter is used to perform group delay compensation [18].

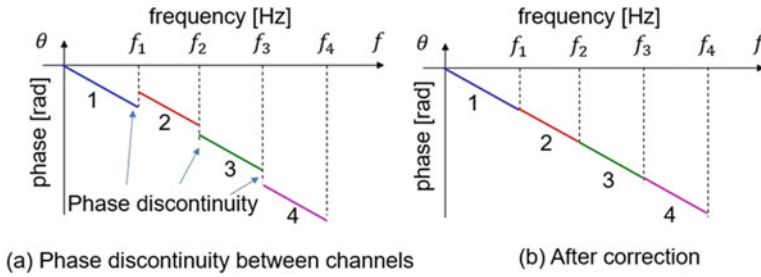
### Differences in Group Delay among Subband Channels and their Compensation

As mentioned above, they can be compensated so that the group delay of the analog filter is constant in the entire band. However, the group delay of the filter by the compensation is different for each subband channel. In addition, since the first subband channel has a filter of a small order than the other subband channels, the group delay by the filter is small. Therefore, the group delays are different among subband channels, and some distortion appears in the overall DAC output waveform. To compensate the group delay, we investigate the sampling timing adjustment for each subband channel to compensate for the group delay differences among subband channels.

### Phase Discontinuity Between Adjacent Subband Channels and their Compensation

In Fig. 13, (a) shows the phase characteristic compensated only by phase-nonlinearity characteristic of synthesis analog filters and differences in group delay among subband channels, and (b) shows the one with also phase discontinuity compensation. In this system, signal band changes by modulation and frequency shift. Filtering





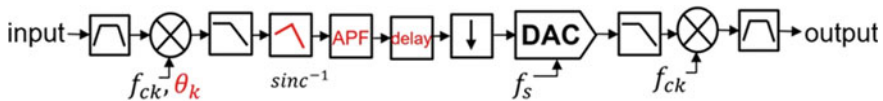
**Fig. 13** Compensation of phase discontinuity between adjacent subband channels

when changing from the original band causes an extra rotation of phase. In Fig. 13a, the phase for each subband channel is linear, but phase differences exist between adjacent subband channels, and the phase characteristic is discontinuous. This can be compensated by adjusting the initial phase of the carrier signal in digital signal processing. Let  $x(n)$  be a signal,  $\cos(2\pi f_c n + \theta)$  be a carrier. A signal by frequency shift is calculated from multiplying  $x(n)$  by  $\cos(2\pi f_c n + \theta)$ . Then discrete fourier transform (DFT) of  $x(n) \cos(2\pi f_c n + \theta)$  is given as follows:

$$\sum_{n=0}^{N-1} x(n) \cos(2\pi f_c n + \theta) = \frac{1}{2} (\exp(j\theta) X(k + p) + \exp(-j\theta) X(k - p)) \quad (4)$$

where  $X(k)$  is DFT of  $x(n)$ ,  $f_c = p/N$  is the carrier frequency in digital signal processing,  $N$  is the number of the sampled data points, and  $p$  is a natural number satisfying  $p = f_c N$ .  $x(n)$  has both upper and lower sideband frequency components because of frequency shift. Based on the initial phase, the high-frequency component rotates by  $\theta$  and the low-frequency component rotates by  $-\theta$ . As shown in the third power spectrum from the left in Fig. 7b, since the high-frequency component is removed by a low-pass filter, the signal components are only at low frequency. That is, by giving the initial phase  $\theta$ , the signal phase can be rotated by  $-\theta$ . As a result, the phases between adjacent subband channels become continuous as shown in Fig. 13b.

Figure 14 shows the block diagram with the compensation described in this section.



**Fig. 14** Block diagram with compensations

## 5 MATLAB Simulation

### 5.1 Simulation Conditions

We have conducted MATLAB simulation to show the validity of the compensation methods for these problems. Figure 15 shows a block diagram of simulation, and the compensations are shown in red there.  $F_{BW}$  is signal band of the input, and  $f_s$  is the sampling frequency of the sub-DACs. In this simulation, the signal bandwidth input to sub-DAC is at most  $f_s/2$ . We give  $F_{BW} = 3f_s/2$ , and the final output band  $F'_{BW}$  is  $0.9 F_{BW}$ .

We use two types of inputs; one is an impulse signal to obtain the amplitude and phase characteristics for frequency. The other is also a 4-tone signal for time-domain waveform comparison. Notice that here the DAC has an infinite word length without quantization error.

The passband ripple of the subband digital filter is 0.2 dB, and its stopband ripple is  $-80$  dB. The analysis digital filter was designed with an FIR filter based on a Kaiser window because the FIR filter is linear phase characteristic and the Kaiser window can change the stopband ripple by changing its parameter. The synthesis analog filter was designed with an elliptic filter, because its stopband can be significantly attenuated in a smaller order than other analog filters. Table 2 shows the orders of the filters used in the simulation. Note that we use two types for all pass filters as compensation of phase-nonlinearity characteristic of smoothing and synthesis analog filters. Since the order is smaller than compensating with one type of filter, the group delay is also smaller. Since the circuit size is not considered in this paper, their orders are high. High orders of digital filters may not be a problem, but high order synthesis analog filters would consume much power. If their orders are lower, their transition bands are wider. Its one remedy is to use higher sampling frequency of sub-DACs;

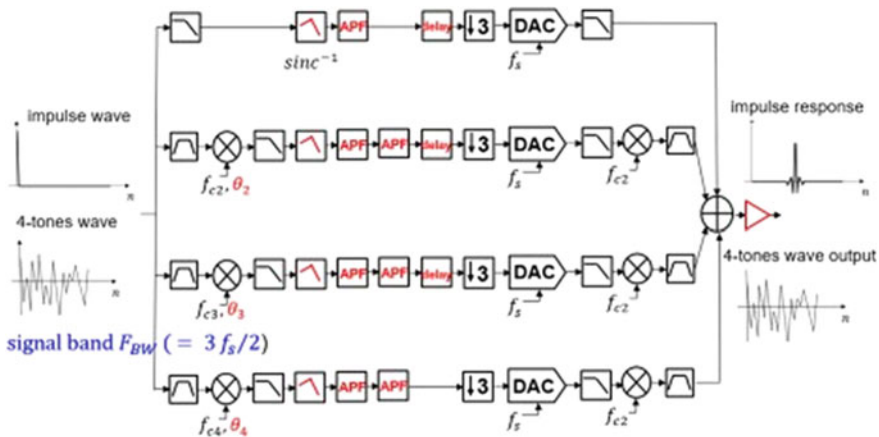


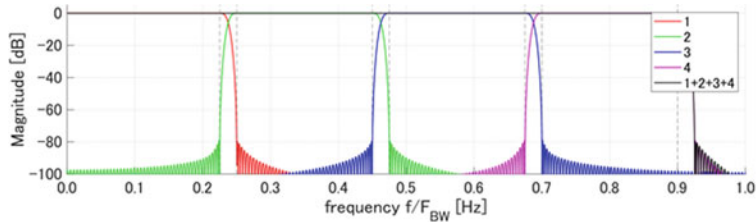
Fig. 15 Simulation block diagram

the input signal bandwidth to sub-DAC becomes can be extended to be wider and then their transition ranges are allowed to be wider, which leads to the lower orders of their corresponding analog filters.

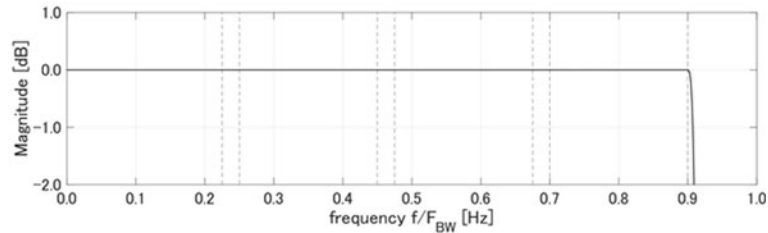
Figure 16 shows the frequency characteristic of subband analysis digital filters. In Fig. 16, (a) shows the filters up to subband channels 1 ~ 4 and their combination, and (b) shows scaling up the pass band of their combination. Note that the dotted line in Fig. 16b shows the range where the signal components between subband channels overlap. The signal component is not attenuated inside the dotted line. Following this, the band splitting is performed using these filters.

**Table 2** Orders of the filters used in simulation

Subband Channel	Digital FIR filter			Digital IIR filter		Analog elliptic filter	
	Subband	Lowpass	Inverse sinc	All-pass (smoothing)	All-pass (Analog bandpass)	Smoothing	Analog bandpass
Ch1	402	–	200	14		13	–
Ch2	402	100	200	14	22	13	20
Ch3	402	100	200	14	20	13	20
Ch4	402	100	200	14	20	13	18



(a) Subband analysis digital filter and their combination



(b) Scaling up the pass band of their combination

**Fig. 16** Frequency characteristic of subband analysis digital filters using simulation

## 5.2 Simulation Verification by Impulse Response

### Signal Attenuation by DAC Zero-th Order Hold Output

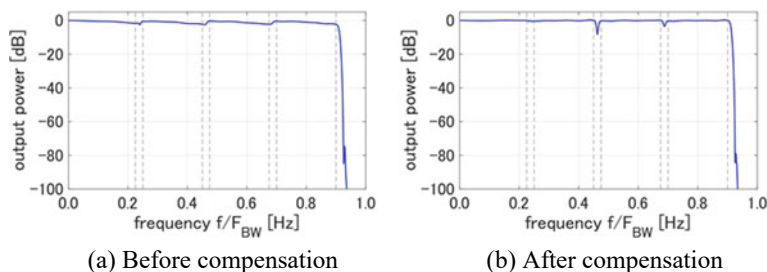
Figure 17 shows the simulation results of before and after compensation for the signal attenuation by zero-th order hold. In Fig. 17, (a) shows the power spectrum without any compensation in Fig. 15, and (b) shows the one with compensation by an inverse sinc filter. In the case of Fig. 17a, a zero-th order hold characteristic appears in the signal component of each subband channel. In the case of Fig. 17b, the compensation makes the gain to be flat in the entire band of each subband channel. However, since the phase compensation is not performed, the output signal is attenuated inside of the dotted line.

### Phase-nonlinearity Characteristic of Synthesis Analog Filters

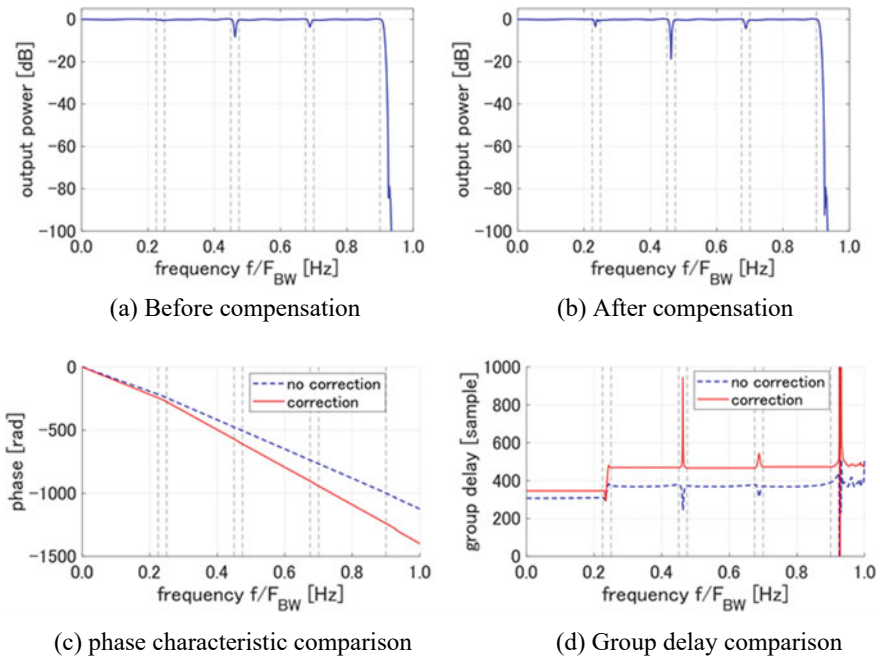
Figure 18 shows the simulation results before and after compensation of phase-nonlinearity characteristic of synthesis analog filters. In Fig. 18, (a) shows the power spectrum with only compensation for signal attenuation by zero-th order hold, and (b) is the one with compensation by all-pass filters, while (c) (d) show the comparison of phase characteristic and group delays in (a) and (b). This simulation is to verify the compensation of phase characteristic, so the power spectrum does not change except inside of the dotted line. As shown in Fig. 18d, before the compensation, the group delay for each subband channel is not straight but curved. On the other hand, after compensation, the group delay is straight. However, the group delays are different among subband channels.

### Differences in Group Delays among Subband Channels

Figure 19 shows the simulation results before and after comparison of group delay difference among channels before and after compensation. In Fig. 19, (a) shows the power spectrum with the above two compensations, and (b) is the one in case with compensation by a delay, while (c) is a comparison of the group delays of (a) and (b). Similarly, this is the compensation of phase characteristic, so the power spectrum does not change except inside of the dotted line. As shown in Fig. 19c,



**Fig.17** Simulation results of before and after compensation for the signal attenuation by DAC zero-th order hold output

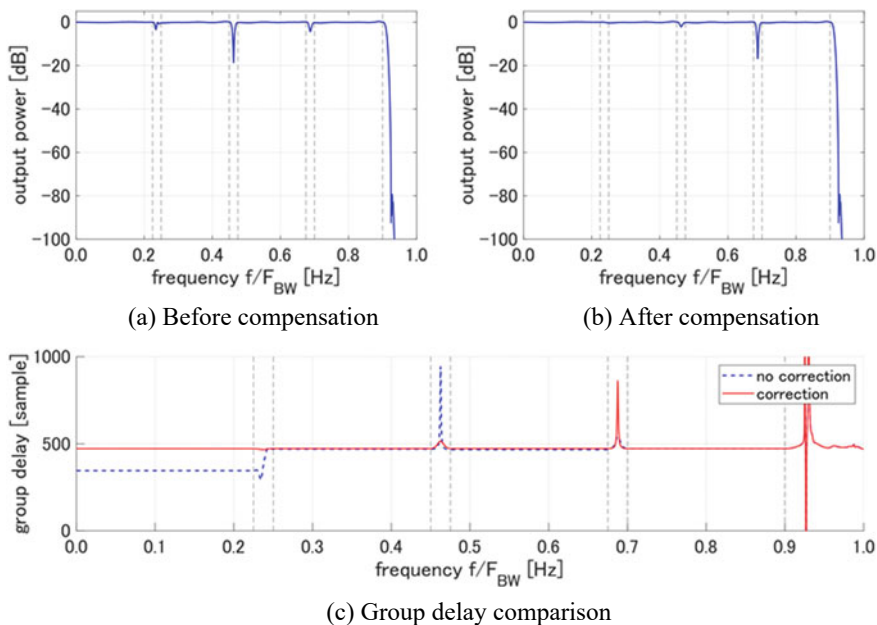


**Fig. 18** Simulation results before and after compensation of phase-nonlinearity characteristic of synthesis analog filters

after compensation, the difference in group delay between among subband channels is reduced substantially. However, since the phases between subband channels do not match, the signal is attenuated inside of the dotted line.

### Phase Discontinuity Between Adjacent Subband Channels

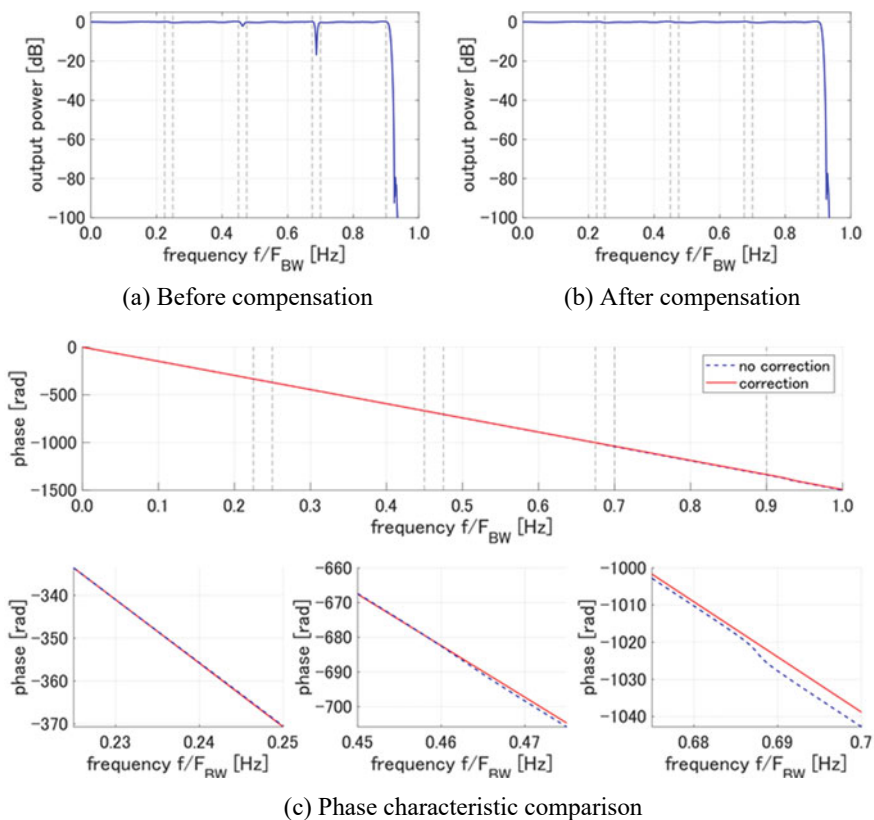
Figure 20 shows the simulation results before and after compensation for the output phase discontinuity between adjacent subband channels. In Fig. 20, (a) shows the power spectrum with the above three compensations, (b) shows the one with the compensation by the initial phase to the carrier signal, (c) shows the comparison of the phase characteristic in (a) and (b). In Fig. 20c, the upper figure is the overall phase characteristic, and the lower is the enlarged display inside the dotted line. Before compensation, the power spectrum is attenuated significantly inside the dotted line. As shown in Fig. 20c, a large deviation appears in the phase characteristic in this range. After compensation, this deviation disappears, and the power spectrum inside of the dotted line is also not attenuated.



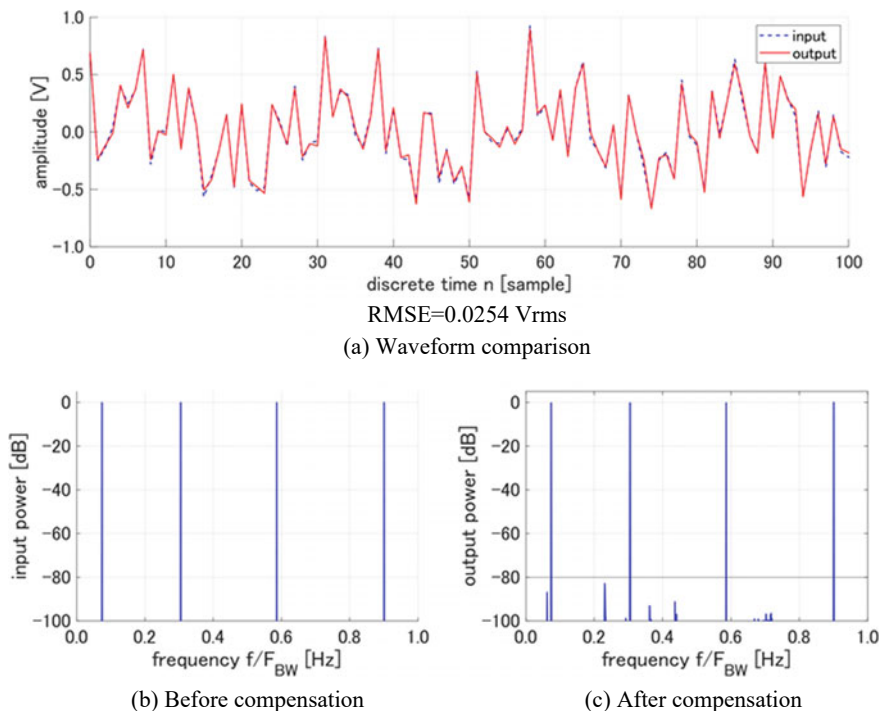
**Fig.19** Simulation results before and after comparison of group delay difference among subband channels

### 5.3 Simulation Verification by 4-tone Input Signal

Figure 21 shows the simulation result using the 4-tone input signal to compare the time-domain waveforms. In Fig. 21, (a) shows the input/output waveforms, while (b) and (c) show their power spectrums. In Fig. 21a, the RMSE of the output waveform is 0.0254 Vrms, and almost no distortion appears in the output waveform. Then, there are no spurious in the output power spectrum greater than the filter stop band ripple of  $-80$  dB.



**Fig. 20** Simulation results before and after compensation phase discontinuity between adjacent subband channels



**Fig. 21** Simulation results using the 4-tone input signal

## 6 Conclusion

We have investigated the basic configuration of the FI-DAC architecture. There, the image components by sub-DACs and modulations are generated, and they are removed with the following synthesis analog filters. Besides, these filters have transition bandwidth, and therefore, the input band of each sub-DAC becomes narrow. In the FI-DAC, in order to make the input band of each sub-DAC be wider, the transition band of the synthesis analog filters should be narrower, and their orders should be higher.

We have investigated the subband processing using digital filters for the band division. Then, by overlapping the transition band of the filters, the overall DAC output is kept not to be attenuated. To have the input band of the sub-DAC, the frequency shifting and down sampling for its input are performed.

There are several fundamental problems in the above series of processing. The compensation methods for these problems have been discussed. Signal attenuation by zero-th order hold of the sub-DAC output can be compensated by applying an inverse sinc filter. Phase-nonlinearity characteristic of synthesis analog filters is compensated by applying all-pass filters. Differences in group delay among subband channels are



compensated by inserting a delay and changing the sampling timing. Phase discontinuity between adjacent subband channels can be compensated by adjusting the initial phase of the carrier signal in digital signal processing.

We have performed MATLAB simulation, which showed changes in the amplitude and phase characteristic of the signal by compensations. The input and output waveforms were compared, and the validity of the compensation method was verified.

In this paper, we have shown especially the configuration that gives an initial phase of the carrier in digital domain to compensate for the following problem: even the phase for each subband channel is linear, the final output has phase discontinuity and its waveform is distorted. Our configuration has solved this problem.

Frequency interleaving can realize a wideband DAC, and hence its application as a wideband signal generator for testing communication equipment is expected.

Remaining is the development of compensation methods for circuit implementation issues. For example, we have investigated the compensation for the difference in group delay among subband channels by changing the sampling timing. However, it can only be compensated by delay values that are integer multiples of the sampling period. The group delay of the synthesis analog filter does not necessarily exist within this compensation range, and compensation up to the minute delay cannot be performed. Therefore, it is necessary to compensate for this delay by using analysis digital filters, whose group delay can be adjusted with the time resolution of fractional of the sampling period as described in [19, 20].

We will examine circuit implementation issues such as variations in elements during manufacturing, minute delays, nonlinear distortion, and phases of carriers, spurious at inputs, and quantization errors in digital signal processing parts. We will construct their compensation algorithms.

## References

1. Kurosawa N, Kobayashi H, Maruyama K, Sugawara H, Kobayashi K (2001) Explicit analysis of channel mismatch effects in time-interleaved ADC systems. *IEEE Trans Circ Syst I: Fundam Theory Appl* 48(3):261–271
2. Kabeya R, Umeda Y, Takano K (2021) Frequency-interleaved ADC with RF equivalent ideal filter for broadband optical communication receivers. In: *IEEE international conference on electronics, circuits and systems*, Dubai, United Arab Emirates
3. Song J, Tian S, Hen Y (2019) Analysis and correction of combined channel mismatch effects in frequency-interleaved ADCs. *IEEE Trans Circ Syst I: Regul Pap* 66(2):655–668
4. Asami K, Kusunoki K, Shimizu N, Aoki Y (2020) Ultra-wideband modulation signal measurement using local sweep digitizing method. In: *IEEE 38th VLSI test symposium*, San Diego, CA
5. Olieman E, Annema AJ, Nauta B, Bal A, Singh PN (2013) A 12b 1.7GS/s Two-times interleaved DAC with  $<-62$ dBc IM3 across Nyquist using a single 1.2V Supply. In: *IEEE Asian solid-state circuits conference*, Singapore
6. Balasubramanian S, Creech G, Wilson J, Yoder SM, McCue JJ, Verhelst M, Khalil W (2011) Systematic analysis of interleaved digital-to-analog converters. *IEEE Trans Circ Syst II: Express Briefs* 58(12):882–886

7. Pupalaikis PJ, Yamrone B, Delbue R, Khanna AS, Doshi K, Bhat B, Sureka A (2014) Technologies for very high bandwidth real-time oscilloscopes. In: IEEE bipolar/BiCMOS circuits and technology meeting, Coronado, CA
8. Schmidt C (2020) Interleaving concepts for digital-to-analog converters: algorithms, models, simulations and experiments. Springer, pp 99–177
9. Okawara H (2005) Analysis of pseudo-interleaving AWG. In: International test conference, Austin, TX
10. Deveugele J, Palmers P, Steyaert MSJ (2004) Parallel-path digital-to-analog converters for Nyquist signal generation. *IEEE J Solid State Circ* 39(7):p1073-1082
11. Jha A, Kinget PR (2008) Wideband signal synthesis using interleaved partial-order hold current-mode digital-to analog converters. *IEEE Trans Circ Syst II: Express Briefs* 55(11):1109–1113
12. Schmidt C, Kottke C, Tanzil VH, Freund R, Jungnickel V, Gerfers F (2018) Digital-to-analog converters using frequency interleaving: mathematical framework and experimental verification. *Circ Syst Signal Process* 37:4929–4954
13. Schmidt C, Kottke C, Jungnickel V, Freund R (2016) Enhancing the bandwidth of DACs by analog bandwidth interleaving. ITG-Fachbericht-Breitbandversorgung in Deutschland, Berlin
14. Galetto AC, Reyes BT, Morero DA, Hueda MR (2021) Background compensation of frequency interleaved DAC for optical transceivers. In: IEEE 12th Latin America symposium on circuits and systems, Arequipa, Peru
15. Galetto AC, Reyes BT, Morero DA, Hueda MR (2021) Adaptive background compensation of frequency interleaved DACs with application to coherent optical transceivers. *IEEE Access* 9:41821–41832
16. Crochiere RE, Rabiner LR (1996) Multirate digital signal processing. In: Prentice-hall signal processing series. Prentice Hall
17. Vaidyanathan PP (1990) Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial. *Proc IEEE* 78(1):56–93
18. Antoniou A (1993) Digital filters: analysis, design, and applications, 2nd edn. McGraw-Hill, Inc.
19. Asami K, Miyajima H, Kurosawa T, Tateiwa T, Kobayashi H (2010) Timing skew compensation technique using digital filter with novel linear phase condition. In: IEEE International test conference, Austin, TX
20. Asami K, Tateiwa T, Kurosawa T, Miyajima H, Kobayashi H (2011) Digitally-assisted compensation technique for timing skew in ATE systems. In: IEEE international mixed-signals, sensors, and systems test workshop, Santa Barbara, CA

# Neural Network Models for Time Series Analysis and Estimation



Louay Al Nuaimy

**Abstract** This paper focuses on time series and regression modeling using neural networks. Based on the recent results of least squares estimation of nonlinear time series, the research proposed a comprehensive and straightforward methodology for parameter estimation (learning process) and model selection (structure selection). This research presents a solution to the problem of pruning in a multi-layer perceptron by means of a step-by-step method using the Schwarz information criterion (SIC) type standard in which can be demonstrate its consistency. In more concise words, this paper deals with neural network modeling for time series and regression analysis based on recent results of least square estimation for nonlinear time series, the work proposed a complete and feasible methodology for parameter estimation (learning process) and model selection (structure selection). This work has solved the pruning problem of multi-layered cognition models using a gradient search method using the SIC criterion which has been shown to be consistent.

**Keywords** Statistical selection · Asymptotic stats · Graded stats · Near-certain determination

## 1 Introduction

Multi-layer perceptron's (MLP) were first introduced to solve complex classification problems. But due to their universal approximator property, they were quickly used as nonlinear regression models, and then for time series modeling and forecasting [1]. However, the estimation and identification of these models use sophisticated techniques, and it is not easy to determine the adequate architecture. Indeed, these models are over-parameterized, the error functions to be minimized have many local minima, and the implementation often turns out to be tricky. Many articles discuss

---

L. Al Nuaimy (✉)

Oman College of Management and Technology, Halban, Sultanate of Oman

e-mail: [loay.alneimy@omancollege.edu.om](mailto:loay.alneimy@omancollege.edu.om)

techniques for pruning unnecessary parameters, especially in the context of regression models, and users have extended the proposed techniques to the case of time series [2]. Most of these papers provide heuristics, but do not place themselves in a rigorous statistical framework [3].

This paper proposes a set of theoretical results established in the context of time series neural models, which extend known results in the context of linear statistical models. In fact, these results are also valid in the context of regression models and mixed models (auto-regressive models also including exogenous variables), but to simplify the presentation, we place ourselves only in the auto-regressive framework [4].

One can therefore consider a family of models called neural auto regressive (NAR) models, defined by [5]:

$$Y_t = f_w(Y_{t-1}, \dots, Y_{t-p}) + \epsilon_t \quad (1)$$

where:

$Y_t$ : Multidimensional case,

$f_w$ : Represents a function implemented by a multi-layer perceptron with a single output unit,  $Y_{t-1}$ :  $i = 1 \dots, p$  are the delays of the series ( $Y_t$ ).

$\epsilon_t$ : the noise at time  $t$ . It can be considered that the following a  $(p, K)$  multi-layer perceptron, with a linear output unit,  $p$  linear input unit,  $K$  hidden units equipped with a sigmoid activation function  $\emptyset$  of hyperbolic tangent type (odd function) Then a model (NAR) is precisely defined by an equation of the type [6]:

$$Y_t = f_w(Y_{t-1}, \dots, Y_{t-p}) + \epsilon_t = d_0 + \text{Sum}(D_j \emptyset(\text{Sum } B_{ij} Y_t - i + B_{0j})) + \epsilon_t \quad (2)$$

where the assumptions are those of Eq. (1). The notations are the usual notations:  $B_i$ :  $1 \leq i \leq p$ ,  $1 \leq j \leq K$ . The parameter corresponding to the weight of the connection between the input unit  $i$  and the hidden unit  $j$ ,  $D_j$ :  $1 \leq j \leq K$  corresponds to the weight of the connection between the concealed unit  $K$  and the output unit,  $B_{0j}$ :  $1 \leq j \leq K$  is the constant associated with the hidden unit  $K$  and  $j$  is the constant corresponding to the output unit [7]. Equation (2) then defines a parametric model having a particular functional form. The study presented asymptotic properties of the parameter estimators for a specific model and a given number  $R$  of observations. These properties are relevant in learning, a statistical methodology that helps in selecting the best model among several options (from a model dominant), which in this case corresponds to the choice of network architecture. The analysis focused on real outputs only, but all the properties presented can be readily extended to the multidimensional case [8].

Let  $W = \{(D_j) \ 0 \leq j \leq k, (B_{ij}) \ 0 \leq i \leq p, 1 \leq j \leq k\}$  the parameter vector. Note that its dimension is  $m = (p + 2)k + 1$  Given  $T + p$  observations,  $(Y_{-p+1}, \dots, Y_0, Y_1, \dots, Y_T)$ .

Within the series, we estimate  $W$  by minimizing the mean of the residual squares (Error function) [9].

$$S_t(W) = \text{Sum}(\text{square}(Y_t - f_w(Y_t - 1, \dots, Y_t - p))), 1 \leq t \leq T \quad (3)$$

Notice that  $W^{\wedge t} = \arg \min_w S_t(W)$  the least squares estimator of  $W$ .

If  $\text{var}(\epsilon_t)$  is assumed  $\sigma^2 > \mu > 0$ , this amounts to a bare minimum [10]:

$$\text{LV}_t(W) = \ln(\text{sum}(Y_t - f_w(Y_t - 1, \dots, Y_t - p))^2) = \ln(S_t(W)) \quad (4)$$

In the case where is Gaussian,  $\epsilon_t$  is exactly what statisticians call the concentrated log-likelihood. We denote:  $W^{\wedge t} = \arg \min_w \ln \text{LV}_t(w)$ .

The two minimizations are equivalent ( $W^{\wedge} = W$ ), but it can have distinguished them to recall how they were obtained their asymptotic properties are almost the same, but it will be seen that in practice it is better to use the second [11]. The calculation of the estimators can be carried out using any minimization method. This work does not deal with the above problem, and it has been assumed that  $W_t$  is the true minimum of the Error function  $S_t(W)$  [12]. In the general case, not necessarily Gaussian, the least squares estimator is a special case of minimum contrast estimator.

The paper is organized as follows: Sect. 2 states some mathematical results on least squares estimators for an NAR model. Section 3 provides theoretical results of almost sure identification. Section 4 presents a new method of almost sure identification of the true model based on a stepwise descent method (Stepwise Statistical Method, SSM) which allows to prune unnecessary parameters. Finally, paragraph 5 shows on a simulated example how this method is used. The last paragraph contains conclusions and perspectives.

## 2 The Identifiability of MLP

To estimate the parameter vector, a fundamental property for obtaining consistency results is the identifiability of the model. This means, that for a function representable by a given MLP, there is only one parameter vector that represents that function [13]. Suppose that a multilayer perceptron of dimension  $(p, k)$ , that is to say with  $R^m$  inputs, four hidden units and an output, is a parametric function on  $L$ , with  $m = (p + 2) * k + 1$ , the model is not identifiable [14]. We can indeed find two different parameter systems that generate the same outputs. This can be achieved, for example, if two hidden units have strict upstream weights identical since it will be enough that the sum of their downstream weights is constant to represent the same function. However, if the model limit to a reasonable set of parameters, MLPs become identifiable, except for trivial operations. This will then lead to a weak identifiability [15].

It is given in the following conditions necessary and sufficient for the model to be identifiable, in the weak sense, in the case of an MLP to a hidden layer, having hyperbolic tangents for activation functions  $\emptyset$ . A first precaution to take is not to artificially use too many hidden units [16]. For example, if it considered that an MLP with a hidden unit whose weight connected to the output is zero, this is totally useless since the input weights to this unit are any, that they do not influence the function

represented by this MLP. We will therefore characterize MLPs that do not have unnecessary hidden units in a manifest way by the notion of irreducible MLPs [17]. From now on, only irreducible MLPs are considered, i.e., not comprising hidden units that are manifestly useless or exactly duplicated. Note that this restriction does not prevent MLPs from being over-configured. There are still trivial transformations that do not change the function represented by the MLP [18]. For example, is on choice lunate cache  $K$ , if we change the sign of all weights  $B_{ij}$  for  $1 \leq I \leq p$  and change the sign of the output weight  $d_j$  associated with this hidden unit, as the hyperbolic tangent function is odd, this will not change the function  $F(W)$  [19].

### 3 The Ability to Generalize

One of the main difficulties involved in the use of increasingly complex functions for the statistical estimation of processes is the phenomenon of overfitting. If a too complex on a little data use as a model that is, the noise ends up from the modeling process that generated the data on which that estimate the model [19]. A bias is thus introduced into the model which strongly compromises the validity of its results on new data from the same process. At that time, it can be said that the model “generalizes” badly. It therefore appeared fundamental to control the complexity of the model to ensure that its error remains low, not only on the data that that observe, but also on future data, not yet observed, coming from the same phenomenon [20]. For example, to use the principle of minimization of the structural risk where this principle defines a compromise between the quality of approximation and the complexity of the functions of approximation. However, the main drawback of the bounds that they are valid only for identically distributed independent random variables. However, in this article dealing with time series and are not in an its indicators [21]. Moreover, the dimension of the models is known only in the case of Multilayer Perceptron’s (MLPs) that have indicator functions on the hidden layer. For MLPs with hyperbolic tangent activation functions, only upper bounds for this dimension exist [20]. Therefore, to address the issue of over-parameterization in our models, we will instead employ a penalty term that considers the number of parameters and the amount of data. This approach, based on the principle of parsimony, shares similarities with the SIC principle, but it differs in terms of theoretical framework and resulting outcomes [21].

It can be assumed here that the choice between several models  $M$  to explain the observed process. How to choose a model properly? The choice must be parsimonious (the fewest possible parameters) but providing a good adjustment (parameters in sufficient number). Using the existing theorems on model selection by means of penalized contrasts, it can be shows the almost sure identification, thanks to a penalization term, when there is a finite number of possible models, all having a common dominant model [22]. More precisely, suppose that there is an upper bound on all possible dimensions of the model. This assumption, although a criterion for penal methods, may seem sound in theory. However, this has no practical consequences

because the process is still limited to the maximum architecture, if only due to the limited computing and memory capacities of computers [23].

Let  $W$  subset of  $R^m$  be a dominant model, whose parameter vector is  $W_{\max} = (w_1, w_2, \dots, w_m)$ . Consider the finite family  $F = \{(w_1, w_2, \dots, W_m)$  for some components can be zero  $Z$ . For  $F$  subset of  $F$ , a sub-model of  $F_{\max}$ , denote  $m(F)$  the number of non-zero parameters, i.e., the dimension of the parameter vector  $W$ , and  $W_f$  the set of possible values of  $W$ , which assumed that the true model is a sub-model of  $W$ , that denote  $F_{\max}$  and the true value of the parameter vector is denoted  $W_0$  of dimension  $m(F_0)$ . Let  $W^{\wedge}t,f$  be the least squares estimator of  $W$  restricted to  $F$ ,  $W^{\wedge}t,f = \text{Arg min } S_t(W)$ ,  $W$  belong to  $W_r$ . And  $S_t(F)$  (instead of  $S_t(w^{\wedge}t,f)$ ) the lower bound corresponding to the error function. where it would also be  $c(t)$  to be a sequence of positive real numbers. Punishable least squares variance, covariance, with penalty rate  $c(t)$  takes shape [23]:

$$\text{CWP}(t, F) = (S_t(F) + (t) * m(F))/T \quad (5)$$

Let  $F^{\wedge} = \text{Arg min } (\text{CWP}(t, F))$  be the estimated model, which results from two successive reductions of a constant  $T$  a reduction over a continuous space, to account for  $W^{\wedge}t,f = S_t(F)$ , and a reduction over a finite space, to account for  $F^{\wedge}t$ . From these definitions, one can derive the following result, the complete proof of which can be found [24]:

$$\text{Lim } t(c(T)/t = 0 \text{ and } \text{lim } t(c(T)/2 \ln \ln T) > \delta_2 A/d \quad (6)$$

In fact, the same results can be demonstrated using the properties of the estimator  $W^{\wedge}t$  and in this case the criterion with penalty used is exactly the usual SIC criterion given by [25]:

$$\text{SIC} = \ln(S_t(F) + \ln t * m(F))/T \quad (7)$$

This is associated with the maximum likelihood estimator when the noise is Gaussian, and it has the advantage that it is not necessary to introduce a constant  $\hat{Y}$ , because the first term in logarithm is quite insensitive to noise variance [4]. On the other hand, this criterion can only be used if  $E \epsilon_t < \infty$ . Two criteria's thus can be obtained (close, but different) whose minimization leads in theory almost surely to the true model if the process start from a dominant model that is a supermodel of the true model [25].

## 4 Practical Search for the Real Model

So, the following method can be suggested to determine the true model by initializing the structure, the process starts by taking all relevant inputs (as obtained from a basic linear model), and one input is a hidden unit [26]. Then gradually add the units in the hidden layer, calculating the criterion SIC at each step, if the value of this criterion

decreases. When the SIC norm remains stable or begins to increase, we stop looking for the model and we take it as the observed dominant  $F_{\max}$  [27]. Note that with this methodology, the SIC criterion is assumed to be a convex function of the number of hidden units [28]. This is what is classically done when SIC type criteria are used in linear models. It seems difficult to justify this hypothesis in theory; however, this method has the advantage of being easy to implement and of giving good results empirically as shown by the example treated in the following section and studies on real data [29].

The parameters are initialized in a very simple way. With a single hidden unit, the coefficients  $B_{i1}$  are taken equal to the values of the linear form, the parameter  $d_0$  is taken equal to the average value of the chain, and the others small and random (for example, between  $-0.5$  and  $0.5$ ) [30]. In fact, many researchers have proposed ways and tricks to properly initialize the parameters, but no effective method has yet been proposed that would allow us to approach the global minimum. Tests with a simulated annealing algorithm, but the computation times are too long, and one can simply repeat several different initializations and keep the best one. Recall that  $W_{\max} = (w_1, w_2, \dots, w_m)$  is the parameter vector associated with the form  $F_{\max}$ . In principle, to estimate the real model, it is necessary to know exhaustively the finite family of all the sub-models of the dominant form  $F$  and to calculate the SIC code for each one. But the number of such subforms is too large, and it is practically impossible to do so [3]. Thus, as in linear regression, a step-by-step statistical method can be proposed: the step-by-step statistical method, SSM) directing research in  $F^\wedge$ . This lineage strategy relies on a naturalistic approach to intended  $W_t$ . Deciding whether or not to remove the weight of  $w_1$  is equivalent to creating a test for the null hypothesis  $w_0$  versus the alternative hypothesis  $w_1 \neq 0$ , this is the Student's test on  $w_1$ , (actually a test Gaussian since the normality of the estimator only the weight is guaranteed when  $T$  is large) [4].

The preceding notes apply specifically in the case of a minimization of the SIC norm, but they also apply approximately to the SIC norm. It is not equivalent to reducing the difference or the difference to logarithms, but it is not very different. Briefly, the procedure for finding the true form is as follows:

1. Define  $F_{\max}$  as described in Sect. 4 and estimate the parameters (Network Training).
2. Calculate for each  $t$  the ratio  $Q_t = w^\wedge t / \sigma^\wedge(w^\wedge t)$ .
3. Find the minimum of these ratios in absolute value.
4. Agree to eliminate  $w_t$  only if the SIC criterion falls.
5. If rejected, stop the disqualification process, and keep the previous form. If accepted, re-estimate the parameters corresponding to the model  $F1_{\max}$ , and repeat step 2 of the model  $F_t_{\max}$  to search for evidence of etymology, etc.

The stopping rule is completely normal because the deletion continues if the SIC criterion is reduced. This provides an objective criterion based on the statistical properties of weight estimators. It is already possible to determine the “small” weight that can be thrown. The SSM algorithm belongs to the family of inverse gradient algorithms widely used in regression analysis. Note that it is also possible to use the



stepwise method to guide the search for a lower world standard for the SIC standard [31]. The SSM segmentation method is close to the previously defined OBD (On-Board Diagnostics and Boundary) algorithm because it chooses the filter parameter to be segmented in the same way. But their algorithm does not provide a criterion to stop the pruning process; it needs a performance evaluation which is performed externally on an external data set. This requires a strategy of dividing the data into a training set and a validation set, sacrificing a certain amount of data that is detrimental to the quality of model evaluation. Here, thanks to the results on the almost certain determination of the model, a theoretical stopping criterion can be obtained: the SIC criterion. The principle is to no longer delete the parameters of the SIC standard. It should be noted that with this method, the maximum amount of data is kept for learning, which makes it possible to make the most of the information they provide [2].

## 5 Practical Example

In this part of the research, the effectiveness of the SIC standard on simulation is tested. The real architecture, the real parameter vector  $S$  = and the variance of the associated Gaussian noise are known. The chosen model (see Fig. 1) is released with 2 inputs, 2 hidden units and 1 output, and 8 parameters (one of the hidden units is not bound to one of the inputs. The function fw0 is written for the real model:

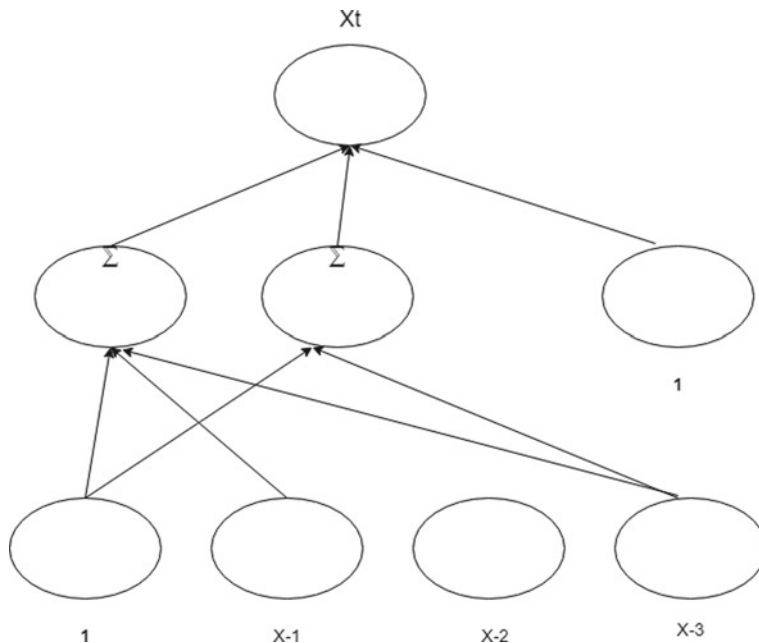
$$X_t = \tanh(-0.5X_t - 1 - 1.5X_t - 3 + 0.5) + \tanh(X_t - 3 - 0.5) + \epsilon_t$$

The noise used is  $\epsilon_t$ . Gaussian with contrast  $\sigma^2 = 0.1$ .

The goal is to find the real structure as well as the right parameters. As this model has few parameters, one can compare the exhaustive search with the SSM methodology. The search will be done among the sub-models of the dominant model described in Fig. 2.

The sequence of 9800 points was simulated 60 times independently. Figure 1 the structure of the real model which contains eight links. The dominant model in use, with 16 connections, is shown in Fig. 2. For the SIC standard, we fix  $\hat{Y} = \sigma^2 = 0.1$ , which is an option that has the advantage of making the penalty duration weight constant to a noise level of the process being studied which on average gives the best practical results. To avoid local minima, parameter estimation is the best obtained among 10 estimations starting from various random initializations.

Table 1 shows the proportion of the three constructs  $A$ ,  $B$ , and  $C$  (underestimation of the SIC) held by the in-depth research on 60 simulations. As expected, the true structure appears in 73% of cases. The other winners  $i$  and  $i - 1$  are very close to the final architecture, but with one more connection and one less connection. They appear, respectively, in 12 and 10% of cases. Table 2 shows the three main architectures selected by SSM for 60 simulations. It can be noted that this strategy provides for the same structures and that the real structure is present in 62% of cases.



**Fig. 1** Real architectural foundation

## 6 The Conclusions

The main findings of this article can be extended in several directions. First, the exogenous variables can be added as inputs to the network, and the so-called nonlinear autoregressive model with exogenous inputs (NARX) is studied. Then, the delays can be fed into the input error to define a nonlinear type of automatic regression moving average model (ARMA). Finally, all results can be extended to the case of a multipart output. He showed in his work that it is necessary to reduce the determinant in this case of the covariance matrix of experimental error. All these results were generated and then tested on simulated examples and on real data, such as the spot series, daily electricity consumption, different real or simulated series, but also recently on ozone level pollution data for several places. Be careful, all this applies only to stationary time series and the use of neural networks does not eliminate the need for traditional preprocessing, it is necessary to reanimate the directions and frequency. There are many programs for obtaining linear recording and estimating neural network parameters from a regression model or time series analysis. The SSM method (based on SIC criteria) is used to select the best architectural design.

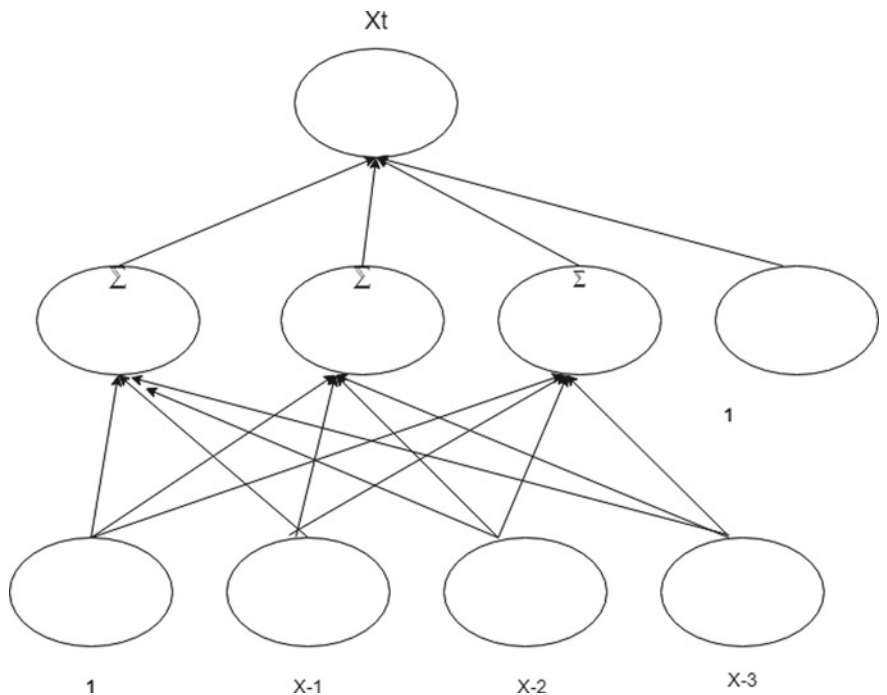


Fig. 2 Architecture after training

Table 1 Comprehensive search performance

Final over 60 simulations	Architecture percentage
A	0.73
B	0.12
C	0.1

Table 2 SSM performance

Final over 60 simulations	Architecture percentage
A	0.62
B	0.22
C	0.16

## References

1. Manouchehrian A, Sharifzadeh M, Moghadam RH (2012) Application of artificial neural networks and multivariate statistics to estimate UCS using textural characteristics. *Int J Min Sci Technol* 22(2):229–236. <https://doi.org/10.1016/j.ijmst.2011.08.013>
2. Zhu Z et al (2018) Performance prediction of an aerobic granular SBR using modular multilayer artificial neural networks. *Neurocomputing* 8(1):1–12. <https://doi.org/10.1016/j.neunet.2018.08.014>
3. Manouchehrian A et al (2018) Walking gait event detection based on electromyography signals using artificial neural network. *Neurocomputing* 8(1):1–12. <https://doi.org/10.1186/s13635-018-0073-z>
4. Zhu Z, Zhu X, Kong F, Guo W (2018) A rapid method on identifying disqualified raw goat's milk based on total bacterial count by using dielectric spectra. *J Food Eng* 239(June):40–51. <https://doi.org/10.1016/j.jfoodeng.2018.06.020>
5. Nguyen T, Nguyen T, Nguyen BM, Nguyen G (2019) Efficient time-series forecasting using neural network and opposition-based coral reefs optimization. *Int J Comput Intell Syst* 12(2):1144–1161. <https://doi.org/10.2991/ijcis.d.190930.003>
6. Cirstea R-G, Micu D-V, Muresan G-M, Guo C, Yang B (2018) Correlated time series forecasting using deep neural networks: a summary of results. Retrieved from <http://arxiv.org/abs/1808.09794>
7. Zhang X, Zhang Q, Zhang G, Nie Z, Gui Z (2018) A hybrid model for annual runoff time series forecasting using elman neural network with ensemble empirical mode decomposition. *Water (Switzerland)* 10(4). <https://doi.org/10.3390/w10040416>
8. Casado-Vara R, del Rey AM, Pérez-Palau D, De-La-fuente-valentín L, Corchado JM (2021) Article web traffic time series forecasting using LSTM neural networks with distributed asynchronous training. *Mathematics* 9(4):1–22. <https://doi.org/10.3390/math9040421>
9. Javeri IY, Toutiaee M, Arpinar IB, Miller TW, Miller JA (2021) Improving neural networks for time series forecasting using data augmentation and AutoML. Retrieved from <http://arxiv.org/abs/2103.01992>
10. Waheeb W, Ghazali R (2016) Multi-step time series forecasting using ridge polynomial neural network with error-output feedbacks. *Commun Comput Inf Sci* 652:48–58. [https://doi.org/10.1007/978-981-10-2777-2\\_5](https://doi.org/10.1007/978-981-10-2777-2_5)
11. Walczak S (2001) An empirical analysis of data requirements for financial forecasting with neural networks. *JMIS*
12. Zhou J, Peng T, Zhang C, Sun N (2018) Data pre-analysis and ensemble of various artificial neural networks for monthly streamflow forecasting. *Water (Switzerland)* 10(5). <https://doi.org/10.3390/w10050628>
13. Hadwan M, Al-Maqaleh BM, Al-Badani FN, Khan RU, Al-Hagery MA (2022) A hybrid neural network and box-jenkins models for time series forecasting. *Comput Mater Continua* 70(3):4829–4845. <https://doi.org/10.32604/cmc.2022.017824>
14. Jin J, Kim J (2015) Forecasting natural gas prices using wavelets, time series, and artificial neural networks. *PLoS One* 10(11). <https://doi.org/10.1371/journal.pone.0142064>
15. Affan MF, Abdullah AG, Surya W (2019) Forecasting of wind speed using exponential smoothing and artificial neural networks (ANN). *J Phys: Conf Ser* 1402(3). <https://doi.org/10.1088/1742-6596/1402/3/033082>
16. Mozo A, Ordozgoiti B, Gómez-Canaval S (2018) Forecasting short-term data center network traffic load with convolutional neural networks. *PLoS One* 13(2). <https://doi.org/10.1371/journal.pone.0191939>
17. Suhartono, Amalia FF, Saputri PD, Rahayu SP, Suprih Ulama BS (2018) Simulation study for determining the best architecture of multilayer perceptron for forecasting nonlinear seasonal time series. *J Phys: Conf Ser* 1028(1). <https://doi.org/10.1088/1742-6596/1028/1/012214>
18. Mapuwei TW, Bodhlyera O, Mwambi H (2020) Univariate time series analysis of short-term forecasting horizons using artificial neural networks: the case of public ambulance emergency preparedness. *J Appl Math* 2020. <https://doi.org/10.1155/2020/2408698>

19. Wu W, An SY, Guan P, Huang DS, Zhou BS (2019) Time series analysis of human brucellosis in mainland China by using Elman and Jordan recurrent neural networks. *BMC Infect Dis* 19(1). <https://doi.org/10.1186/s12879-019-4028-x>
20. Tsakiri K, Marsellos A, Kapetanakis S (2018) Artificial neural network and multiple linear regression for flood prediction in Mohawk River, New York. *Water (Switzerland)* 10(9). <https://doi.org/10.3390/w10091158>
21. Hansen JV, Nelson RD (2003) Forecasting and recombining time-series components by using neural networks. *J Oper Res Soc* 54(3):307–317. <https://doi.org/10.1057/palgrave.jors.2601523>
22. Tadayon M, Iwashita Y (2020) A clustering approach to time series forecasting using neural networks: a comparative study on distance-based vs. feature-based clustering methods. Retrieved from <http://arxiv.org/abs/2001.09547>
23. Isfan M, Menezes R, Mendes DA (2010) Forecasting the Portuguese stock market time series by using artificial neural networks. *J Phys: Conf Ser* 221. <https://doi.org/10.1088/1742-6596/221/1/012017>
24. Lara-Benítez P, Carranza-García M, Luna-Romera JM, Riquelme JC (2020) Temporal convolutional networks applied to energy-related time series forecasting. *Appl Sci (Switzerland)* 10(7). <https://doi.org/10.3390/app10072322>
25. Mahto AK, Alam MA, Biswas R, Ahmad J, Alam SI (2021) Short-term forecasting of agriculture commodities in context of Indian market for sustainable agriculture by using the artificial neural network. *J Food Qual* 2021. <https://doi.org/10.1155/2021/9939906>
26. Hirata Y, Aihara K (2017) Improving time series prediction of solar irradiance after sunrise: comparison among three methods for time series prediction. *Sol Energy* 1(3):149–294
27. Al-Nuaimy L (2005) Enhanced artificial neural networks model based on a single layer linear counter propagation for prediction and function approximation. *Egypt Comput Sci J* 27(2):46–54
28. Al-Nuaimy L (2016) Muscat securities market index (MSM30) prediction using single layer linear counterpropagation (SLIC) neural network. In: 2016 3rd MEC international conference on big data and smart city (ICBDSC), pp 1–5. <https://doi.org/10.1109/ICBDSC.2016.7460366>
29. Al-Nuaimy L (2003) Feedback matching to predict the time series. In: The first scientific conference for computer in Irbid Privat University
30. Septiarini TW, Taufik MR, Afif M, Rukminastiti Masyrifah A (2020) A comparative study for Bitcoin cryptocurrency forecasting in period 2017–2019. *J Phys: Conf Ser* 1511(1). <https://doi.org/10.1088/1742-6596/1511/1/012056>
31. Zhu S et al (2018) Artificial neural network enabled capacitance prediction for carbon-based supercapacitors. *Mater Lett* 233:294–297. <https://doi.org/10.1016/j.matlet.2018.09.028>

# An Approach for Test Impact Analysis on the Integration Level in Java Programs



Muzammil Shahbaz

**Abstract** Test impact analysis is an approach to obtain a subset of tests impacted by code changes. This approach is mainly applied to unit testing where the link between the code and its associated tests is easy to obtain. On the integration level, however, it is not straightforward to find such a link programmatically, especially when the integration tests are held into separate repositories. We propose an approach for selecting integration tests based on the runtime analysis of code changes to reduce the test execution overhead. We provide a set of tools and a framework that can be plugged into existing CI/CD pipelines. We have evaluated the approach on a range of open-source Java programs and found  $\approx 50\%$  reduction in tests on average, and above 80% in a few cases. We have also applied the approach to a large-scale commercial system in production and found similar results.

**Keywords** Test impact analysis · Regression test selection · Continuous integration · Continuous testing

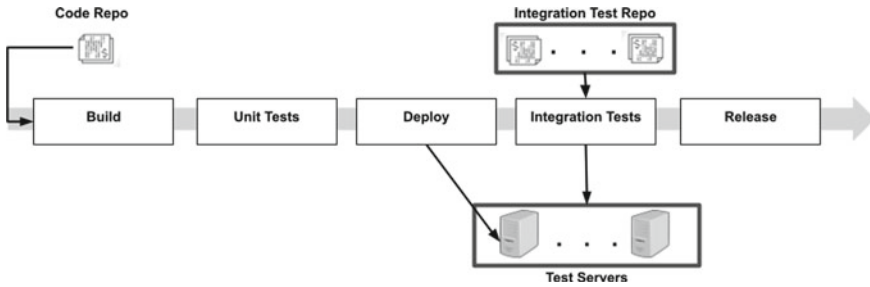
## 1 Introduction

Continuous integration/delivery (CI/CD) [11] is a modern way of building and releasing software in a shorter development lifecycle. Developers continuously implement small changes into their code and merge back to the main branch of the repository as often as possible and many times a day. These changes are validated by the automated system of building, testing, and deployment of software, which makes the integration easier and delivery quicker. Running automated integration tests is a vital phase of CI/CD that makes sure that the new changes to the code do not break the software or acceptance criteria whenever new merges are committed to the main branch. Moreover, they also serve as a regression test suite, verifying that no bugs are introduced into the existing behavior by the new changes.

---

M. Shahbaz (✉)

Thales UK Ltd Cheadle Heath, Stockport SK3 0XB, UK  
e-mail: [muzammil.shahbaz@uk.thalesgroup.com](mailto:muzammil.shahbaz@uk.thalesgroup.com)



**Fig. 1** Typical stages of a CI/CD pipeline

Figure 1 shows a typical CI/CD pipeline of a software component that is moved across various stages in the development lifecycle before it is released. The pipeline is triggered when some code changes are committed into a repository of a version control system. The code goes to the build phase where it is compiled, followed by running the unit tests. Given there are no failures in the build, the component moves to the next phase where it is deployed in staging or a test server. Thereby, the integration or acceptance tests are checked out from the test repository and run on the component along with the other components in the system collectively. When all the integration tests are passed, the component is tagged or promoted to release.

## 1.1 Motivation

As software matures with time, the integration test suite tends to grow to the extent that running all tests takes a considerable amount of time. This is especially true when the software is delivered via iterative process, where the first release might contain a minimum functionality, but its codebase grows manifolds due to adding new features and fixing defects. This increases the rate of code changes, for instance, Google's code churn rate—a commit every second on average—produces 800 K builds with 150 million tests runs daily [16]. Even with the company's massive compute resources, it is not cost-effective to test each code commit individually at this rate. In practice, developers delay their merges until tests are completed in the pipeline for each of the prior changes. This eventually slows down the whole pipeline process and diminishes the benefit of automation. Developers tackle this situation by either disabling some tests, or deferring them to the very end of the release cycle, which again results in lower productivity and quality.

## 1.2 Test Impact Analysis

The basic principle of *test impact analysis (TIA)* [7, 9, 19] is to determine the subset of tests that is impacted by the code change. Thus, running only the subset results into faster execution of the pipeline.

Finding the subset of impacted tests is easy on a component level where code and unit tests are kept in the same repository. Modern IDEs and tools for structural coverage can efficiently compute call graphs of unit tests either statically or dynamically. These call graphs can trace the tests back to the methods changed, and then, only those unit tests can be selected to run.

On the integration level, however, it is a different story. The integrated tests cover the whole system and execute many components at once or in a particular order. There might not be a direct relation between an integration test and a method changed in one of those components. This is a typical case of **microservices** [12], where a monolithic application is replaced by a suite of small services that are built independently. They are implemented in different repositories and run on different servers and communicate via different network protocols. Thus, there is no easy way of determining the (indirect) dependencies between the tests and the new commits in various components.

We propose an approach for test selection that combines static and dynamic analysis of Java components irrespective of their organization into repositories or system topology. We have implemented the approach into existing CI/CD pipelines, so that it only runs a subset of tests required to validate the code being committed without losing quality, i.e., the outcome of the tests that are not selected will not be affected by the changes. Trivially, it results into faster execution of the pipeline because for a given code commit, our approach selects and runs only the relevant tests required to validate that commit.

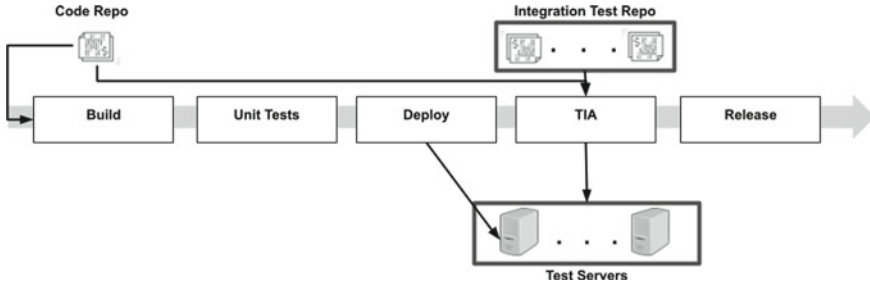
The rest of the paper is organized as follows. Section 2 explains the approach with formal settings. Section 3 presents the experimental evaluation. Section 4 discusses the existing works related to this topic. Section 5 concludes the paper with a note on future works.

## 2 Methodology

### 2.1 Overview

In order to perform the impact analysis, we need to understand the relationship between the existing tests and the source code. Our approach is to “sniff” the method invocations at the test execution runtime. At the end of testing, we produce a map that links each test with the associated list of methods invoked. For instance, we have a map of three tests and the methods they call:





**Fig. 2** Stages of the CI/CD pipeline with TIA adjusted

$\text{map}(T1) = \{A.m1, B.m2, C.m1\}$

$\text{map}(T2) = \{A.m2, B.m1\}$

$\text{map}(T3) = \{B.m2, C.m2\}$

That means, the test T1 calls the method m1 of class A, m2 of class B, and m1 of class C. Similarly, the test T2 calls the method m2 of class A and m1 of class B, and so on. We store this map in the integration tests repository. In the next code commit, we compute the difference with the previous commit to analyze code changes. Thereby, we determine which tests are actually impacted by looking at this map and the methods that are changed in the last commit. Suppose the next commit changes the method m2 of the class B, then we perform the lookup into the map and pick tests T1 and T3; the only tests that are impacted. We leave T2 as it is not relevant for this change. Our approach is *safe* [26] in the sense that the set of selected tests contains all tests whose behavior may have been affected by the change in the methods.

Figure 2 shows the pipeline with TIA adjusted. We define the concepts used in the approach in the following section and then explain the implementation stages of the approach.

## 2.2 Formal Settings

We assume the source code consists of classes, and each class has a set of implemented methods. We consider explicit constructors as methods. We also assume a set of tests that call a subset of those methods when they are executed.

Let  $M$  be the set of all methods and  $T$  be the set of all tests. We assume that  $M$  and  $T$  are non-empty sets, i.e., there are one or more methods in the source code and one or more tests that execute a subset of those methods. For this purpose, we define a function  $\lambda : T \times M \rightarrow \{1, 0\}$ , such that  $\lambda(t, m) = 1$  iff the method  $m$  is executed by the test  $t$ , and  $\lambda(t, m) = 0$ , otherwise. We also use a *map* for each test and the methods it executes defined as  $T \rightarrow 2^M$  such that each  $t \in T$  maps to a set  $\{m_1, \dots, m_n\} \subseteq M$  where  $\lambda(t, m_i) = 1$  for  $1 \leq i \leq n$ .

A method declaration has six components: name, arguments, return type, modifiers, exception list, and an implementation body [15].

A method is called **modified** if the method is (1) changed, (2) removed, or (3) added. Each of these types are explained below.

**Changed** A method is changed if its modifiers, exception list or implementation body is changed. Thus, we create the following rule to track method changes.

**Rule 1** *A method is marked modified if any of the following changes occur*

- *implementation body, e.g., adding/removing/changing statements.*
- *modifiers, e.g., adding/removing access modifiers (including `static`) or keywords like `synchronized`, `throws` or `abstract`.*
- *exception list, e.g., adding/removing checked exceptions.*

Changes to return types do not pose a particular challenge to our methodology. This is because the return statement in the implementation body is changed almost every time the return type is changed. The only exception to changing the return type without changing the return statement is the *covariant* return type [15], which does not impact the semantic change without changing the implementation.

Changes to a method name or arguments force changes into other methods which call the changed method. In this case, the other methods will be marked as modified as per Rule 1.

If a method has not been called by any other method, but changing its name or arguments cause it to override/overload an existing method, then it can alter the behavior of the program. Therefore, any changes to the name or arguments are considered as the method is removed and a new method is added. Both of these cases are covered in the next sections.

**Removed** A method is removed if its whole declaration is deleted, or changes to its name or arguments occur. Thus, we create the following rule to track method removals.

**Rule 2** *A method is marked modified if any of the following changes occur*

- *declaration is removed.*
- *name is changed.*
- *arguments (including number, order, or types) are changed.*

All methods having explicit references to a removed method would require changes to its implementation. Those methods will also be marked *modified* as per Rule 1. Removing a method may also cause some tests to be invalid if they have an explicit reference to the method. Any adjustments to tests due to a source code change are not in the scope of our methodology.

**Added** Any new methods likely result into changes to existing methods or tests. However, if the new method overrides or overloads an existing method, it may alter the behavior of tests without changing the existing code.

Program	Tests	Mappings
<pre> class A {     void foo() {     }     void bar(Object obj) {     } }  class B extends A {     void foo() {     } } </pre>	<pre> // T1 test1() {     A a = new A();     a.foo(); }  // T2 test2() {     A a = new B();     a.foo(); }  // T3 test3() {     A a = new B();     a.bar("hello"); } </pre>	<pre> // map(T1) { A.A(), A.foo() }  // map(T2) { B.B(), A.A(), B.foo() }  // map(T3) { B.B(), A.A(),   A.bar(Object) } </pre>

**Fig. 3** Example: original program (left) and its associated tests with mappings (right)

If the new method overrides an existing method, the new method may be called in lieu of the existing method due to dynamic binding.<sup>1</sup> In this case, we mark the existing method as *modified*. We create the following rule to cover this case.

**Rule 3** *If a method  $\alpha$  is a new method such that it overrides an existing method  $\beta$ , then  $\beta$  is marked modified.*

If the new method overloads an existing method such that it matches the name, number, order, and type of arguments with another method in the class hierarchy, then it can change the static binding of the method in the other class.<sup>2</sup> This is because the binding of such a method depends upon the compile-time types of the arguments [15]. Therefore, all methods in the class hierarchy that has the same name, number, order, and type of arguments are marked as *modified*. We create the following rule to cover this case.

**Rule 4** *If a method  $\alpha$  is a new method such that it matches the name, number, order, and (super or sub) type of arguments with a method  $\beta$  in any class in the hierarchy, then  $\beta$  is marked modified.*

## 2.3 Example

We explain the approach with the help of the example in Fig. 3. The original program consists of class *A* and its subclass *B*. Three tests, *T1*, *T2*, and *T3*, and their associated maps of method calls can also be seen. The program goes through four versions of changes as shown in Fig. 4. The figure also shows “modified methods” in each version, “selected tests” based on those methods, and “new mappings” calculated

<sup>1</sup> Version 1 in Fig. 4 provides an example of such a case.

<sup>2</sup> Version 2 in Fig. 4 provides an example of such a case.

Version 1: Overridden <i>B.bar(Object)</i> added	Version 2: Overloaded <i>A.bar(String)</i> added	Version 3: Changed <i>B.foo()</i> changed	Version 4: Removed <i>A.foo()</i> removed
<pre> class A {     void foo() {     }     void bar(Object obj) {     } }  class B extends A {     void foo() {     }     void bar(Object obj) {     } } </pre>	<pre> class A {     void foo() {     }     void bar(Object obj) {     }     void bar(String text) {     } }  class B extends A {     void foo() {     }     void bar(Object obj) {     } } </pre>	<pre> class A {     void foo() {     }     void bar(Object obj) {     }     void bar(String text) {     } }  class B extends A {     void foo() {         bar("hello");     }     void bar(Object obj) {     } } </pre>	<pre> class A {     void bar(Object obj) {     }     void bar(String text) {     } }  class B extends A {     void foo() {         bar("hello");     }     void bar(Object obj) {     } } </pre>
Modified Methods			
<i>A.bar(Object)</i>	<i>A.bar(Object)</i> <i>B.bar(Object)</i>	<i>B.foo()</i>	<i>A.foo()</i>
Selected Tests			
<i>T3</i>	<i>T3</i>	<i>T2</i>	<i>T1</i>
New Mappings			
<pre> // map(T1) { A.A(), A.foo() }  // map(T2) { B.B(), A.A(), B.foo() }  // map(T3) { B.B(), A.A(),   B.bar(Object) } </pre>	<pre> // map(T1) { A.A(), A.foo() }  // map(T2) { B.B(), A.A(), B.foo() }  // map(T3) { B.B(), A.A(),   A.bar(String) } </pre>	<pre> // map(T1) { A.A(), A.foo() }  // map(T2) { B.B(), A.A(), B.foo(),   A.bar(String) }  // map(T3) { B.B(), A.A(),   A.bar(String) } </pre>	<pre> // map(T1) N/A — T1 cannot execute  // map(T2) { B.B(), A.A(), B.foo(),   A.bar(String) }  // map(T3) { B.B(), A.A(),   A.bar(String) } </pre>

**Fig. 4** Example: illustration of four versions of the original program in Fig. 3 and test selection with new mappings in each iteration

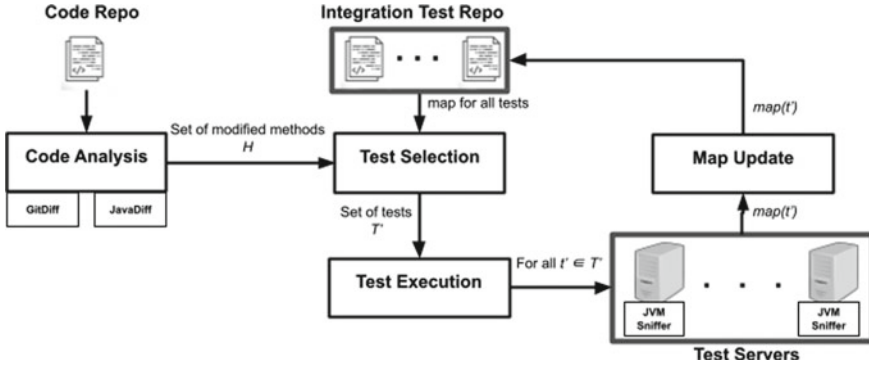
after running those tests. We have shown the map of all tests in each iteration for completeness.

Version 1 adds a new method *B.bar(Object)* that overrides *A.bar(Object)*, which marks it a modified method (as per Rule 3). We find this method in *map(T3)* and therefore *T3* is selected. The new map for *T3* is obtained after running the test on Version 1 that now includes *B.bar(Object)*.

Version 2 overloads *A.bar(Object)* by adding *A.bar(String)*. Note that the new method has the same name, number, order, and type of arguments as the original method. The only difference is that the new method uses a subtype of the existing method. This alters the behavior of *T3* which now resolves to the new method (instead of *B.bar(Object)*) due to the explicit *String* type argument (i.e., “hello”) in the test. We mark *A.bar(Object)* and *B.bar(Object)* as modified methods (as per Rule 4). The latter is included in *map(T3)*, which was updated after Version 1. Thus, *T3* is selected once again, and the new map is updated that now includes *A.bar(String)*.

Version 3 changes the body of *B.foo()*, which is therefore a modified method (as per Rule 1). As this is included in *map(T2)*, *T2* is selected and map is updated consequently.

Version 4 removes *A.foo()*, which is therefore a modified method (as per Rule 2). As this is included in *map(T1)*, *T1* is selected. However, this test cannot be executed because of the explicit reference to the non-existent method and hence needs revision.



**Fig. 5** Stages of jTIA: (1) code analysis, (2) test selection, (3) test execution/map update

## 2.4 Implementation Stages

We have implemented our TIA approach into a pipeline plugin called jTIA, in addition to a set of tools [21, 22] supporting the different stages of the approach. Figure 5 illustrates these stages with the explanation below.

**Code Analysis:** The first stage of jTIA performs code change analysis by checking out the source code repository. The list of methods  $H$  that are modified in the last commit is obtained and passed on to the next stage.

Our current implementation deals only with GIT repositories for identifying code changes. For this purpose, we have implemented a tool called *GitDiff* [21], which is responsible for managing GIT repositories (e.g., retrieving change logs). *GitDiff* retrieves the list of changes in the last commit by executing GIT commands<sup>3</sup> and then parse the output to find the new/changed/deleted Java files.

Our tool *JavaDiff* [21] is responsible for analyzing Java files without regard to formatting, comments, whitespaces, or method ordering. The tool implements the four rules defined in Sect. 2.2. For a changed or deleted Java file, the previous version is checked out<sup>4</sup> and compared to obtain the list of any methods modified. For new Java files, changes in the existing class hierarchy are reported for any overloaded or overridden methods. For this purpose, it uses the call graph analysis of the *Soot* Java optimization framework [24].

**Test Selection:** The second stage of jTIA selects tests based on the list of modified methods from the previous stage. The integration tests repository is checked out, and the mapping for each test to its associated methods is obtained. The test that contains a modified method in its mapping is selected. Any test that has no mappings, such as newly added tests, is also selected. This means that all tests are selected when the pipeline is triggered for the very first time. Any test that has been changed since its

<sup>3</sup> `git diff-tree --no-commit-id --name-only -r HEAD`

<sup>4</sup> `git show HEAD:<file>`

last execution, or previously failed will have its map deleted, i.e., all changed and failing tests will also be selected.

Formally, if  $H \subseteq M$  is the set of modified methods and  $T$  is a set of tests, then the set of selected tests  $T' \subseteq T$  is computed in the following way. For each test  $t \in T$  and a modified method  $h \in H$ ,  $t$  is added to  $T'$  if

1.  $map(t) = \emptyset$ , or
2.  $h \in map(t)$

**Test Execution and Map Update:** The third stage of jTIA runs tests in  $T'$  one by one on the test servers. For each test  $t' \in T'$ ,  $map(t')$  is recomputed in order to capture any changes after the commit.

Our tool *JVMSniffer*[22] is responsible for capturing the method calls for each running test. The tool uses Java Debug Interface API<sup>5</sup> to hook into the target JVM in the debug mode. It provides the ability to trap target events, including method entry and exit events. Therefore, it is able to capture all method invocations on the target JVM objects including types and values of the arguments. It also provides an option to filter method calls based on the package name, which is useful in collecting the information only related to the system under test and avoiding method call traces from third-party or system libraries.

*JVMSniffer* is deployed per JVM on each test server. Before running a test, jTIA starts *JVMSniffer* on each server that attaches itself to the target JVM. jTIA then starts test execution, and *JVMSniffer* starts recording all method invocations during the test execution. For each method invocation, it records the name of the package, class, method, its argument list, and return type. When the test is complete, jTIA stops all instances of *JVMSniffer* and collates the method invocation records from each server. Finally, jTIA updates the test map with the list of method calls and checks into the integration tests repository.

### 3 Evaluation

In this section, we evaluate our approach and discuss results, threats to validity of our approach and its limitations.

#### 3.1 Assumptions

We assume each Java component in the system which has at least one public class that implements at least one public method. All tests run under identical conditions, i.e., our test environment, resource configuration, environment variables, and external

---

<sup>5</sup> <https://docs.oracle.com/javase/8/docs/technotes/guides/jpda/>.

**Table 1** Case studies with versions and lines of code in thousands (kLoC)

Case studies	Version	kLoC		
		Total	src	Test
Apache Commons Col.	4.4	92	51	41
Apache Commons Lang	3.6	112	59	53
Eclipse Collections	10.0.0	327	171	156
Netty	3.10	105	87	18
RxJava	2.2.0	329	134	195
Seata	0.5.0	32	24	8
OkHttp	4.0.0	47	11	36
MyBatis	3.4.0	61	25	36
UML Reverse Mapper	1.4.4	2	1	1
Symja Parser	1.0.0	9	7	2

dependencies remain unchanged to produce deterministic results if we run the same test suite multiple times on the same code base.

### 3.2 Case Studies

The experiments have been conducted on well-known open-source projects from GitHub<sup>6</sup> that were selected based on their popularity and repository size. Table 1 lists the case studies with their versions and size in terms of lines of code.

### 3.3 Research Questions

*RQ1:* Are the selected tests from the proposed approach as effective as the whole test suite?

To analyze the effectiveness of our approach, we used mutation testing that provides a reliable metric to measure the quality of a test suite [1]. We used *PIT* mutation testing tool [5] in our experiments. The mutants are created based on code coverage, i.e., only the methods covered by the tests are mutated. PIT generates mutants via number of mutation operators targeting various semantics in Java. The details of each operator can be seen on PIT's Website.<sup>7</sup> The selected operators in our experiments are

<sup>6</sup> <https://github.com>.

<sup>7</sup> <http://pitest.org/quickstart/mutators/>.

- Default:** Conditional Boundary, Increments, Invert Negatives, Math, Negate Conditionals, Void Method Calls, Empty/False/True/Null/Primitive Returns.
- Optional:** Constructor Calls, Inline Constant, Non Void Method Calls, Remove Conditionals, Remove Increments.

First, we ran the original program with PIT in order to compute the mutation coverage with the whole test suite. Then, we changed the tool's configuration to run only the tests selected by our approach and computed the mutation coverage again. We compared the mutation coverage between the two runs by looking at the number of mutants killed. Ideally, the mutation coverage of the selected test suite should be equal to the mutation coverage of the entire test suite, otherwise the selected test suite is not complete with respect to the changes introduced to the program due to mutations.

We used default settings of operators to generate as many mutations possible depending on the size of the project. There are multiple reasons for which a mutant can be considered as killed, all of which were considered as fault detection.

- The test was executed and failed.
- The test was executed and caused a memory error, e.g., stack overflow.
- The test terminated due to a predefined timeout (i.e., 3 s).

*RQ2:* What is the gain in terms of the number of tests reduced?

The effectiveness of our approach is measured by the ratio of number of tests selected over the total number of tests. That is:  $\text{Gain} = 1 - (100 \times |T'|/|T|)$ .

### 3.4 Results

*RQ1:* Are the selected tests from the proposed approach as effective as the whole test suite?

Table 2 shows the number of mutations generated for each case study in the first column. The second column shows the mutants killed by the whole test suite. The third column shows how many mutants were missed by the reduced test suite.

The results confirm that all mutants that were killed by the whole test suite were also killed by the selected test suite. The only exception was *Apache Commons Collections* where 3% mutants survived. The detailed analysis showed that PIT added some mutations to the static initializers in the Java code. The fact that the *JVMSniffer* only traces method calls, the tests that cause those blocks to be executed were not selected. When adding those tests manually, the number of mutants missed dropped to zero.

*RQ2:* What is the gain in terms of the number of tests reduced?

Table 3 shows the total number of tests in the whole test suite and the selected test suite in the first and second columns, respectively. The third column shows the percentage reduction in tests. As per the results in Table 2, running only the



**Table 2** Comparison of mutation coverage

Case studies	# of mutations	Killed by whole test suite (%)	Missed by selected tests (%)
Apache Commons Col.	348	37	3
Apache Commons Lang	1360	53	0
Eclipse Collections	29	53.5	0
Netty	1363	33	0
RxJava	329	72	0
Seata	192	17	0
OkHttp	543	55.67	0
MyBatis	169	84	0
UML Reverse Mapper	27	55	0
Symja Parser	182	32	0

selected test suite is sufficient for the changes introduced due to mutations. There is a significant reduction in the number of tests averaging  $\approx 50\%$ . Eight out of ten case studies achieved double digit reduction (and above 80% in a few cases). The only exception where no reduction was possible is *Symja Parser*, which is a math utility library. Due to the nature of functions implemented in this library and the fact that we have used math related operators in PIT, almost all methods were mutated. Therefore, all tests in the whole test suite were selected by our approach.

### 3.5 Threats to Validity

The case studies used in the experiments are real-life projects, but they are not representative of a large integrated system of components. As with any type of evaluation, additional case studies will be required for gaining further confidence in our approach. Having said that, our objective was to evaluate the approach for test selection based on code changes, for which, these projects provide a sufficient ground. All of these projects are in active development and widely used in the open-source community. We believe that they provide realistic and diverse examples of what our approach needs to be able to handle in practice.

One threat to validity might come from the mutation strategy used in our experiments. In order to reduce bias, we used the default settings of PIT's operators for all case studies. We repeated each experiment 3 times and found the same result. This is due to the fact that PIT always generates same mutants for an unchanged code. Another related threat is the presence of equivalent mutants. PIT tries best to ignore such mutants using a built-in bytecode matching library and removes mutants that

**Table 3** Tests minimization gain against the whole test suite

Case studies	# of tests	# of selected tests	Gain (%)
Apache Commons Collections	3900	744	80.92
Apache Commons Lang	3039	1814	40.31
Eclipse Collections	178	61	87.24
Netty	1213	956	21.19
RxJava	3571	1917	46.32
Seata	245	47	80.81
OkHttp	1241	1125	9.34
MyBatis	878	617	29.73
UML Reverse Mapper	67	16	76.12
Symja Parser	83	83	0
Average	14,593	7380	49.43

are trivially equivalent. Nevertheless, equivalent mutants do not pose a particular problem in our experiments as we compare results on the same set of mutants.

Another threat may come from the implementation of *JVMSniffer* that how accurately it computes the mapping between tests and method executions. The tool uses Java Debug Interface API that has been around since Java 1.4 and considered stable. To minimize the risk, we checked the mapping with the call graph generated by *IntelliJ IDEA*<sup>8</sup> and found no discrepancies.

### 3.6 Limitations

The implementation of our approach has following limitations.

**Sequential execution:** The approach is designed to run tests in a distributed environment, and multiple JVMs are supported. However, tests must execute sequentially, otherwise *JVMSniffer* cannot distinguish which method invocation is caused by which test.

**Method level changes:** Changes to field initializations, static blocks, and constants are currently not considered. This limitation is not fundamental to our approach as most regression test selection techniques [13, 17, 18, 20] are applied on the method level, in contrast to coarser-grained approaches [4, 6], which are based solely on Java file/class level changes.

## 4 Related Works

Traditional regression test selection techniques compute detailed program changes by traversing control flow graphs of two versions of a program and their associated tests at different code granularities, such as statement block, method, class, or file levels [26]. Then, the tests which overlap the elements of the graph are selected for execution. Such techniques require a lot of processing and incur non-trivial overhead [27]. Therefore, researchers have also proposed techniques that are applied only on the method level granularity with the combination of static and dynamic analysis.

One of the early works in this realm is *Chianti* [20] that relies on the computation of structural differences between the two versions of a Java program. It computes abstract syntax tree (AST) of the source program, followed by identifying changes using predefined rules. It generates control flow graph (CFG) to associate tests with the method changes and determines a subset of tests relevant to the changes. Our approach is similar to *Chianti* with respect to applying rules to cater the language specific semantics and the concepts of object-oriented programming (e.g., polymorphism). Our approach is different in how the test selection process works. We do not

---

<sup>8</sup> <https://www.jetbrains.com/idea/>.

compute AST or CFG to determine the test association, which could be computationally expensive. We obtain this information dynamically from the test execution and create the mapping (or association) between the tests and the methods. This fits well with the CI/CD methodology where the code merges are frequent, and therefore, computing such trees or graphs could have a detrimental effect on the build timescales [27]. In our case, there is no overhead of creating this mapping. The only pre-processing required in our approach is the identification of method changes at the bytecode level (via *Soot* optimization framework [24]). We argue that this is a significantly less overhead compared to computing a full CFG of the source code as well as the tests. The main reason is that we use the CFG analysis only if a newly added method has the same method signature as an existing method; so the method override/overload rules, i.e., rules 3/4, could be applied.

In contrast to static analysis techniques, dynamic techniques have been used for impact analysis to support regression testing [13, 17, 18]. Orso et al. [17, 18] collect execution traces from program instrumentation. The execution traces consists of all methods that were called during program execution. For each code change, they compute an approximate dynamic slice based on the execution traces that traverse the change. The impact set is the union of the slices computed for each change. Separately, they select an initial set of tests that traversed at least one change based on coverage information. Then, they use the impact set from the execution traces to assess whether, according to the user supplied information (which the authors called “field data”), the initial set of tests is adequate, i.e., there exists at least one test that traverses the affected method after traversing the change. This step is based on the intuition that the executions that traverse changed parts of the code are more likely to traverse the affected methods. The main difference between their work and ours is that their goal is to determine which affected methods are not tested by the given set of test cases. On the contrary, our aim is to find the test cases that are affected by those changes. Secondly, their technique is not safe as they rely on coverage data from static analysis [20]. In contrast, we collect test mapping from execution traces. Thirdly, they rely on program instrumentation, whereas we do not require any changes to the source code for our implementation.

Lehnert et al. [14] used UML state machines to identify impacted test cases. In their approach, a test path is a sequence of state machine transitions, which is used to test the class that corresponds to the state machine. For every method changed, the corresponding transition in the state machine which uses this method in its events, guards, or actions is marked as modified. The test sequences that contain this transition are then selected. The underlying hypothesis is that the changes can be identified through design models as the equivalence between the UML models and the implementation [10]. Therefore, they introduced specific rules to perform impact analysis on the changed model elements and to build traceability links between model elements and test cases. Cazzola et al. [3] noted that UML behavioral models such as state machines are often incomplete or quickly outdated, which limits the applicability of existing model-based approaches. As an improvement, they proposed UML class diagrams as a more reliable modeling source and applied fuzzy logic to deal with inconsistencies. Similarly, Sun et al. [23] used *OOCMDG*—a dependency graph

of classes, methods, and variables to create impact rules. The rules identify the impact set according to the change types and then computing the union of all change types for each entity. However, their approach does not work for method body changes and statement changes [14]. In contrast to all these works, we do not rely on any design models or documentation to identify changes. Moreover, our aim is to select test cases that are already implemented as a code rather than represented as abstract models.

Azizi and Do [2] used information retrieval (IR) techniques to select tests during regression testing. They identify common keywords in the difference of two program versions and construct IR queries using the keywords. They also tokenize the source code of the tests to compute their diversity and build a graph to represent the relationship among tests annotated with the diversity score. Then, they collect tests whose cosine similarity to the queries is higher than a predefined threshold. Finally, they select the most diverse tests among the collected tests. As normally the case with IR techniques, this solution only works if the source code of the program and tests use mostly the same keywords.

Symbolic execution techniques have also been used in regression test selection [8, 25]. Yang et al. [25] leveraged symbolic execution with static analysis to identify program statements relevant to changes in the program. They used static analysis to determine the modified program statements; then, only the *affected* parts of the program were explored during symbolic execution. While the goal of this technique is to make symbolic execution efficient in program analysis, Guo et al. [8] argued that this may leave out some paths if the affected paths are equivalent to some paths in the previous version. They demonstrated their findings on an example of a concurrent program that is executing two threads, where a change in the code of one thread is not identified in the other thread. Hence, the path is not selected in the symbolic exploration. Our approach is not affected by this shortcoming because the test executing the two threads is selected if a change in the code of any of the threads is detected. Nevertheless, the most important drawback of the symbolic execution approach is the algorithmic complexity that can be very expensive [26].

## 5 Summary and Future Works

This paper presents an approach for test impact analysis on the integration level, where the source code is distributed in multiple repositories. The approach computes program changes statically and selects the impacted tests based on their association with method calls dynamically. We have used mutation testing to evaluate our approach on a number of open-source Java projects. Firstly, we ran the mutated programs with the whole test suite then executed the tests selected by our approach to compute the effectiveness. The results achieved  $\approx 50\%$  of test minimization gain with the same mutants killed as with the whole test suite.

Our approach is based on test selection at the method level, but it can be applied at different levels of granularity, such as statement, class, or even module level. A sim-

ple solution would be to extend the mapping to store additional information besides method calls for each test. However, coarse-grained levels may lead to select unnecessary tests, whereas finer-grained levels may incur more processing overhead [6, 27]. We shall investigate these levels to find the best trade-off between precision and efficiency in the context of CI/CD pipelines. Another interesting area of research is test prioritization to make delivery pipelines more cost-effective. We are looking into techniques to prioritize the selected tests that are more likely to fail based on historical test executions in the pipeline.

## References

1. Andrews JH, Briand LC, Labiche Y (2005) Is mutation an appropriate tool for testing experiments? In: Proceedings of the 27th international conference on software engineering. ICSE '05. Association for Computing Machinery, pp 402–411
2. Azizi M, Do H (2018) Retest: a cost effective test case selection technique for modern software development. In: 29th IEEE international symposium on software reliability engineering, ISSRE 2018. IEEE Computer Society, pp 144–154
3. Cazzola W, Ghosh S, Al-Refai M, Maurina G (2022) Bridging the model-to-code abstraction gap with fuzzy logic in model-based regression test selection. *Softw Syst Model* 21(1):207–224
4. Celik A, Vasic M, Milicevic A, Gligoric M (2017) Regression test selection across JVM boundaries. In: Proceedings of the 2017 11th joint meeting on foundations of software engineering. ESEC/FSE 2017. ACM, pp 809–820
5. Coles H, Laurent T, Henard C, Papadakis M, Ventresque A (2016) Pit: a practical mutation testing tool for Java (demo). In: Proceedings of the 25th international symposium on software testing and analysis. ISSTA 2016. ACM, pp 449–452
6. Gligoric M, Eloussi L, Marinov D (2015) Practical regression test selection with dynamic file dependencies. In: Proceedings of the 2015 international symposium on software testing and analysis. ISSTA 2015. ACM, pp 211–222
7. Gousset M (2011) Test impact analysis in visual studio 2010. *Visual Studio Magazine*
8. Guo S, Kusano M, Wang C (2016) Conc-ise: incremental symbolic execution of concurrent software. In: Proceedings of the 31st IEEE/ACM international conference on automated software engineering. ASE 2016. ACM, pp 531–542
9. Hammant P (2017) The rise of test impact analysis. <https://tinyurl.com/y42xpf2b>
10. Heger C, Heinrich R (2014) Deriving work plans for solving performance and scalability problems. In: Computer performance engineering. LNCS, vol 8721. Springer, Cham, pp 104–118
11. Humble J, Farley D (2010) Continuous delivery: reliable software releases through build, test, and deployment automation. Addison-Wesley Professional
12. Indrasiri K, Siriwardena P (2018) Microservices for the enterprise: designing, developing, and deploying. Apress, Berkeley, CA
13. Law J, Rothmel G (2003) Whole program path-based dynamic impact analysis. In: ICSE '03. IEEE Computer Society, pp 308–318
14. Lehnert S (2011) A taxonomy for software change impact analysis. In: Proceedings of the 12th international workshop on principles of software evolution and the 7th annual ERCIM workshop on software evolution. ACM, pp 41–50
15. Lindholm T, Yellin F, Bracha G, Buckley A (2014) The Java virtual machine specification, Java SE 8 edition, 1st edn. Addison-Wesley Professional
16. Memon A, Gao Z, Nguyen B, Dhanda S, Nickell E, Siemborski R, Micco J (2017) Taming google-scale continuous testing. In: Proceedings of the 39th international conference on soft-

- ware engineering: software engineering in practice track, ICSE-SEIP 2017. IEEE Computer Society, pp 233–242
17. Orso A, Apiwattanapong T, Harrold MJ (2003) Leveraging field data for impact analysis and regression testing. *SIGSOFT Softw Eng Notes* 28(5)
  18. Orso A, Apiwattanapong T, Law J, Rothermel G, Harrold MJ (2004) An empirical comparison of dynamic impact analysis algorithms. In: *Proceedings of the 26th international conference on software engineering. ICSE '04*. IEEE Computer Society, pp 491–500
  19. Peng Z, Chen T, Yang J (2022) Revisiting test impact analysis in continuous testing from the perspective of code dependencies. *IEEE Trans Softw Eng* 48(06):1979–1993
  20. Ren X, Shah F, Tip F, Ryder BG, Chesley O (2004) Chianti: a tool for change impact analysis of java programs. *SIGPLAN Not* 39(10):432–448
  21. Shahbaz M (2020) Integration TIA. <https://tinyurl.com/36e9tphj>
  22. Shahbaz M (2020) JVM Sniffer. <https://tinyurl.com/5baurdyj>
  23. Sun X, Li B, Tao C, Wen W, Zhang S (2010) Change impact analysis based on a taxonomy of change types. In: *Proceedings of the 34th annual IEEE international computer software and applications conference, COMPSAC 2010*. IEEE Computer Society, pp 373–382
  24. Vallée-Rai R, Co P, Gagnon E, Hendren L, Lam P, Sundaresan V (2010) Soot: a Java bytecode optimization framework. In: *CASCON first decade high impact papers. CASCON '10*. IBM Corp., pp 214–224
  25. Yang G, Person S, Rungta N, Khurshid S (2014) Directed incremental symbolic execution. *SIGPLAN Not* 24(1)
  26. Yoo S, Harman M (2012) Regression testing minimization, selection and prioritization: a survey. *Softw Test Verif Reliab* 22:67–120
  27. Zhang L (2018) Hybrid regression test selection. In: *Proceedings of the 40th international conference on software engineering, ICSE 2018*. ACM, pp 199–209

# ANN-Based Modeling and Control of a Pick and Place Manipulator



Mohamed Essam Mostafa, Aya Essam Mostafa, Hossam Hassan Ammar,  
and Raafat Shalaby

**Abstract** Pick-and-place robots are used in the manufacturing industry. This paper will focus on one of the most used functionalities in the industry. In addition, it will focus on this functionality and how to achieve more precision with the least cost. The robot is used for packaging, picking, placing factory jobs, and inspection. This paper is focusing on the design of the linear pick-and-place robot. The target of this paper is to tackle the design, construction, control, and analysis of the linear pick-and-place robot. The main target of this paper is to find the best design and the most cost-effective way to develop it. It will be done by comparing several delta robots designs to find which one will be the most economical. It will be done via a survey.

**Keywords** System identification · Delta robots · Artificial neural network · Training workplace · Computer vision

---

M. E. Mostafa (✉) · H. H. Ammar · R. Shalaby  
School of Engineering and Applied Science, NU, Giza, Egypt  
e-mail: [M.Essam@nu.edu.eg](mailto:M.Essam@nu.edu.eg)

H. H. Ammar  
e-mail: [h.ammar@ieee.org](mailto:h.ammar@ieee.org)

R. Shalaby  
e-mail: [rshalaby@nu.edu.eg](mailto:rshalaby@nu.edu.eg)

A. E. Mostafa  
School of Information Technology and Computer Science, NU, Giza, Egypt  
e-mail: [Ay.Essam@nu.edu.eg](mailto:Ay.Essam@nu.edu.eg)

R. Shalaby  
SESC Center, School of EAS, NU, Giza, Egypt  
Faculty of Electronic Engineering, Menofia University, Menouf, Egypt



## 1 Introduction

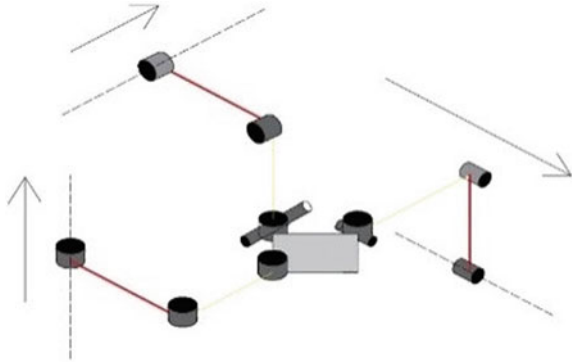
The delta robot is one of the famous robots for pick and place. It is one of the parallel robot families. In this paper, we are concerned with one of the most used functionalities in the industry. Delta robots usually do the pick-and-place tasks [1]. This paper will focus on this functionality and how to achieve more precision with the least cost. It is used for packaging, picking, factory jobs, and inspection. The efforts in this paper will focus on the design of the linear pick-and-place robot. The target of this paper is to tackle the design, construction, control, and analysis of the linear pick-and-place robot. The main target is to find the best design and the most economical way to develop it [2]. It will be done by comparing several designs of the delta robot to find which one will be the most cost-effective [3]. It will be done via survey. The expected outcome is a pick-and-place robot with improved precision, in addition to the design being new and straightforward. In this context, the work method will be based on designing several designs with simplicity in mind to achieve the required simplicity after this phase comes to the analysis phase [4]. The main aim of this phase is to produce the most suitable material that can withstand the operational conditions while being safe enough and, most importantly, affordable to do so. Several analyzes will be done on varied materials, and based on the results, the material will be picked. The third and final phase before manufacturing is the control of the robot. This phase is crucial as it must be done precisely to ensure perfect robot control. This robot is controlled using computer vision [5]. The main advantage of the developed robot is its economic impact after reaching a more straightforward design and choosing the appropriate material. We would have cut the manufacturing and running costs, and the expected outcome is that many factories will be able to add an extra number of robots to their production lines [6].

## 2 Methodology

The Tripteron is a parallelism mechanism constructed of three legs, each of which consists of a Four-DOF serialized mechanism with connections connecting from base to platform. The prismatic joint is anchored at the bottom by three hinge joints whose axis are aligned but not orthogonal to the prismatic joint's axis [7]. Their three legs' final hinge joints are attached to the movable forum such that their axis are orthogonal. The mechanism is the result of a synthesis aiming at uncovering the kin of paralleled translational mechanisms. Despite there are other Tripteron types, the orthogonal Tripteron is the most important [8] (Fig. 1).

The axis of the prismatic joints is orthogonal to one another, and an axis of the passive revolute joints on a particular limb is parallel toward a course of the leg's prismatic joint. The platform's mobility is controlled by an actuator through one of the Cartesian planes. The manipulator is singularity-free and disconnected: its Jacobian matrix equals the identity matrix, and its workspace becomes a parallelepiped for

**Fig. 1** Schematic representation for tripteron



the appropriate geometric configuration of the legs [9]. Its kinematics is like that of a sequent seeming Cartesian manipulator, and that is made up of three orthogonal prismatic sliders connected in a chain. Tripteron inverse kinematic issue may be expressed as

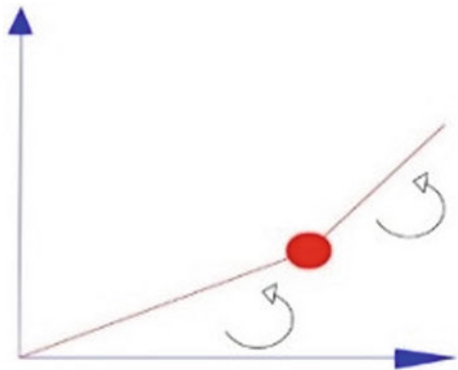
$$\mathbf{P}_1 = \mathbf{x}, \mathbf{P}_2 = \mathbf{y}, \mathbf{P}_3 = \mathbf{z} \quad (1)$$

where and represent the actuated joint coordinates, and  $x$ ,  $y$ , and  $z$  represent the moving platforms' Cartesian coordinates.

## 2.1 Dynamic Modeling

The Newton–Euler formula underpins the dynamic models presented in this study. All moving bodies' free-body diagrams are installed, and the Newton and Euler equations constructed for movable every movable object. Nevertheless, rather of simply assembling all the equations into a large matrix system, it will be demonstrated that, with a careful solution approach, the procedure may be limited to solving a small number of often separated equations (Fig. 2).

The Tripteron manipulator's construction is shown in Fig. 1. The three prismatic, revolute, revolute, revolute, and revolute (3-PRRR) limbs link a moveable platform to a stationary base is shown in the figure. Point E is the center of the end-effector concerning each edge [10, 11]. The link lengths are represented by the initial point of every prismatic joint, which is specified by dpi, as well as the sliding displacement is determined by three local coordinate systems, the origins of which are located at Any limb constrains the end-effector from a rotation movement about the local and global axes due to the three parallel revolute joints located at points. The combined impacts result in three-redundant restraints on the rotational of the moving platform, which eventually constrains the rotation of the moving end-effector. The moving platform now has three translational degrees of freedom. A rotary actuator powers the first

**Fig. 2** Leg of the tripteron

revolute joint within each limb in the rotary actuation method. All other joints are passive, whereas the actuated joints provide motion [9].

## 2.2 Inverse Kinematics of the End Effector

The inverse kinematics is End Effector to calculate the joint angles  $\theta_{i1}$ ,  $\theta_{i2}$  and  $\theta_{i3}$  given the end-effector position  $E = [E_x E_y E_z]^T$  as input. From Eq. (1), the position vector of Ci with reference to each local coordinate system as follows

$${}^1C_1 = \begin{bmatrix} E_y - L_E \\ E_z \\ \mathbf{0} \end{bmatrix}, {}^2C_2 = \begin{bmatrix} E_z \\ E_x - L_E \\ \mathbf{0} \end{bmatrix}, {}^3C_3 = \begin{bmatrix} E_x - L_E \\ E_y \\ \mathbf{0} \end{bmatrix} \quad (2)$$

From Eq. (2), the input joint angle  $\Theta_{i1}$  for each limb found as follows:

$$\Theta_{i1} = \tan^{-1}\left(\frac{C_{iy}}{C_{ix}}\right) + \cos^{-1}\left(\frac{L_{i1}^2 + C_{ix}^2 + C_{iy}^2 - L_{i2}^2}{2L_{i1}\sqrt{C_{ix}^2 + C_{iy}^2}}\right) \quad (3)$$

where  $i = 1, 2$  and  $3$  which is the limb number. The following equation is generated because of the finite reach of each limb, and it consists of the minimum and maximum limitations of a serial chain and thus must be satisfied:

$$(L_{i1} - L_{i2})^2 \leq C_{ix}^2 + C_{iy}^2 \leq (L_{i1} + L_{i2})^2 \quad (4)$$

### Force analysis for Jacobian

The velocities of Ci can be obtained by differentiating Eq. (2) for time yields relative to local coordinate systems as follows:

$${}^1\dot{C}_1 = \begin{bmatrix} \dot{E}_y \\ \dot{E}_z \\ \mathbf{0} \end{bmatrix}, {}^2\dot{C}_2 = \begin{bmatrix} \dot{E}_z \\ \dot{E}_x \\ \mathbf{0} \end{bmatrix} \text{ and } {}^3\dot{C}_3 = \begin{bmatrix} \dot{E}_x \\ \dot{E}_y \\ \mathbf{0} \end{bmatrix} \quad (5)$$

The linear velocity of  $C_i$  relative to the local coordinate system can be written in terms of the joint speeds as follows:

$$\begin{aligned} {}^i\dot{C}_{ix} &= -\dot{\theta}_{i1} L_{i1} \sin(\Theta_{i1}) - \dot{\theta}_{i2} L_{i2} \sin(\Theta_{i2}) \\ {}^i\dot{C}_{iy} &= \dot{\theta}_{i1} L_{i1} \cos(\Theta_{i1}) + \dot{\theta}_{i2} L_{i2} \cos(\Theta_{i2}) \end{aligned} \quad (6)$$

Eliminating  $\dot{\Theta}_{i2}$  in Eq. 6 yields:

$${}^i\dot{C}_{ix} \cos(\Theta_{i2}) + {}^i\dot{C}_{iy} \sin(\Theta_{i2}) = L_{i1} \sin(\Theta_{i2} - \Theta_{i1}) \dot{\theta}_{i1} \quad (7)$$

Obtaining an input–output velocity relationship by substituting Eq. 5 into 7 as follows:

$$\mathbf{J}_x \dot{\mathbf{P}} = \mathbf{J}_{q1} \dot{\theta}_1 \quad (8)$$

where

$$\dot{\mathbf{E}} = [\dot{E}_x, \dot{E}_y, \dot{E}_z]^T, \quad \dot{\theta}_1 = [\dot{\theta}_{11}, \dot{\theta}_{22}, \dot{\theta}_{33}]^T \quad (9)$$

$$\mathbf{J}_x = \begin{bmatrix} \mathbf{0} & \cos(\theta_{12}) & \sin(\theta_{12}) \\ \sin(\theta_{22}) & \mathbf{0} & \cos(\theta_{22}) \\ \cos(\theta_{32}) & \sin(\theta_{32}) & \mathbf{0} \end{bmatrix} \quad (10)$$

$$\mathbf{J}_q = \begin{bmatrix} L_{11} \sin(\theta_{12} - \theta_{11}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & L_{21} \sin(\theta_{22} - \theta_{21}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & L_{31} \sin(\theta_{32} - \theta_{31}) \end{bmatrix} \quad (11)$$

If  $\sin(\theta_{i2} - \theta_{i1}) \neq 0$ , then the inverse velocity transformation can be written as:

$$\dot{\theta}_1 = \mathbf{J} \quad (12)$$

where the Jacobian matrix  $(J) = J_q^{-1} J_x$ . If  $\det(J_q) = 0$  because we can't determine the joint velocities, so the inverse kinematic singularity occurs when the first and second links of any limb are collinear which is achieved by

$$\sin(\theta_{i2} - \theta_{i1}) = 0 \quad (13)$$

the direct kinematics singularity occurs when  $\det(J_x) = 0$  or according to the following equation:

$$\text{Cos}(\Theta_{12}) \text{Cos}(\Theta_{22}) \text{Cos}(\Theta_{32}) + \text{Sin}(\Theta_{12}) \text{Sin}(\Theta_{22}) \text{Sin}(\Theta_{32}) = 0 \quad (14)$$

Equation 14 can be concluded that their many direct kinematic singularities may exist.

## 2.3 Model-Based Neural Network

Neural networks, also named as artificial neural networks or simulation neural networks, are the foundation of deep learning techniques, and these have been named and built following the human brain to mimic how physical neurons communicate with one another. Node layers in artificial neural networks include three levels first is input, then hidden and contains one layer or more, and the last one is output [12, 13].

Each node seems to have a connected node, a limit, and a weight associated with it. When a node's output exceeds a certain limit, it is in active mood and starts sending data to the network's next level. Or else, no data is transmitted to a subsequent level of the network. To adapt and improve over time, neural networks require training data. Nevertheless, when those learning algorithms have been fine-tuned improved accuracy, they may be used to create effective tools, which enables us to recognize and combine inputs much more quickly than if we did it manual. For example, speech recognition and image recognition tasks can be accomplished in moments [14, 15].

Types of neural network

There are sundry types of neural networks, everyone serves a certain function. Most frequent types of neural networks are feedforward neural network, convolutional neural networks, and recurrent neural networks.

Rosenblatt invented the first form of neural network, perceptron's, in 1958. These represent the most basic type of neural network, consisting of a single neuron.

**Feedforward neural network** (FNNs) sometimes referred as a multi-layer perceptron (MLPs), is an example of an artificial neural network. These are made up of three layers same as artificial neural networks.

**Convolutional neural networks** (CNNs) identify patterns in images by employing linear algebra methods such as matrix multiplication.

**Recurrent neural networks** (RNNs) use feedback loops to adapt from time-series data, making them useful in anticipating upcoming events.

This robot's successful operation involves detecting an object's position via computer vision. Then the movement of the three motors should be calculated accordingly and sent to the microcontroller. To achieve this, there are two ways: inverse kinematics or a neural network. In this case, a 6–12–3 multilayer feedforward neural network was used. It consisted of three neurons input layers, each for the position along the three axes ( $x$ ,  $y$ ,  $z$ ), and an error variable for each axis. The second layer is the hidden layer which consists of twelve neurons with a sigmoid activation function.

The third layer is the output layer consisting of three neurons: each neuron outputs the required rotation for each Motor [15–17].

### 2.3.1 Training the ANN

To train the ANN, a dataset was collected. This dataset records each one as the complete operation started moving from home to picking up and finally to return to the home concerning the target object location (Tables 1 and 2).

**Table 1** The data used to train the ANN

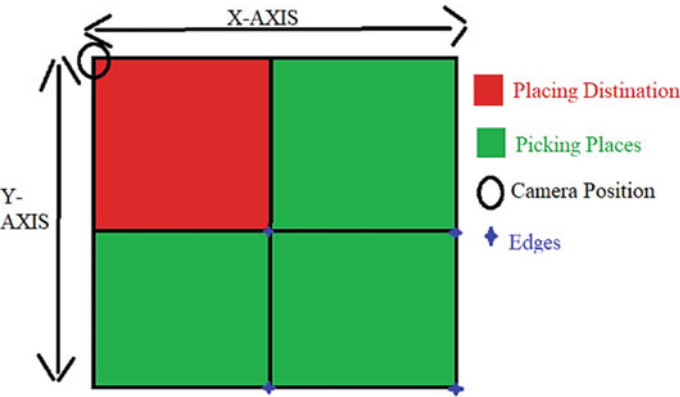
LX	LY	LZ	X (mm)	Y (mm)	Z (cm)
72	72	− 69	294	251	20
30	44	− 70	222	259	20
26	30	− 70	209	245	20
44	0	− 69	216	195	20
29	0	− 70	197	204	20
14	4	− 70	178	221	20
28	4	− 70	196	212	20
42	4	− 70	215	204	20
51	35	− 70	244	238	20
61	35	− 70	256	232	20
− 14	35	− 70	256	284	20
72	− 63	− 70	217	103	20
72	− 2	− 70	249	179	20
44	53	− 70	244	268	20
28	12	− 70	202	226	20
60	12	− 70	242	206	20
60	45	− 69	261	248	20
23	45	− 69	212	273	20
22	12	− 70	193	231	20
− 4	12	− 70	162	247	20
− 4	42	− 70	175	289	20
− 17	29	− 70	152	282	20
7	30	− 68	182	267	20
− 12	6	70	156	245	20
9	26	− 70	182	265	20
0	17	− 70	165	259	20
− 8	36	− 70	165	290	20

**Table 2** The frequencies of different modes

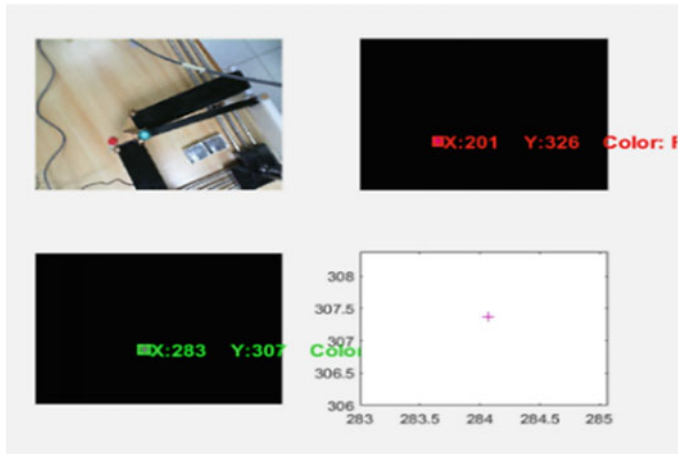
Mode	Frequency [HZ]
1	5.1312
2	5.7219
3	10.88
4	18.598
5	23.286

**2.4 Control Methodology**

We decided to divide the working space into four similar regions in the shape of squares to give a position axis ( $X$ – $Y$ ) for the boundary of each. According to the axis for each square edge, the MATLAB code is appended to detect whether the object we need to pick is inside one of those squares in the figure. If the object is detected to be out of all squares, the arm will not move as safely and avoid arm breaking. However, if the object is inside one of the picking places (green squares), the code will automatically calculate in which region the object is. Consistently, the code will recognize in which square the object is and automatically send it to the Arduino-specific  $LX$  and  $LY$ . Which is the distance the motor should move the cart on each axis ( $X$ ,  $Y$ ), therefor the End-Effector should lfeffector to pick the object and then place it in the placing destination (red square) by giving specific  $LX$  and  $LY$  as well, finally return to its home position waiting to detect another object and restart the whole process again (Fig. 3).



**Fig. 3** Represents the working space of the pick and place robot



**Fig. 4** Concept of operation

## 2.5 Computer Vision System

Computer vision is how a computer gains a high-level understanding of a digital image or a video. It is a field of artificial intelligence that teaches the computer how to perceive, interpret, and understand information from digital images and videos to communicate with the outside world. Computer vision is implemented to give feedback, which is successful control.

### 2.5.1 Concept of Operation

Detection and feedback are constructed by fixing a camera on the robot at a place where the working space is entirely inside the region of view. The camera is programmed using MATLAB software to detect three assorted colors (RGB) and a frame for each to give a specific axis and location; so, The End Effector (Green) position concerning the object we need to pick and place (Red) given as shown in Fig. 4.

### 2.5.2 Vision System Construction

We fix the Wood stand (1) at the top of the motor housing (3) of the Z-axis. Afterward, we embed the camera (2) to the wood stand while directing the lens perpendicular to the working space. Additionally, we install the external camera into MATLAB after connecting the USB to the laptop. Finally, we run the code that will be discussed later and confirm that the end-effector (4) and the object (5) are detected inside a boundary box with its axis ( $X$ – $Y$ ) location, as shown in Fig. 4.



### 2.5.3 Camera Calibration

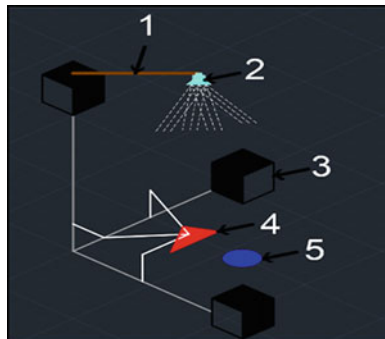
The relationship between the region of view and the space between the camera and the object. Thus, the region of view which is between the two dash lines increases with the space between the camera and the object; the closer the object to the camera, the bigger it appears; correspondingly, a big boundary box takes place and vice versa; as well as that act on the detected location ( $X$ – $Y$ ). In the MATLAB detection code, the axis we get is in pixels. Thus, we need to calibrate it from pixels to centimeters. The region of view of the robot is fixed because the camera is fixed at the top of the Z-axis; hence, they can get the constant calibration factor as shown in the equation below.

The calibration constant in the  $x$  direction is equal to the calibration constant in the  $y$  direction. As pixel is square with equal sides.

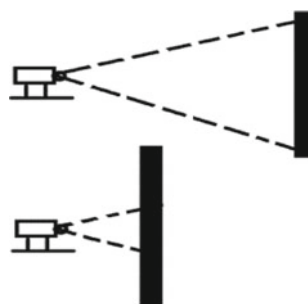
$$K = \frac{X}{Y} \quad (15)$$

$K$  is the constant calibration factor,  $X$  is the  $x$  distance (CM) of the region of view that manually measured using a meter, and  $Y$  is the number of pixels in the  $x$ -axis of the camera (Resolution) (Figs. 5 and 6).

**Fig. 5** Vision system construction



**Fig. 6** Camera calibration





**Fig. 7** Camera calibration process

## 2.6 Camera Calibration Process

See Fig. 7.

## 3 Results

**Analysis:** This section is about the analysis of the configurable parallel robot. All analyses were done on ANSYS 19.2. The analysis is divided into three main sections: Rigid dynamics analysis, static structural analysis, and modal analysis.

### 3.1 First Rigid Dynamics Analysis

This section focuses on the robot's velocity and Acceleration and shows the blocks' translation and rotation of the joints. The Total deformation analysis shows the deformation of the robot: Aluminum, chromium, and polyamide (Fig. 8).

### 3.2 Second Static Structural Analysis

This section focuses on the stresses on the robot and the total deformation for the robot. Second, we defined the fixed support of the robot, then add force with 50 N at the end effector to calculate the weight can carry, and added three joints with a displacement of 20 to show the end effector can carry the weight when it's moved (Fig. 9).

### 3.3 Third Modal Analysis

This section focuses on the effect of vibration on the robot and shows the total deformation at different frequencies (Fig. 10).

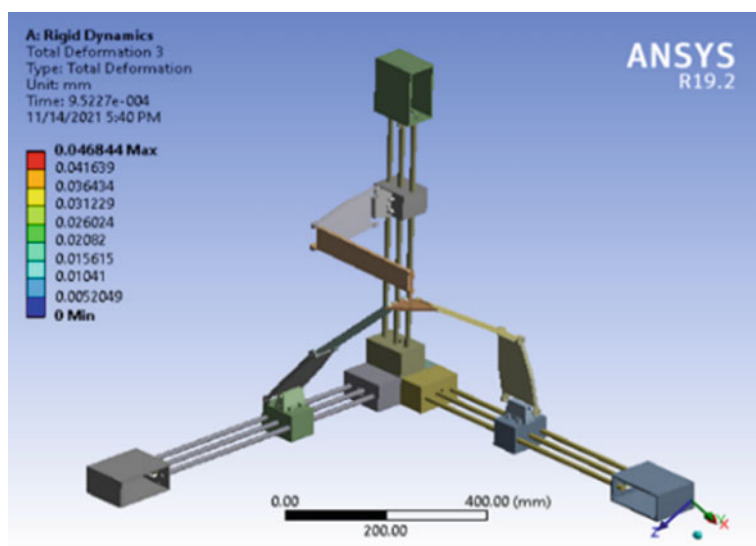


Fig. 8 First position

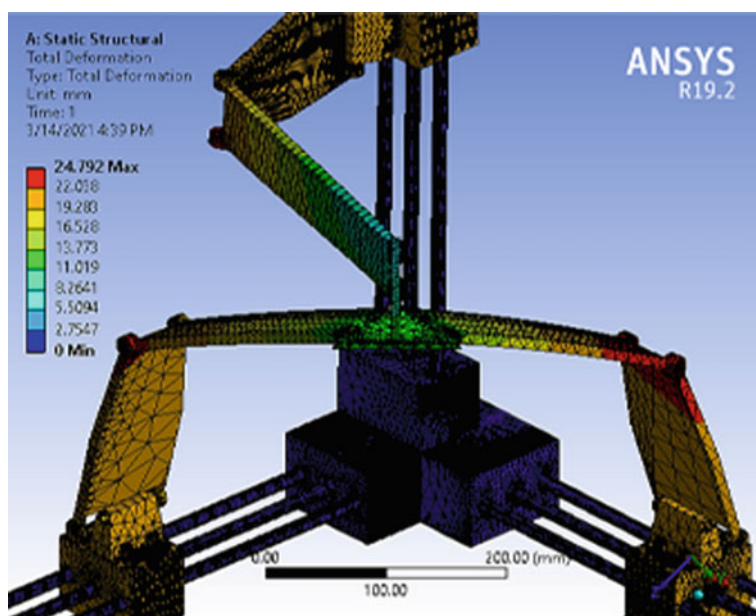
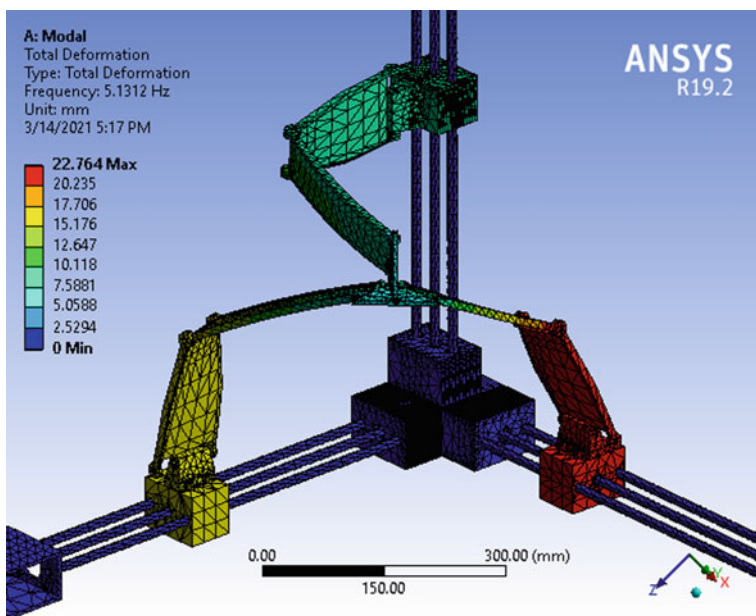


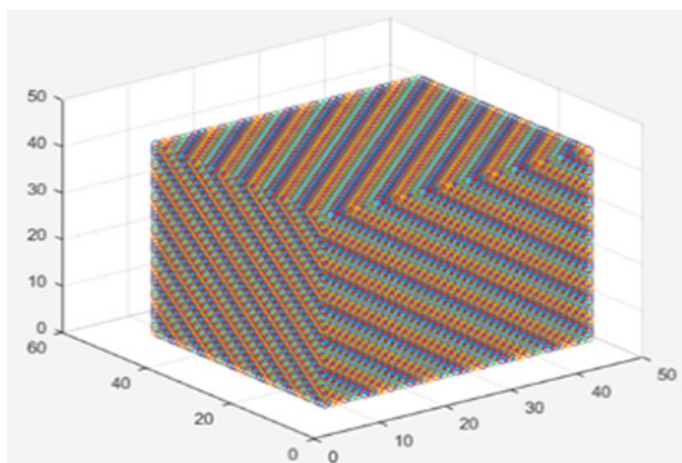
Fig. 9 Total deformation



**Fig. 10** Total deformation at frequency 5.1312 Hz

### Workspace for the End Effector.

See Figs. 11, 12, 13, 14 and 15.



**Fig. 11** Workspace of the robot

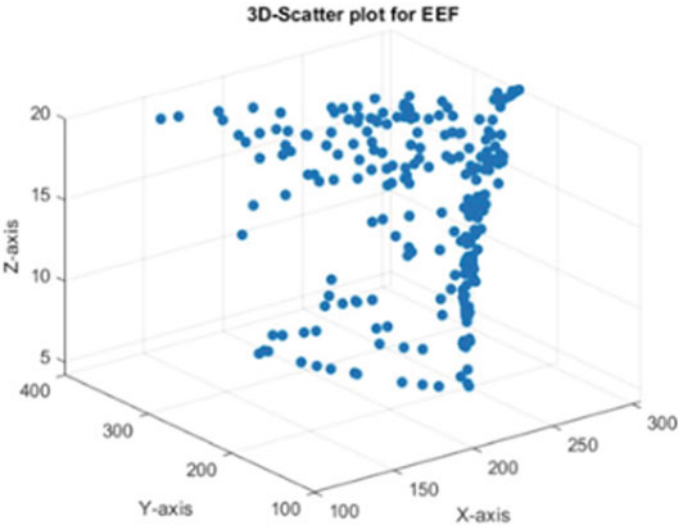


Fig. 12 3D-scatter plot for end effector

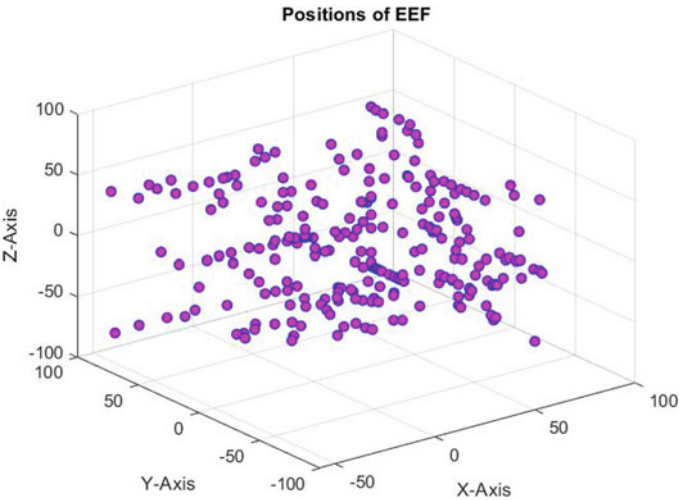
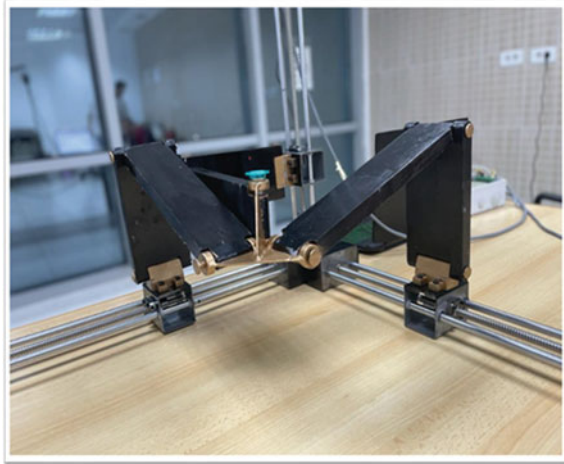


Fig. 13 3D-scatter plot for position of EEF

**Fig. 14** The end effector after manufactured



**Fig. 15** Linear pick and place robot



## 4 Conclusion

Pick-and-place robots are used extensively in industrial applications such as inspection, assembly, and packaging. These robots are important as they have improved accuracy and the best processing speed, which saves time consistency, eliminates, and reduces injuries, and most importantly, they are cost-effective. Although they are not a new trend, they have always been under research for enhancements and still need many optimizations. The target of this paper is to introduce the design and implementation of an accurate pick-and-place robot that is sufficiently fast for the job to be done efficiently. Several stress analyses were done, and based on the results, varied materials were assigned for each part depending on the working environment. The robot will be composed of three materials Aluminum, Artelon, Sheet Metal, and Chrome. To control the robot in the next phase, we will use computer vision for object

detection. For sensory Data of the End Effector, the end effector will have an IMU Sensor attached to it to measure acceleration and Gyroscopic Data for controlling purposes.

## References

1. Yang C, Ye W, Li Q (2022) Review of the performance optimization of parallel manipulators. *Mech Mach Theory* 170:104725
2. Meng Q, Li J, Shen H, Deng J, Wu G (2021) Kinetostatic design and development of a non-fully symmetric parallel Delta robot with one structural simplified kinematic linkage. *Mech Based Des Struct Mach* 51(7):3717–3737. <https://doi.org/10.1080/15397734.2021.1937213>
3. Lin J, Luo CH, Lin KH (2015) Design and implementation of a new DELTA parallel robot in robotics competitions. *Int J Adv Robot Syst* 12:1–10
4. Alvares AJ, Gasca EAR, Jaimes CIR (2018) Development of the linear delta robot for additive manufacturing. In: *Proceedings of the 2018 5th international conference on control, decision and information technologies (CoDIT)*, Thessaloniki, Greece, 10–13 April 2018, pp 187–192
5. Wang G, Xu Q (2017) Design and precision position/force control of a piezo-driven microinjection system. *IEEE/ASME Trans Mechatron* 22:1744–1754
6. Gao X, Liu Y, Zhang S, Deng J, Liu J (2022) Development of a novel flexure based XY platform using single bending hybrid piezoelectric actuator. *IEEE/ASME Trans Mechatron* 27:1–11
7. Hultdt T, Stenius I (2019) State-of-practice survey of model-based systems engineering. *Syst Eng* 22:134–145.11
8. Howell LL, Magleby SP, Olsen BM (2013) *Handbook of compliant mechanisms*. John Wiley & Sons Incorporated, Hoboken, NJ, USA, p 311
9. Clark L, Shirinzadeh B, Bhagat U, Smith J, Zhong Y (2015) Development and control of a two DOF linear–angular precision positioning stage. *Mechatronics* 32:34–43.14
10. Jin L, Li S, Yu J, He J (2018) Robot manipulator control using neural networks: a survey. *Neurocomputing* 285:23–34
11. Le TD, Kang HJ (2014) An adaptive tracking controller for parallel robotic manipulators based on fully tuned radial basic function networks. *Neurocomputing* 137:12–23
12. Mitrovic A, Nagel WS, Leang KK, Clayton GM (2020) Closed-loop range-based control of dual-stage nanopositioning systems. *IEEE/ASME Trans Mechatron* 26:1412–1421
13. Roshni N, Kumar TKS (2017) Pick-and-place robot using the centre of gravity value of the moving object. In: *2017 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, Srivilliputtur, India, pp 1–5
14. Toquica JS, Oliveira PS, Souza WS, Motta JMS, Borges DL (2021) An analytical and a deep learning model for solving the inverse kinematic problem of an industrial parallel robot. *Comput Ind Eng* 151:106682
15. Li S, Shao Z, Guan Y (2019) A dynamic neural network approach for efficient control of manipulators. *IEEE Trans Syst Man Cybern Syst* 49:932–941
16. Jiang Y, Yang C, Na J, Li G, Li Y, Zhong J (2017) A brief review of neural networks based learning and control and their applications for robots. *Complexity* 2017:1895897
17. Carlos LF et al (2018) A soft computing approach for inverse kinematics of robot manipulators. *Eng Appl Artif Intell* 74:104–120. <https://doi.org/10.1016/j.engappai.2018.06.001>

# Toward Learning Analytics in a Distributed Learning Environment



Dijana Oreski , Vjeran Strahonja, and Darko Androcec

**Abstract** The learning and teaching process are widely supported by various virtual learning environments (VLE), often combined in a distributed learning environment. The distributed approach offers learning at scale, making use of a range of available technologies. The nature of distributed learning environment offers various potentials for learning analytics as well as various challenges since data on student activities and achievements are distributed. In this paper, we examine how to empower student activity and achievement by using data collected in the LMS with data collected in other parts of the distributed learning environment with the aim of improvement for learning analytics and learning design. This study takes the first step towards an integrated approach by investigating a case of one course, Information systems development taught at the University of Zagreb. The course was taught by using Moodle learning management system combined with YouTube videos and Oracle Apex environment. Research is focused on the patterns in students' behavior on Moodle for the specific groups of resources and activities related to the Oracle Apex, some of which are created on YouTube. Research results indicate great potential in integrating multi-sources of e-learning data. One of the main issues remains data integration with Oracle Apex and the application of integrated data for the development of predictive models, which will be investigated in future work.

**Keywords** Learning analytics · Distributed learning environment · Oracle Apex · LMS data · YouTube analytics

---

D. Oreski (✉) · V. Strahonja · D. Androcec

Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, Varazdin, Croatia

e-mail: [dijana.oreski@foi.unizg.hr](mailto:dijana.oreski@foi.unizg.hr)

V. Strahonja

e-mail: [vjeran.strahonja@foi.unizg.hr](mailto:vjeran.strahonja@foi.unizg.hr)

D. Androcec

e-mail: [darko.androcec@foi.unizg.hr](mailto:darko.androcec@foi.unizg.hr)



# 1 Introduction

In recent years, especially during the COVID-19 pandemic, universities are increasingly using virtual learning environments in teaching. There are numerous environments available, and often they are combined in the distributed learning environment. The distributed learning environment consists of several systems, for example, LMS (e.g., Moodle, Blackboard), cloud platform for application prototype development (e.g., Oracle APEX), modeling tools (e.g., SQL Database Modeler), cloud/Internet hosted distributed version control system (e.g., GitHub), online learning resources (e.g., Oracle Academy, YouTube...), etc. The combination of environments enables more informed choices regarding resource allocation, student support, and assessment. At the same time, such an approach also raises several concerns. The extraction, integration, analysis, and use of student data in the distributed learning environment are complex from various perspectives.

The usage of LMSs is acquiring importance in education because they provide flexible integration platforms for organizing a huge amount of learning resources of different types. Moodle is the most used platform in higher education, and the easiest to use [1, 4]. Moodle enables different types of resources and activities to be uploaded and created by teachers in these systems and to be used by students. Activity refers to a feature in Moodle that imposes students' interactions with teachers or other students. Typical activities are assignments, forums, quizzes, and lessons. On the other hand, the resource is a type of feature presented to the student by a teacher. Typical resources are files, pages, links, or videos. Each resource and activity refer to certain content of the course. Videos can be created on YouTube. Youtube was ranked 2nd technological tool used for educational purposes in the survey performed by Trabelsi et al. [18]. Students perceive YouTube as a meaningful part of their university learning since helps them understand the concepts. YouTube videos linked at the LMS and analyzed in this paper are related to the Oracle Apex. Oracle Apex is a web-based development tool for building web-based database applications [12]. Oracle Application Express (Apex) is an application development platform that enables user's development of web applications based on data within the Oracle database [2].

Despite a broad acceptance of the various benefits of learning analytics within an LMS environment, as far as we know, there are no studies performing learning analytics in a distributed environment. This study seeks to do the first step toward developing a learning analytics framework in a distributed learning environment. Such an approach can provide a deeper insight into the types of students' activities, and teachers can make informed decisions about learning design and student support.

In this paper, we analyze the course Information systems development, taught at the University of Zagreb, Faculty of Organization and Informatics. The course was taught by using three systems: (i) LMS Moodle, (ii) YouTube, and (iii) Oracle Apex.

The paper is structured as follows. Section 2 overviews relevant research papers on a given topic. Section 3 explains the data and research methodology. Section 4 gives research results with a discussion. Section 5 concludes the paper with guidelines for future research on this matter.

## 2 Literature Review

Learning analytics is described by Gašević et al. [9] as the process of analyzing digital traces (log data) of the student's interactions with their online learning activities to identify patterns of learning behavior [9]. Educational institutions accepted LMSs, and their application led to the accumulation of huge amounts of data which are one of the main data sources for learning analytics. LMSs have grown exponentially recently and have a strong influence on the overall learning process. Many papers so far have analyzed these data to develop models of student achievement. This has led to numerous findings related to the topic. Some of them are listed below.

The study of Black et al. [3] aims to determine whether are logs of student activity within online courses related to student perceptions. Results indicated a significant level of course data logs predictive power and emphasized the need to take into consideration the content of the courses [3]. Deepak [6] examined the importance of the Moodle features implemented in the Moodle at Kajaani University of Applied Sciences from teachers' perspectives. Deepak investigated what features are mostly used by the teachers. The results showed that Moodle is mostly used for delivering course content and creating activities [6]. In the recent research study of Feldman-Maggor et al. [8], students' description based on learning patterns were performed. Factors of students' success were examined in an online environment. A data mining approach was applied to the data from undergraduate online courses. Submission of an optional assignment and the students' video opening numbers were shown to be the most important predictors. This study showed how important for student success are their choices. Conijn et al. [5] reviewed variables extracted from the LMS. Their research concluded that it is hard to draw general conclusions about student performance prediction, and each setting and each domain requires analysis [5]. Oreski and Kadoic performed an analysis of log data from a blended course in the field of information communication technology (ICT). Results indicated no significant correlation between students' behavior and course participation. However, differences between male and female students are identified [13]. The same authors further analyzed log files of the LMS Moodle at the Faculty of Organization and Informatics at the University of Zagreb in terms of students' success. The results of the student's behavior, based on logs in Moodle are interpreted considering student course completion [10]. In their recent work, Kadoić and Oreški [11] expanded research by combining LMS data with YouTube video data striving to identify patterns between video viewing and LMS behavior. Surprisingly few studies have examined the use of YouTube videos for information technology education. There are studies examining YouTube usage for teaching and learning arts [7], music [14], medicine [16], chemistry [15], and tourism [17]. As presented, the literature review indicated numerous papers dealing with analytics of LMS Moodle data or YouTube analytics. However, there is a lack of integration of various approaches.

### 3 Research Design

In a distributed learning environment, data on student activities and achievements are distributed. LMS logs only contain data about activities within the LMS because each part of the distributed learning environment stores data that was created as a result of the user's activities and the actions taken.

The main research question investigated in this paper is:

RQ: How to empower student activity and achievement data collected in the LMS with data collected in other parts of the distributed learning environment with the aim of learning analytics and learning design improvement?

With this motivation, our paper takes a first step toward achieving that goal by proposing a data fusion of two systems hence leading to enriching data analysis. In concrete, these sources are Moodle of the Faculty of Organization and Informatics at the University of Zagreb (Croatia), which is enriched with YouTube video analytics. The proposed approach strives to properly integrate e-learning data.

Data for this study were collected from 187 students in the course Information systems development, which was taught at the University of Zagreb, Faculty of Organization and Informatics in the winter semester of the academic year 2021/2022. Information systems development is the obligatory course in the third year of the undergraduate study program Information and business systems. In the course, students should acquire the necessary theoretical knowledge and practical skills for designing complex information systems, and applying the methodology of system engineering. Students become familiar with approaches, processes, methods, techniques, and development environments that are the basis of modern information system development methodologies. Special attention is put on modeling as a basis for designing. In this context, models of information system requirements, information system architecture, data and processes, and business models are studied. In addition to primary processes and methods of development, secondary ones are also studied, in the field of project management. It should be emphasized that students must apply the acquired theoretical knowledge and skills in a problem situation, creating a given project in a team, with the use of tools for modeling, prototyping, and collaboration. In this context, Oracle Apex is used in the course.

The course was taught as a hybrid consisting of lectures, seminars, and laboratory exercises combined with various activities at LMS Moodle. At the end of the semester, students' logs were downloaded from Moodle. Reporting feature at the LMS, Moodle enables all students to click on the course. Different types of logs can be extracted at the level of each student, as well as overall. Once downloaded, the log data are prepared with Microsoft Excel using the pivot tables option, thus creating data logs for the whole semester. Students were grouped by their names, and each row consisted of data for one student. A summary of the student's activities in the LMS was obtained focusing on the following variables: ERA model for Oracle Apex (provided to students as a link), Oracle Apex 1st video (provided to students as a link), Oracle Apex 2nd video (provided to students as a link), Oracle Apex

2nd assignment, (provided to students as an assignment), Oracle Apex 3rd video (provided to students as a link), Oracle Apex 3rd assignment (provided to students as an assignment), Oracle Apex 4th video, Oracle Apex 4th assignment (provided to students as an assignment). Topics of the videos were as follows:

1. Oracle Apex 1st video—Creation of tables, primary and foreign keys
2. Oracle Apex 2nd video—filling in tables and creating an initial application with simple forms
3. Oracle Apex 3rd video—Editing of simple forms, static and dynamic lists of values
4. Oracle Apex 4th video—master-detail forms, views, classic and interactive reports

Video content was based on course content and was explicitly intended for the students in the course. Videos were created by the teacher at the course and were placed on the YouTube channel. A YouTube channel should be created by teachers wanting to upload videos to YouTube. When logging in with their username and password, the teacher has access to its full YouTube Analytics data. Such data were extracted and used in this research.

## 4 Research Results

Table 1 gives an overview of the overall number of openings for each of the observed resources and activities as well as a unique number of users that opened resources and activities.

The graph in Fig. 1 visualizes overall logs per activity and resource. Logs to activities (second, third and fourth assignment submissions) stand out compared to the logs to the resources.

Logging into the assignments stands out when compared with logging in to the resources such as videos or links.

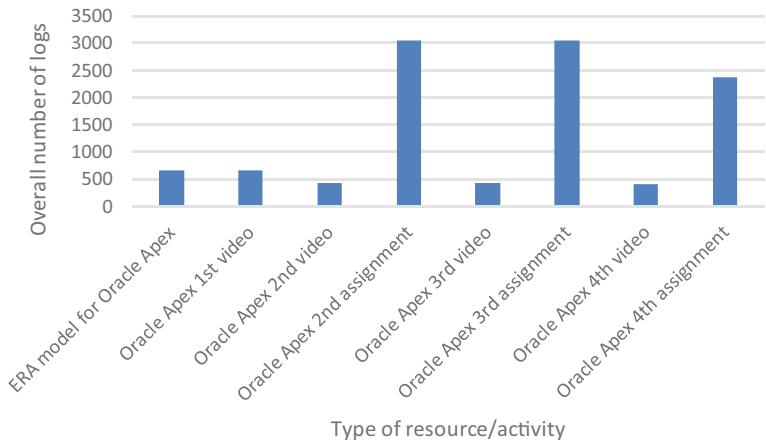
The graph in Fig. 2 shows views of the course materials by unique users. It is notable slightly higher number of views for each activity (assignment) related to the resource (link to the video linked to the assignment) (Fig. 3).

The highest number of logs is achieved on Sunday. Sunday was the day in the week when the assignment submission deadline was set up. If we take a deeper look at logs distribution by days in the week (Fig. 4), we can see that logs to assignments submission were the highest on Sundays, whereas views to the resources were also high on the days of the labs. Most of the groups had labs on Mondays.

In this paper, learning analytics is combined with YouTube analytics to get a deeper view of students' behavior patterns. Table 2 presents analytics of four videos whose links were attached to LMS Moodle. Video duration is provided along with the Average percentage of views. The average percentage of views varies from 0.28 to 0.36.

**Table 1** The overall number of logs and unique users

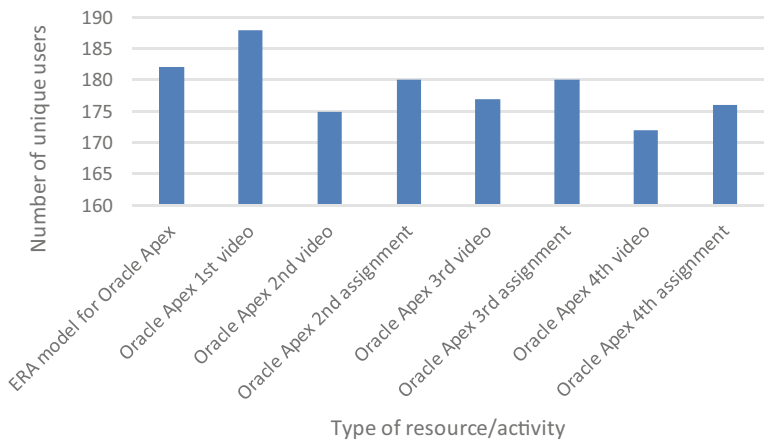
Type of resource/activity	Number of logs	Number of unique users
ERA model for Oracle Apex	670	182
Oracle Apex 1st video	658	188
Oracle Apex 2nd video	437	175
Oracle Apex 2nd assignment	3061	180
Oracle Apex 3rd video	430	177
Oracle Apex 3rd assignment	3044	180
Oracle Apex 4th video	406	172
Oracle Apex 4th assignment	2369	176
ERA model for Oracle Apex	670	182



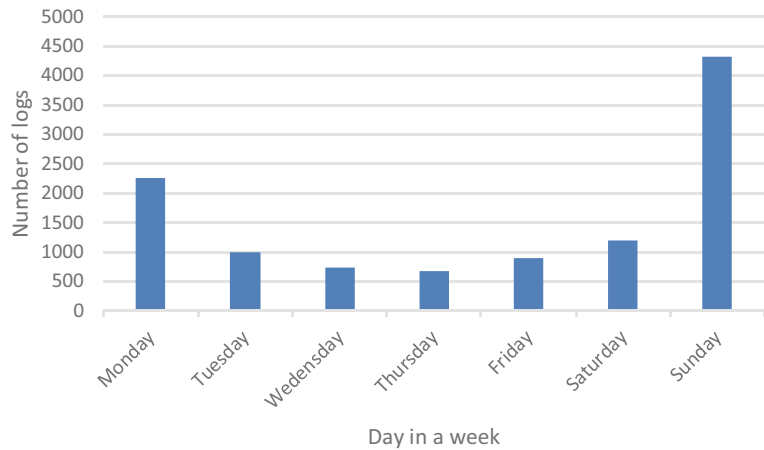
**Fig. 1** Overall logs per activity and resource

Analytics regarding video complexity is provided in Table 3. Complexity is measured twofold: (i) complexity measured subjectively by the teacher, (ii) complexity of the video measured as a sum of the videos’ spikes and dips.

The complexity of the topic is set up as a three-item scale: low, medium, and high were used by lecturers to evaluate the complexity of the topic. The second complexity measure involves spikes and dips. The number of spikes in the video is

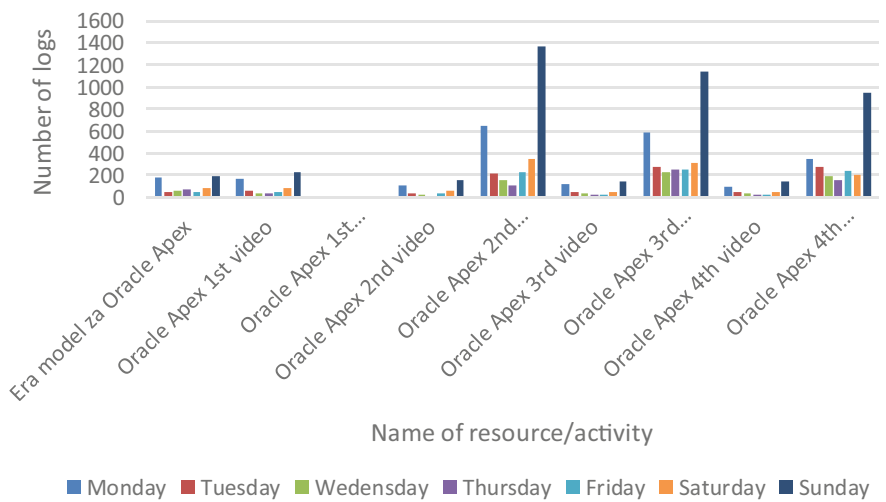


**Fig. 2** Unique users per activity and resource



**Fig. 3** Distribution of logs by day in a week

overtaken by YouTube Analytics: “Spikes appear when more viewers are watching, rewatching or sharing a moment or moments of a video”. The number of dips in the video indicates mean viewers are skipping or leaving your video at that specific part”. The complexity of the content is calculated as the sum of the number of dips and the number of spikes. The measure is adopted by Kadoić and Oreški [11]. Their assumption was that considering the definitions of the dips and spikes, the sum of them would be about the complexity of the content.



**Fig. 4** Logs by day and activity

**Table 2** YouTube analytics: video views

Video	Duration	The average percentage of views
Oracle Apex 1st video	22:30	0.29
Oracle Apex 2nd video	19:53	0.28
Oracle Apex 3rd video	24:40	0.28
Oracle Apex 4th video	19:29	0.36
Oracle Apex 1st video	22:30	0.29

**Table 3** YouTube analytics: video complexity

Video	Complexity	Number of dips	Number of piks	Complexity
Oracle Apex 1st video	Small	3	3	6
Oracle Apex 2nd video	Small	2	5	7
Oracle Apex 3rd video	Medium	2	3	5
Oracle Apex 4th video	Medium	4	5	9

## 5 Conclusion

This paper shows the results of analyzing student course materials created on YouTube and linked to the LMS Moodle regarding Oracle Apex course content. Students opening behavior and assignments submission were investigated. We have

analyzed data from comprehensive to individual perspectives and conducted analysis using different categories of time to detect temporal dependencies. Contributions of this paper relate to identifying patterns in student behavior when accessing the different types of sources at the LMS Moodle in combination with YouTube videos. Furthermore, this paper can be used as the first step towards a data integration strategy for applications in distributed learning environments where current LMSs and other academic data sources are used to support teachers and students with learning analytics.

There are various challenges to be addressed in this study which serve as guidelines for future work. As for future work, we plan to include data from Oracle Apex to incorporate an integrated approach from different e-learning systems used in the course. In this regard, another future activity is an alignment of data sources as well as resolving potential GDPR issues when implementing such integration. Finally, integrated data would serve as input into machine learning algorithms to develop predictive models of student success. At the end, the time complexity of the proposed framework should be calculated.

**Acknowledgements** The authors would like to acknowledge the support given by the European Commission through the Action Erasmus + Better Employability for Everyone with APEX (project ID 2021-1-SI01-KA220-HED-000032218), co-funded by the Erasmus+ programme of the European Union. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

## References

1. Alexander B (2006) Web 2.0: a new wave of innovation for teaching and learning? *Educause Rev* 41, 32–44
2. Austwick T (2013) Using Oracle Apex securely. *Netw Sec* 2013(12):19–20. [https://doi.org/10.1016/S1353-4858\(13\)70139-9](https://doi.org/10.1016/S1353-4858(13)70139-9)
3. Black EW, Dawson K, Priem J (2008) Data for free: using LMS activity logs to measure community in online courses. *Internet Higher Educ* 11(2):65–70. <https://doi.org/10.1016/j.iheduc.2008.03.002>
4. Cavus N, Momani AM (2009) Computer aided evaluation of learning management systems. *Proc Soc Behav Sci* 1(1):426–430. <https://doi.org/10.1016/J.SBSPRO.2009.01.076>
5. Conijn R et al (2017) Predicting student performance from LMS data: a comparison of 17 blended courses using Moodle LMS. *IEEE Trans Learn Technol* 10(1):17–29. <https://doi.org/10.1109/TLT.2016.2616312>
6. Deepak KC (2017) Evaluation of Moodle features at Kajaani University of applied sciences—case study. *Proc Comput Sci* 116:121–128. <https://doi.org/10.1016/J.PROCS.2017.10.021>
7. DeWitt D, Alias N, Siraj S, Yaakub MY, Ayob J, Ishak R (2013) The potential of Youtube for teaching and learning in the performing arts. *Proc Soc Behav Sci* 103:1118–1126
8. Feldman-Maggor Y, Blonder R, Tuvi-Arad I (2022) Let them choose: optional assignments and online learning patterns as predictors of success in online general chemistry courses. *Internet Higher Educ* 55:100867. <https://doi.org/10.1016/J.IHEDUC.2022.100867>



9. Gašević D, Dawson S, Siemens G (2015) Let's not forget: learning analytics are about learning. *TechTrends* 59(1):64–71
10. Kadoic N, Oreski D (2018) Analysis of student behavior and success based on logs in Moodle. In: Proceedings of 2018 41st international convention on information and communication technology, electronics and microelectronics, MIPRO 2018, pp 654–659. <https://doi.org/10.23919/MIPRO.2018.8400123>
11. Kadoić N, Oreški D (2021) Learning analytics of YouTube videos linked to LMS Moodle. In: 2021 44th international convention on information, communication and electronic technology (MIPRO). IEEE, pp 570–575
12. Monger A, Baron S, Lu J (2009) More on Oracle APEX for teaching and learning. In: The 7th international workshop on teaching, learning and assessment of databases, 6 July 2009. University of Birmingham, pp 3–12
13. Oreski D, Kadoic N (2018) 'Analysis of Ict Students' Lms engagement and success', economic and social development (Esd 2018). In: 35th international scientific conference, (35th international scientific conference on economic and social development-sustainability from an economic and social perspective. WE-Conference Proceedings Citation Inde, pp 434–442
14. Rudolph TE, Frankel J (2009) YouTube in music education. Hal Leonard Corporation
15. Smith DK (2014) iTube, YouTube, WeTube: social media videos in chemistry education and outreach. *J Chem Educ* 91(10):1594–1599
16. Stellefson M, Chaney B, Ochipa K, Chaney D, Haider Z, Hanik B, Chavarria E, Bernhardt JM (2014) YouTube as a source of chronic obstructive pulmonary disease patient education: a social media content analysis. *Chron Respir Dis* 11(2):61–71
17. Tolkach D, Pratt S (2021) Travel professors: a YouTube channel about tourism education & research. *J Hosp Leis Sport Tour Educ* 28:100307
18. Trabelsi O, Souissi MA, Scharenberg S, Mrayeh M, Gharbi A (2022) YouTube as a complementary learning tool in times of COVID-19: self-reports from sports science students. *Trends Neurosci Educ* 29:100186

# Influence and Optimization of Power Grid ERP System Permission Management on Enterprise Internal Control



Zhu Zuoping , Zhang Wei , Huang Yao, and Chen Tianxiao

**Abstract** To further explore the impact of ERP (Enterprise Resource Planning, ERP) system authority management on the internal management control of power grid enterprises, further standardize the internal management processes and standards of power grid enterprises, improve the internal production and management efficiency of power grid enterprises, and ensure the security of important data information in the ERP system of power grid enterprises. This paper first analyzes the influence of authority management of power grid ERP system on the internal management control of enterprises based on the background of authority management of Hunan Xiangtan Power Grid ERP system, then put forward feasible suggestions for enterprise automation optimization from the aspects of building enterprise internal control system, asset life-cycle management, internal control environment construction, supply chain cost management, improving personnel's awareness of internal control, and improving power grid ERP system authority management scheme under the power grid ERP system environment. Introduce ERP system to power grid company and related industry enterprises to participate in the modern management of enterprises, and improve the production management efficiency of enterprises, at the same time, provide reference opinions to ensure the security of important data information of enterprises.

**Keywords** Power grid ERP system · Internal control · Control optimization · Authority management · Information safety

## 1 Introduction

In recent years, with the rapid development of computer, Internet communication, enterprise informatization and system integration and other related technologies, China's power grid and related industry enterprises have gradually changed from

---

Z. Zuoping · Z. Wei (✉) · H. Yao · C. Tianxiao  
State Grid, Xiangtan Power Supply Company, Xiangtan 411007, China  
e-mail: [553066105@qq.com](mailto:553066105@qq.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_17](https://doi.org/10.1007/978-981-99-3091-3_17)

215

traditional single production management mode to modern intelligent enterprises based on computer informatization, networking and system integration. More and more power grid and related industry enterprises have introduced enterprise ERP systems in the actual production management process. At the same time, a series of authority management measures are taken to realize the security management of enterprise information, so as to improve the management level of enterprise internal control and ERP system information security. Because of this, the research on the application of ERP system in enterprise internal control and the authority management of ERP system has become a research hotspot in recent years. For example, Dai et al. [1–5] provided research countermeasures for enterprise internal control management based on ERP system environment from different dimensions; Bao et al. [6–10] studied and discussed the risk prevention and control, internal financial management and other issues of power grid enterprises based on ERP system under the background of big data era. However, in the process of transformation and development of actual production management, the introduction of ERP information system will not only have an impact on the original internal management system, management and control processes, but also have an important impact on the data and information security of enterprises due to the existence of a series of user authorization problems in the ERP system itself. The research on authority management of enterprise ERP system has become another research hotspot in the field of enterprise management informatization. For the authority management of enterprise ERP information management system, domestic and foreign scholars have done a series of related research work. For example, Zhang et al. [11] realized two-level access control of information and resources in the production management system based on the RBAC model through the resource coding mechanism. Lu et al. [12] designed a security transmission and authorization management system based on the public key infrastructure and Rose's access control principle. They added the grouping permission principle to the RBAC model, and adopted the combination of centralized permission and distributed permission management to make the model more flexible. Yuan et al. [13] analyzed the limitations of the existing RBAC0 model, established an extended RBAC model that directly faces users across roles, and designed the relevant functional modules required for access control in the document management system in view of the complexity of a certain type of equipment document management system and the difficulties in reasonable and effective organization of personnel authority management and allocation. Wang et al. [14] used ASP comprehensively, NET MVC architecture mode and EF technology have designed and implemented a rights management system. The deep decoupling of each layer of the system makes the system modules have the characteristics of “high cohesion, low coupling”, which improves the development efficiency of the management information system. You et al. [15] designed a role permission control model based on cloud computing and Docker technology based on Docker technology, combining task controller function, project controller function and user controller function. Peng et al. [16] designed a security mechanism for permission control based on RBAC. According to the user type and role, they realized the authorized access to business data and functional modules through role permission control and module allocation control. In order to

help enterprises improve production management efficiency and data information security, this paper, based on previous research results, takes the influence and optimization of authority management of Hunan Xiangtan Power Grid ERP system on enterprise internal control as the research background, on the one hand, analyzes the impact of the introduction of power grid ERP system on enterprise internal control, and then puts forward management and control optimization measures; on the other hand, this paper analyzes the impact of the power grid ERP system itself on the security of important data information of enterprises due to improper user authority configuration, and puts forward technical improvement measures that can effectively improve the security of system authority management.

## **2 Business Management Analysis of Power Grid ERP System**

The main business management contents of the ERP system of power grid enterprises include six core contents of enterprise projects, materials, equipment, manufacturing, sales and financial management. Therefore, to strengthen the authority management of the ERP system of power grid enterprises, it is first necessary to analyze the business management contents of the ERP system, with strengthening control as the actual goal, improving financial management as the core, and comprehensively analyze projects, materials, equipment, manufacturing. The specific contents of the six major businesses, such as sales and finance, form a complete content management system, and then comprehensively establish a comprehensive information management platform that can support all aspects of the daily production and operation activities of power grid enterprises. The following is an analysis of the six core business contents that need to be included in the ERP system of power grid enterprises. Strengthening the authority management of the power grid enterprise ERP system is an important measure to ensure the security of important data information in the power grid enterprise system.

### **1. Financial management business**

In financial management, it is generally based on enterprise budget items to implement real-time monitoring of the whole process of production and operation activities, and advance the specifications and accounting content of financial management to production and operation according to the relevant pre-budget, control, analysis and other stages, so as to integrate production and finance, thus improving the informatization level of enterprise financial management.

### **2. Material management business**

Starting by saving the overall cost of ownership, optimizes the existing procurement management mode, organization and process, establish a new procurement management system based on ERP system, and ensure the intensive operation effect of the supply chain.

### 3. Project management business

Standardize the project number and project category, establish a project centered comprehensive business system of cost budget, production financial management, etc. In this process, according to the ERP system management process, achieve closed-loop management effects such as project management at an early stage, implementation, decision-making, investment and post-evaluation.

### 4. Equipment management business

Establish an equipment maintenance management mode with work order as the maintenance carrier. According to the management idea of the whole life management cycle, carry out hierarchical management of the equipment use process from purchase to scrap, so as to enhance the upper and lower level monitoring level of the enterprise's equipment operation.

### 5. Manufacturing management business

The main content of production and operation activities is to formulate major production plans and material use plans according to customer needs. Through demand planning and capacity demand planning, the whole process of production is tracked, which provides an effective reference for enterprise cost accounting and effectively improves the overall business level of the enterprise.

### 6. Sales management business

In the process of sales management under the ERP system, it is necessary to standardize the management of customer basic data, establish a customer-centric operation mechanism, pay attention to customer relations, optimize enterprise organization and business processes, and then provide convenient customer service, improve customer satisfaction, and improve the comprehensive competitiveness and economic benefits of the enterprise.

## **3 The Analysis of Power Grid ERP System Authority Management**

### ***3.1 The Analysis of Power Grid ERP System Authority Management Content***

The users of the power grid ERP system, the roles set by users, and the authority of role configuration constitute the three basic elements of power grid ERP system authority management [17]. The user management, role and authorization management and

internal control of the system after the operation of the power grid ERP system belong to the scope and content of system authority management.

### 1. User management

User management includes adding user accounts (IDs) and changing, revoking and freezing account information. Users refer to users of all power grid ERP systems, including system administrators. Each user has a unique ID in the system. When creating an ID, the system will synchronously complete user information registration, including user name, work unit, department, job title, contact information and other basic information. To revoke an account is to delete a user's ID from the system so that he or she no longer has the ability to access the system. Blocking the account is to temporarily lock the user ID to temporarily lose the system accessibility, and then open the user's access to the system when necessary.

### 2. Role and authorization management

A role is a group of business operation permissions, described by the ERP system role code. A role is a collection of one or more transaction operation codes. A user can assign multiple roles, and multiple users can also be granted the same role. Permissions are assigned to users through roles. Therefore, authorization management is to grant a role with specific business operations and transaction processing permissions to a user or user group. Only after a user is granted a role can he log in to the system and run various transaction codes in his role to process business. When the user's position (job responsibility) is adjusted or the business requirements are changed, the user needs to be authorized again, including revoking the old permission and granting the new permission.

### 3. Internal control of power grid ERP system

The internal control of the power grid ERP system refers to the process of carrying out the internal control work related to the ERP system, regularly and irregularly carrying out the ERP system test and inspection, and constantly improving and perfecting the business process according to the ERP system control specifications. Among them, the most important is permission test, including access permission test and responsibility separation test.

Access permission test is to determine the scope of resources and functions that users can access in the system, and to find out whether the permissions that users have in the system exceed their work needs from the perspective of control. For example, the authority administrators of the ERP system are divided into ordinary administrators, senior administrators and super administrators. Generally, ordinary administrators are authorized to display, freeze, unlock, assign and delete user IDs. In addition to the general administrator functions, the senior administrator is also authorized to create and modify user IDs. In addition to the normal administrator and senior administrator permissions, the super administrator also has the permission to delete the user ID. According to the principle of different system access permissions, ordinary administrators cannot access the special permissions of senior

administrators and super administrators. Similarly, senior administrators cannot have the user ID deletion permission of super administrators.

The separation of duties test is to check whether the system user has the loopholes and problems in the authority allocation to deal with incompatible business functions according to the principle that incompatible business functions in the business process must be completed by different people, that is, the authority mutual exclusion check, to avoid the risk of fraud because one person has the authority to operate incompatible business. For example, in the material demand plan processing business process, the demand plan creation function and the demand plan approval function are incompatible, or they are two mutually exclusive business operations. According to the principle of separation of post responsibilities, these two functions must be undertaken by two people, respectively. Once this principle is violated, permission allocation will be out of control.

### ***3.2 The Analysis of Power Grid ERP System Authority Management Principle***

#### **1. The Principle of separation of duties**

The principle of responsibility separation is the most basic and important principle of ERP system authority management. The authority granted to users shall not have incompatible operation authority to avoid fraud risk and unauthorized modification of business, financial data and relevant information due to one person's authority to operate incompatible business.

#### **2. The principle of Business driven**

The implementation and application of ERP system have greatly improved the processing and transmission efficiency of enterprise operation information. However, if users are authorized without restriction, it will greatly increase the risk of information leakage in enterprise operations, and at the same time, increase the possibility of business confusion and management out of control.

#### **3. The principle of testing before authorization**

In order to avoid the mutual exclusion of roles and incompatibility of business in the system, user permissions must be tested for mutual exclusion and sensitive permissions under the internal control requirements of the ERP system. After the test is passed, it shall be submitted to the leader for approval. Finally, users shall be assigned roles. If the test fails, authorization cannot be granted.

#### **4. The principle of allocating user permissions on demand and minimizing permission settings**

The user permission application shall meet the business requirements of the position, and shall not exceed the actual position and responsibility scope, so as to avoid mutual exclusion and confusion of user permissions.

## **4 The Influence of Power Grid ERP System Authority Management on Enterprise Internal Control**

The power grid ERP system is an enterprise resource planning integrated management platform with information technology, network communication technology, enterprise system integration technology as the core. Its role runs through all aspects of the enterprise's internal production and operation activities. It can only improve the enterprise's actual management efficiency, reduce its production costs, and create huge benefits for the enterprise. But at the same time, the introduction of advanced power grid ERP information management system will inevitably have a certain impact on the original internal control management mode of enterprises. This section will discuss the influence of the power grid enterprise on the internal control management system, scope, form and key points of the internal control management, as well as the improper authorization of the ERP system on the internal control of the enterprise after the introduction of the power grid ERP system.

### ***4.1 The Traditional Internal Control System Cannot Keep Up with the Development Needs***

The traditional information management and data processing are based on the traditional information management system and its operation process. After the introduction of the power grid ERP system, the links between information management and data processing increase, leading to changes in the internal control of enterprises. At the same time, in the actual operation process of the enterprise, due to the mistakes in the management process, such as weak management awareness and other issues, it will also lead to the loss of important data information in the enterprise ERP system.

### ***4.2 The Scope of Internal Control Continues to Expand***

The internal control in the power grid ERP system is different from the traditional internal control system. In addition to maintaining the integrity of the basic user information in the system, the basic analysis and decision-making ability of the system, and the legitimacy of the user role authority configuration, it is also necessary to ensure the working efficiency of the overall operation of the system, not just the user post responsibility authority management that the traditional internal control management needs to focus on.



### ***4.3 Traditional Internal Control Forms Have Changed***

In the environment of power grid ERP system, the accuracy of data information no longer needs to be ensured by traditional internal control means such as bill management. Users only need to input the original data into the computer, which can realize the comprehensive sharing of enterprise internal information, and automatically generate a large number of required system bills, greatly improving the efficiency. Therefore, in the internal control of enterprises, the traditional form of data information control should be transferred to the computer, such as the preparation and verification of enterprise financial statement information.

### ***4.4 The Focus of Traditional Internal Control Has Shifted***

In the power grid ERP system, most of the production and operation activities of enterprises are completed by computers, and the internal control of enterprises will inevitably change relative to the traditional internal control. Therefore, the internal control of enterprises needs to focus on information input, human-computer interaction, computer connection control, structural organization, etc.

### ***4.5 Unreasonable Authorization Control of Power Grid ERP System Will Affect the Internal Control of the Enterprise***

According to the ERP system user authority management content and setting principles, the user's application requirements, access requirements, operating system and business processing requirements are directly controlled by the ERP system authority. In the authorization management work, if the provisions of the authorization principle are violated, the system operation and business will be confused, and the potential risk of illegal tampering of operation fraud, business data and financial data will be greatly increased, which will inevitably result in out of control management. Therefore, as the most important work content of the operation and maintenance of the grid ERP system, authority management directly affects whether the grid ERP system can operate safely, efficiently and stably, and also affects the quality of the internal control work of the enterprise.

## **5 Power Grid ERP System Authority Management and Enterprise Internal Control Optimization**

### ***5.1 Internal Control and Self-control Optimization of Power Grid Enterprises***

In view of the negative impact that the poor self-management of the internal control of power grid enterprises may have on the production, operation and management of enterprises, this section discusses the optimization of the enterprise's self-management and control from the aspects of building the enterprise's internal control system under the power grid ERP system environment, asset life-cycle management, internal control environment and system construction, supply chain cost management, and improving personnel's awareness of internal control.

#### **1. Building the internal control system of enterprises**

In order to achieve the optimization measures of power grid enterprise ERP system authority management and control, it is necessary to formulate strict rules and regulations of the enterprise itself, establish and improve the internal control system of the enterprise, and improve the system use management system and system structure framework to meet the top-level needs of enterprise self-regulation optimization.

#### **2. Strengthen enterprise asset lifecycle management**

The whole process life-cycle of enterprise asset management includes the planning and design stage, the purchase and construction stage, the operation and maintenance stage, the renewal and transformation stage, the disposal and upgrading stage and other important links. In order to improve the enterprise's self-control and optimize the efficiency of the enterprise's self-management adjustment, it is necessary to set up appropriate monitoring mechanisms in these important links to ensure the scientific rationality of process management and further affect the overall efficiency.

#### **3. Strengthen the construction and management of enterprise internal control environment**

Enterprises should pay special attention to the internal environment construction while improving the external environment construction, and constantly improve the comprehensive quality of all managers. The enterprise leadership should organize relevant corporate culture activities, continue to strengthen the communication and trust between the enterprise management and employees, improve the internal cohesion of the enterprise, and then establish the unique corporate culture of the enterprise. The enterprise shall also timely organize employees to participate in management training and learning, so as to ensure that employees understand the internal control system of the power grid enterprise as well as the basic architecture and usage of

the power grid ERP system, so that the power grid ERP system can better serve the purpose of production, operation and management.

#### 4. Improve supply chain cost management

In the power grid ERP system, the power grid enterprise can timely publish the qualification requirements and bidding announcement of each bidder, and through the ERP system, it is responsible for purchasing the relevant data of raw material suppliers and project contractors in the project and coordinating them. In this process, it can effectively realize the integration and connection effect of funds, information and logistics in each link of the supply chain, and effectively reduce the cost problems such as time and funds caused by collaboration. At the same time, the implementation of ERP system has realized the resource integration of upstream enterprises and information sharing among enterprises. After being put into use, accounting vouchers can not only be automatically transferred to the financial system when business occurs, but also be integrated through the report generation function developed by the system. Here, through direct extraction and generation of data between different functions and enterprises, the problem of separation between modules within the enterprise is solved, the real-time and accuracy of enterprise financial information is significantly improved, and the standardized and intelligent supply chain management of engineering projects is realized.

#### 5. Strengthen training and enhance the awareness of managers to actively participate in the internal control of enterprises

Enterprises should pay attention to improving the awareness of managers to actively participate in the internal control of enterprises, strengthen the management skills training of employees in a timely manner, and gradually promote the information construction, so that managers can understand the internal control methods of ERP system and information platform. In the training process, enterprises need to ensure that managers can distinguish between the main performance management and non-core indicators, and truly find out and deal with the internal control contradiction of budget performance management. Secondly, when strengthening managers' awareness of internal control, we should ensure that managers understand the internal control and budget performance management of the enterprise, so as to achieve the strategic objectives of the enterprise; Through the ERP management system, timely find the problems that need to be solved, and make effective feedback and adjustment.

### ***5.2 Power Grid ERP System Authority Management and Enterprise Internal Control Optimization***

According to the content and principles of power grid ERP system authority management, the following targeted authority management optimization measures are

proposed in this paper, aiming at the possible adverse effects of improper power grid ERP system authority management on enterprise internal control management.

1. Set up system administrator, system auditor and system confidentiality officer, respectively, to perform their duties and ensure system security.

The system administrator is responsible for responding to the requirements from the end user and performing corresponding operations in the system, including account application, account locking and unlocking, authorization and change of authority. At the same time, it is responsible for implementing the normative review and management of evidence forms to ensure that the forms are effectively approved. The system auditor is responsible for the authority mutual exclusion test and inspection to ensure that the applied authority complies with the internal control management regulations and authority management principles. At the same time, he is responsible for collecting and sorting out user flow information and post change dynamics, sorting out business responsibilities, assisting the system administrator in sorting out user accounts and clearing permissions according to user dynamics, closing, revoking or freezing unnecessary idle accounts in a timely manner, deleting users' old permissions in a timely manner, and adding new permissions. The system secrecy officer is responsible for verifying the legitimacy of the user's identity, ensuring that no account is opened for illegal applicants, so as to ensure the access security, information security and business security of the system.

2. Establish a standardized power grid ERP system authority management system, strengthen the authority authorization control, make the authority application process clear, and the authority allocation reasonable, compliant, and controllable.

After the operation of the power grid ERP system, it is very important to establish a mature ERP system authority operation and maintenance management system for the power grid ERP system management. In the actual work, we can not maintain the authority completely through technical means, but also need to establish the corresponding authority management system to standardize the authority management, so as to achieve the management process and standardization. Use the management system to restrict the authority management work and reduce the negative effect of the subjective consciousness of users and administrators.

3. Design permission control solutions.

Power grid ERP system authority control is divided into three levels: transaction code level, organization level and authority object field value level. To perform business operations, users must have the corresponding transaction code execution permission and the corresponding organization operation permission to complete business. According to the actual situation of the enterprise, different organizational structures are designed in the power grid ERP system according to the business characteristics. The highest organizational level in the system is the company code. You can limit users' business to different company codes by assigning different company code values to roles. After the user has the execution permission of the transaction code and

the corresponding organization operation permission, the system further controls the corresponding operations according to the permission object, such as adding, modifying, deleting, etc. Therefore, each authorization object must be specifically set to complete the final control of permissions.

4. A clear system of user post responsibilities has been established to ensure that the division of user post responsibilities is reasonable, orderly and free of conflicts, and that there is no mutual exclusion of business processing permissions.

According to the principle of “one person, one post”, the user account, permission application and permission allocation are managed in the power grid ERP system, which can quickly and efficiently solve the problem of confusion of user permission authorization and serious mutual exclusion of roles. In fact, under the condition that enterprises are constantly pursuing refined and flat management, minimizing human resource costs and maximizing benefits, the phenomenon of “one person with multiple posts” is very common, and it is inevitable that users’ post responsibilities conflict and business processing permissions are mutually exclusive. Therefore, without a clear job responsibility system, it is impossible to ensure that users are reasonably divided and compliant in terms of job responsibilities, which will eventually lead to out of control user authority distribution and disordered management.

5. The power grid ERP system application department and authority management department shall establish an effective communication and coordination system with the personnel management department to form a smooth user information feedback mechanism.

The user account of the power grid ERP system corresponds to the user’s actual position. The user’s position and business responsibilities determine the role and authority that his/her account should be assigned in the system. Therefore, the business department and the personnel management department need to establish an effective communication and coordination system and information feedback mechanism. When personnel positions and responsibilities are adjusted, the personnel management department should timely feed-back relevant information to the business department and authority management department, so that the business department and authority management department can react quickly, revoke and freeze user IDs in a timely manner, or change or delete user permissions, to minimize the system operation risk.

6. In accordance with the internal control specifications of the power grid ERP system, the accounts of the power grid ERP system shall be sorted regularly to improve the effective utilization rate of the accounts.

User account management of power grid ERP system is the most basic work in authority management. To do a good job in power grid ERP system authority management and ensure clear and standardized management procedures, account management should be standardized first. Therefore, according to the internal control specifications of the power grid ERP system, it is very necessary to regularly sort out

the accounts of the power grid ERP system to improve the effective utilization of accounts.

7. The internal control management department is involved in the power grid ERP system authority management, and controls the authorization management of the power grid ERP system from the internal control management system level.

The internal control management department is the business guidance department for the internal control management of the power grid ERP system. The internal control management department is responsible for assisting in the operation and maintenance of the power grid ERP system. However, according to the current situation of the operation and maintenance management of the ERP system, the internal control department is basically separated from the operation and maintenance of the ERP grid system, and often requires the executive department to rectify after the internal control test detects problems. Therefore, the internal control management department should perform the responsibility of assistance and really participate in the management of power grid ERP system authority.

## 6 Conclusion

Scientific ERP implementation method and management are the key to the successful application of ERP, and efficient ERP operation and maintenance management system is the reliable guarantee and means for the safe, efficient and stable operation of the system. In recent years, with the rapid development of computer, Internet communication, enterprise informatization and system integration and other related technologies in China, the management application level of ERP system in all walks of life in China has achieved initial results, but it still faces many new challenges. For example, in the process of introducing ERP system to participate in the enterprise's modernization, electronic informatization and comprehensive intelligent processing, how to overcome the compatibility problems between the ERP system management module and the enterprise's traditional management model and system, and how to configure user roles and grant role permissions in the ERP system operation management process to better protect the security of important data information in the system. In order to overcome the incompatibility between the ERP system management module and the traditional management mode of the enterprise, and ensure the security of the important data information of the enterprise during the operation and maintenance of the ERP system, this paper takes the authority management of the ERP system of Xiangtan Power Grid in Hunan Province as the background, and analyzes in detail the business management content, authority management content and principles of the ERP system of the power grid, as well as the impact of the authority management of the ERP system of the power grid on the internal control of the enterprise, it also puts forward targeted suggestions on the optimization management of enterprise self-control, which can provide effective reference suggestions for the practical application and management of ERP system in relevant enterprises.

However, the views in this paper are only observed and analyzed based on the actual application of power grid ERP system authority management in Xiangtan Power Grid Company in Hunan Province. Therefore, some of the views in this paper may not have universal adaptability, and this paper only describes some macro-control views and automatic control management optimization suggestions in the actual application of power grid ERP system in Xiangtan Power Grid Company in Hunan Province. The paper does not provide a more detailed and feasible implementation plan of ERP system authority management technology. These shortcomings will be the next research work in this paper.

**Acknowledgements** This work was supported by the Authorization Integration Management Automation Technology Research Project of State Grid Xiangtan Power Supply Company under Grant. (No. 5216C0220007).

## References

1. Dai J (2021) Countermeasures for improving enterprise internal control in ERP system environment. *Enterprise Reform Manag* 9:11–12
2. Ye H (2021) Research on enterprise internal control based on ERP system. *Econ Res Guid* 16:96–98
3. Li AH (2021) Research on internal control of small and medium-sized real estate enterprises based on ERP environment. *Enterpr Reform Manag* 2:25–26
4. Song SS (2021) Research on internal control of foreign trade enterprises based on ERP system environment. *Chinese Market* 29:179–180
5. Li FQ (2020) Research on enterprise internal control from the perspective of ERP. *Heilongjiang Sci* 11(1):116–117
6. Bao Q (2021) Research on systematic risk prevention and control strategy of bidding procurement of power grid enterprises based on the perspective of internal control. *Bus Account* 3:104–106
7. Han ZW (2020) Research on financial internal risk control of power grid enterprises in the age of big data. *Caixun* 17:144
8. Lei AH (2020) Discussion on the construction of enterprise accounting internal control system—taking power grid enterprises as an example. *Chinese Bus Rev* 18:152–153
9. Zheng Y (2020) Analysis and countermeasure research on internal control of financial risk of power grid enterprises in the age of big data. *Account Study* 25:184–185
10. Wang YY (2020) Analysis of internal control of power grid enterprises based on risk management. *Caixun* 31:104
11. Zhang S, Yan M, University XA et al (2016) Application and implementation of production management system based on RBAC authorization management. *Microcomput Appl*
12. Lu G, Zhao L, Yang K (2015) The design of the secure transmission and authorization management system based on RBAC. *IEEE*
13. Yuan D (2017) Design and implementation of document management system based on extended role-based access control model. *Ship Electron Eng*
14. Wang JX, Wang HZ (2021) Design of authority management system based on MVC. *Electron Compon Inform Technol* 5(12):165–172
15. You L, Sun H, Gong D (2022) Research and design of docker technology based authority management system. *Comput Intell Neurosci* 2022:5325694–5325694
16. Peng SX, Peng P (2021) RBAC based B/S structure student fee system security mechanism. *J Shantou Univ* 36(01):12–20

17. Ren ZK, Yang M (2015) Impact of ERP system authority management on enterprise internal control and management and control optimization. *Petroleum Plan Des* 26(02):43–46



# Ensemble Feature Selection and Classification of Medical Dataset Using K-Nearest Classifier with Swarm Intelligence



Ebtesam Shadadi, Saahira Banu Ahamed, Latifah Alamer,  
Mousa Khubrani, Iman Mohammad Alqahtani, and Aisha Sumaili

**Abstract** Medical Field are growing tremendously and using various IT related services for data analysis and diagnoses. The numerous amounts of computer assisted systems are available to measure accurate results. The clinical dataset consist of huge amount of disease information and which marks negative effects with respect to features. Feature selection is major key player removing redundant information. It is also increase the decision making result effective manner. The important step in feature selections is classification and dimensionality reduction. Here we select medical information such as multiple attributes, data mining approach and disease prediction using deep convolution neural networks. Various data analytics techniques are classified to predict the disease using feature selection method. In this paper we evaluate principle component analysis, support vector machine, factor analysis and ranking methods are compared with our proposed CNN-based ensemble feature selection using K-nearest classifier. Our proposed method is followed as standard testing features with parameter optimization. Compare with existing method our method has 96% of accuracy result by using TensorFlow simulator.

**Keywords** Feature selection · Classification · CNN · K-nearest classifier · Swarm intelligence

---

E. Shadadi · S. B. Ahamed (✉) · M. Khubrani  
Department of Computer Science, College of Computer Science and Information Technology,  
Jazan University, Jazan, Saudi Arabia  
e-mail: [sahamed@jazanu.edu.sa](mailto:sahamed@jazanu.edu.sa)

M. Khubrani  
e-mail: [mmkhubrani@jazanu.edu.sa](mailto:mmkhubrani@jazanu.edu.sa)

L. Alamer · A. Sumaili  
Department of Information Technology and Security, College of Computer Science and  
Information Technology, Jazan University, Jazan, Saudi Arabia  
e-mail: [laalamer@jazanu.edu.sa](mailto:laalamer@jazanu.edu.sa)

I. M. Alqahtani  
Computer and Information Systems Department, King Khalid University, Abha, Saudi Arabia  
e-mail: [alqahtani@kku.edu.sa](mailto:alqahtani@kku.edu.sa)

# 1 Introduction

Medical experts and researchers are using various machine learning techniques to predict and analysis the huge volume of dataset. The large scale volume of medical dataset is having number attributes and dimensions. Dimensions can affect the analytics and diagnose the results. So it increases the running time and turnaround time so process can wait still existing execution will complete [1, 2]. Curse of dimensionality is the major problem in high dimension medical dataset so feature selection is an important technique to handle the dataset. Researchers are analyzed various data analytics applied for pre-processing the dataset [3, 4].

The feature selection is applied in dataset and analysis the dimension of each attribute. In this case over filtering [5], reduce the dataset size [6], complexity, performance [7] are major factors considered for processing. It has two categories such as extraction and selection. In this case heart, cancer and diabetes medical dataset are taken from UCI Data Repository [8]. The high probabilities are taken such as heart attacks [9], hypertension [10] and anemia diseases [11]. Multiple tests are taken for evaluating disease and predicting accurate results [12]. Table 1 shows that description of medical dataset.

The medical dataset contains 20 attributes with 800 instances. The different classifiers are used such as neural network [13], decision tree [14], support vector machine [15] based on selected or pre-processed dataset. The accuracy can be measured based on attributes and features. In existing cases apriori-based genetic algorithm is used for structured or query optimized dataset. Feature selection is exponential approach with

**Table 1** Medical dataset description

Attribute	Description
Age group	A # Value: below 30 B # Value: 31–50 C # Value: Above 50
Gender	A # Value: Male B # Value: Female C # Value: Others
Corresponding specification (Cp)	A# angina, B# non angina C# pain, D# asymptotic
Test blood pressure results	Category 1: 0–120 Category 2: 121–150 Category 3: 150–max
Cholesterol	Category 1: 0–200 Category 2: 201–250 Category 3: 250–max
Cardio result	A# Value 1: unsloping B# Value 2: downslopping
Result	Value 1 - 3: normal Value 4–6: detect as beginning Value 7-above: abnormal

exhaustive approach with  $2^n$  subset dataset with  $n$  number of possibilities [16]. In this paper, metaheuristic approach as ensemble k-nearest neighbor classifier with swarm intelligence feature selection algorithm is proposed for evaluating medical dataset. For classification support vector machine with CNN approach can be done [17, 18]. The results are compared with existing methods. This paper structure as follows as, Sect. 2 gives about various problem statements literatures, Sect. 3 describes feature selection and description, Sect. 4 discuss about simulations and experiments setup using TensorFlow and Sect. 5 gives conclusion and future plans.

## 2 Related Works

Medical and Healthcare professional are used and stored huge volume of patient data in computerized storage systems. It could be extract the data using decision support method using data mining and statistical approach. For example Wonk et al., the risk factors of heart diseases are age, blood pressure, higher cholesterol, diabetes, family history and hypertension [10]. More number of physical activities are required for reducing the risk factors. Various researchers are suggested different data analytics techniques can be applied for prediction of accuracy such as naive Bayes, neural network, kernel process, support vector machine. Runge et al. suggested the healthcare professional diagnosis the heart disease with respect to logistic regression approach has 77% [19].

Gian Hug et al. Classified Conception clustering method is proposed with highest accuracy index of 90% with respect to neural network optimization [20]. CleverLand Dataset can be used for testing the results [21]. Principle Component analysis method has good classification factor and cross validation approach is applied for predicting accuracy. The various literatures are reported as there is no strong classifier for predicting the accuracy factor especially medial dataset. The maximum result is achieved as 77% using neural network [22].

Manikandan et al. the size of dataset is large so need to concentrate the problem in the case of constructing dataset and applying pre-processing. While applying dimensionality reduction serious issues to be considered for selecting features, it creates major issue with respect to association and normalization [23]. The successful association rule mining is required for reducing dimension and redundant information. Most of decision support system has best classifier for reducing dimension with good accuracy index [24].

Xia Guo Well et al. Ant bee colony feature selection is used for detecting liver disease and hepatitis. Support vector machine was applied for supervised and unsupervised learning such as clustering [25] and classification [26]. The feature selection obtained from clusters and divides two parts of centers with respect to optimal and trained features. Based on literatures SVM parameters provides good accuracy factor for feature selection [27]. Above literatures, we need good classifier and simulator required for predicting accuracy index with minimal execution time [28].

### 3 Ensemble Feature Selection and Classification

In our proposed system has three attributes feature selection methods. Classification is represented for providing solution for handling huge volume of dataset. Here we proposed support vector machine with the combination of CNN and K-nearest classifier method. The feature set is very small we can get accuracy index easily but huge volume of dataset means we need to do various pre-processing step for reducing dimensions. One of the major techniques is principle component analysis which transforms original features into represented space with accurate dimensions. It preserves swarm intelligence optimization technique for feature extraction.

The pre-processing is applied to select the attributes with significant features, classified tasks, dominant factor and dimensions. We can apply multiple techniques means it reduce the accuracy and major issue is execution time always increases. So the original dataset always remind as keep and multiple copies of dataset can be used for evaluations. In this case the below steps are used to process the selection and extraction.

Step 1: Any format of dataset can be normalized and converted to pre-processing stage.

Step 2: Select the attributed and use classifier for reduction of dimensions.

Step 3: Dataset can be modeled with less feature and size also reduced with optimization index.

Step 4: Allocate the classifier values to simulator for generating trained and test dataset.

Step 5: Simulate the environment with our proposed system.

Step 6: Accuracy is compare with existing methods such as factor analysis, ranking and SVM classifier.

The support vector machine has different classes with hyperplane dataset and instance. We used KNN classifier feature selection se used divide and conquer decision tree process for reducing dimensions. The entire instances are recorded with respect to possible solutions. If representation of training dataset instance is  $n$  with the vector  $xi \in D$  ( $D$  is the feature  $i = 1, 2, \dots, n$ ) the corresponding instance class  $yi$ .

So the hyperplane reduction has

$$(w \cdot xi + a) = 1 \text{ and } (w \cdot yi + b) = 1 \quad (1)$$

where  $w$  is the dataset and  $a, b$  are the attribute instance with probability factor.

In this way dimension can be obtained as reverse as follows based on clusters

$$(w \cdot yi + a) = -1 \text{ and } (w \cdot xi + b) = -1 \quad (2)$$

To obtain this the instance values are recorded as support vector which is represented as

$$yi(w \cdot xi + b) \geq 1 \text{ and } xi(w \cdot yi + a) \geq 1 \quad \text{for } 1 \leq i \leq n \quad (3)$$

Based on above results wrongly classified dimension values are identified and removed the labels. So we can easily add outlier labels and mistakes feature. It is real time model to imperfect classified results are defined as follows:

$w \cdot (xi + yi) \in 1$ ,  $\epsilon i > n < 1$  it is slack variable belongs to quadric equation process

$$\frac{1}{N} |w| = \sum_{i=0}^{n-1} \frac{W(xi + yi)}{N} \quad (4)$$

where the wrongly classified instance is stored in misclassified dataset which is not permitted for evaluation stage. The hyperplane values are selected and considered for dimensionality reduction. The both models have linearity property index using SVM classifier. So the kernel process is applied for more complex feature selections. The higher dimension space values are mapped with kernel sub faction and each condition are obtained from radial basis representation as follows:

$$K(x, y) = ||w|| + (a + b)/n|2| \quad (5)$$

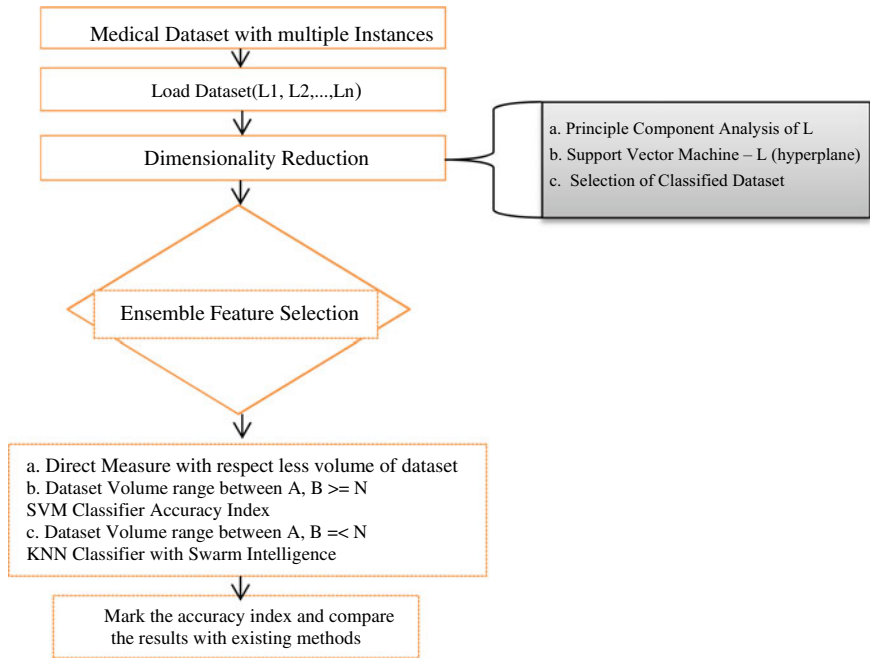
The SVM model can be added with KNN classifier which is shown in below algorithm where the optimal value is represented as  $(C, \gamma)$ . The proposed method with swarm intelligence feature can be represented as Fig. 1.

Let  $S = \{i_1, i_2, \dots, i_n\}$  set of items and Turnaround Time  $Tr = \{t_1, t_2, \dots, t_n\}$  for all the active  $n$  transactions. The subset items are selection based on kernel distribution index so the property as marked as support count

$$\begin{aligned} X_{\text{support}} &= \text{Count}(X) / Tr \\ X_{\text{confidence}} &= X_{\text{support}} \left( Tr \sum_{i=0}^{n-1} W \right) \end{aligned} \quad (6)$$

Based on above association rule each dataset items are recorded and appears in dimensionality reduction index calculations.

From the above support and confidence item set which has  $P(x \in y) = P(x) \cdot P(y)$  so the correlation item set correlation index of each selected values index (Table 2).



**Fig. 1** Process diagram for ensemble feature selection and classification

**Table 2** Principle components analysis condition for feature selection

Principle component analysis has three factor techniques such as best, mean square and linearity index. The covariance matrix generated based on below conditions
a. The input data is normalized and large dataset attributes are classified by using clustering
b. K-nearest orthogonal vectors are normalized by using hyperplane
c. Sorted the order using least significance to strength values using covariance index calculation
d. Each sorted components are stored in space like weaker and stronger indexing representations
e. The strongest components are taken for approximation indexing

#### 4 K-Nearest CNN Classifier with Swarm Intelligence Algorithm for Feature Selection

K-nearest CNN classifier is the approach for calculating nature inspired population in the category of swarm intelligence. The ideation process with inspiration of each dataset instance feature can be selected. This algorithm is suitable for clustering, classification, path planning, agent interaction and optimization. This method is very suitable for our proposed method for evaluating feature selection with agent modeling. This combined approach for reducing dimensions, feature identity and clustering. This case K-means clustering represented as m clusters and n solutions

are obtained. The below pseudocode is the new solution for generating features (Table 3).

The probability  $P$  represented as the one solution from  $X_{\text{selected}}$  valued and new one is saved as  $^{\text{New}}X_{\text{selected}}$ . So the new values are recorded as,

$$^{\text{New}}X_{\text{selected}} = X_{\text{selected}} + \zeta * n(\mu, \sigma) \quad (7)$$

From above selected values  $n(\mu, \sigma)$  is the means and variance factor from Gaussian random forest identifier representation. So the  $\zeta$  is obtained as

$$Z = \text{Log}(\text{sig}(0.5 + \text{Max}(i) - \text{Current}(i)/K) X_{\text{rand}} X(f) \quad (8)$$

In this work Swarm intelligence is used for selecting feature and metaparameters is obtained from binary solution from lower and upper bounders as 0 and 1, respectively. It is continuous function in the range between lower and upper index [0,1]. The threshold can be set as following function,

$$\text{Threshold } (T_{\text{hr}}) \begin{cases} 1 - n >= t \geq N \\ t = 0 \end{cases} \quad (9)$$

**Table 3** Algorithm—feature selection and accuracy index

---

**Algorithm: K-Nearest CNN Classifier—Feature Selection**

---

*Initialization phase*

---

Select Initial Count-Population Index n random variables

Reduce and Divide population Index n into m clusters using K-means algorithm Examine the solutions and select best m from centers

$r = r_{\text{and}}(0,1)$  if  $r_{\text{and}}(0, 1) < P_a$  then

Select r cluster center and replace F random variable  
end if

repeat

Generate New Variable N

if  $r < P_b$  then

Select the Cluster  $C_x$  with probability  $P_a$  and  $P_b$

$r_{\text{and}}N = r_{\text{and}}N(0,1)$  if  $r1 < P_{bi}$  then

Change the  $C_x$  cluster and Random variable Nn

else

Select clusters  $C_N \in N$

$r_{\text{and}}C = r_{\text{and}}C(0,1)$

if  $r_{\text{and}}C < P_c$  then

Solution from Cn

end if

Compare New vs Old Solution Random Index R

Continue

Until New solution X

Until iteration number = Max (X)

Record the Index

---

From the above results accuracy is obtained from number of chosen feature and cross validation which obtained as follows:

$$F(X) = (100 - \text{Accuracy\_Index}) + \sum \text{noFeature/SelectedFeature} \times 10^{\log(\text{No. of Feature})/N} \quad (10)$$

## 5 Simulation Results

The proposed algorithm is evaluated by using TensorFlow simulator with Intel i5 CPU at 3.5 Ghz, 8 GB RAM, GPU Computing, Windows 10 OS. The medical dataset selected from UCI repository with cancer, diabetes and disorder feature. The detailed dataset are shown in Table 4 and each dataset divided by using deep learning dimensionality reduction features.

For the data analytics process WEKA tool is used for classification for different dataset features. The data is pre-processed and discretization is done by using decision tree approach. Figure 2 shows that dimensionality reduction with attributes index (Table 5).

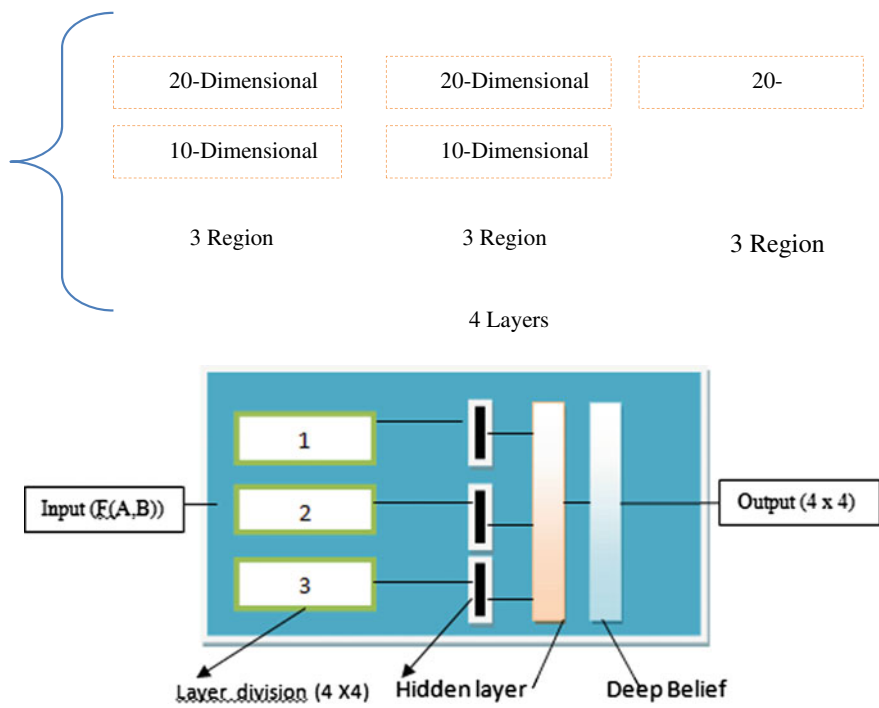
From Table 6, the accuracy is calculated using TensorFlow simulator. Number iteration is applied to each dataset based on nodes, classes and features. So we can more accurate results. Based on above table representation the average accuracy index in 96% is obtained. The proposed KNN-based swarm intelligence classifier has 5, 10, 20, 50 nodes with respect size n and the parameter can be varied based on  $20 \times 10$  dimensionality reduction possibilities. In this paper we compare the accuracy index with various existing methods. The below Table 7 shows that various classification and prediction index factors.

From the above Table 7 shows that the comparison of existing methods such as support vector machine, factor analysis, ranking framework and proposed KNN classifier methods. The performance compared with accuracy index in percentage and turnaround time in ms. Compared with existing methods our proposed methods has good accuracy index and less turnaround time (Figs. 3 and 4).

**Table 4** UCI repository medical dataset features

S. No	Dataset	Classes	Instance	Features
1	Diabetes	4	1243	98
2	Cancer	4	989	81
3	Liver	5	890	102
4	Heart	7	1561	112





**Fig. 2** Deep belief network generation of medical dataset features

**Table 5** Deep belief network generation for simulations

Trained inputs (A, B)	Selected feature–hyperplane
Tool	KNN and TensorFlow
layers	4 layers and 3 region deep belief network
Connected components	Classes selection
Hidden components	Instance representation
GPU	3.5 Ghz deep swarm intelligence
Dimensionality	20 × 10
Dataset	Diabetes, cancer, heart, liver

**Table 6** Accuracy results obtained from TensorFlow

Dataset	Iterations	Nodes	Classes	Features	Accuracy
Diabetes	1	5	4	98	95
	2	10	4	81	96
	3	20	5	102	95
	4	50	7	112	94
Cancer	1	5	4	98	94
	2	10	4	81	95
	3	20	5	102	95
	4	50	7	112	98
Liver	1	5	4	98	96
	2	10	4	81	97
	3	20	5	102	96
	4	50	7	112	96
Heart	1	5	4	98	97
	2	10	4	81	95
	3	20	5	102	92
	4	50	7	112	94

**Table 7** Comparison of existing feature selection methods with our proposed approach

Dataset	Support vector machine		Factor analysis		Ranking framework		KNN classifier	
	Accuracy	Turn around time	Accuracy	Turnaround Time	Accuracy	Turn around Time	Accuracy	Turn around Time
Diabetes	88	0.98	78	1.87	81	1.92	95	0.67
Cancer	86	1.21	75	2.23	83	2.28	96	0.72
Liver	85	1.56	79	2.87	85	2.91	96	1.02
Heart	84	2.43	81	2.90	84	3.21	95	1.87

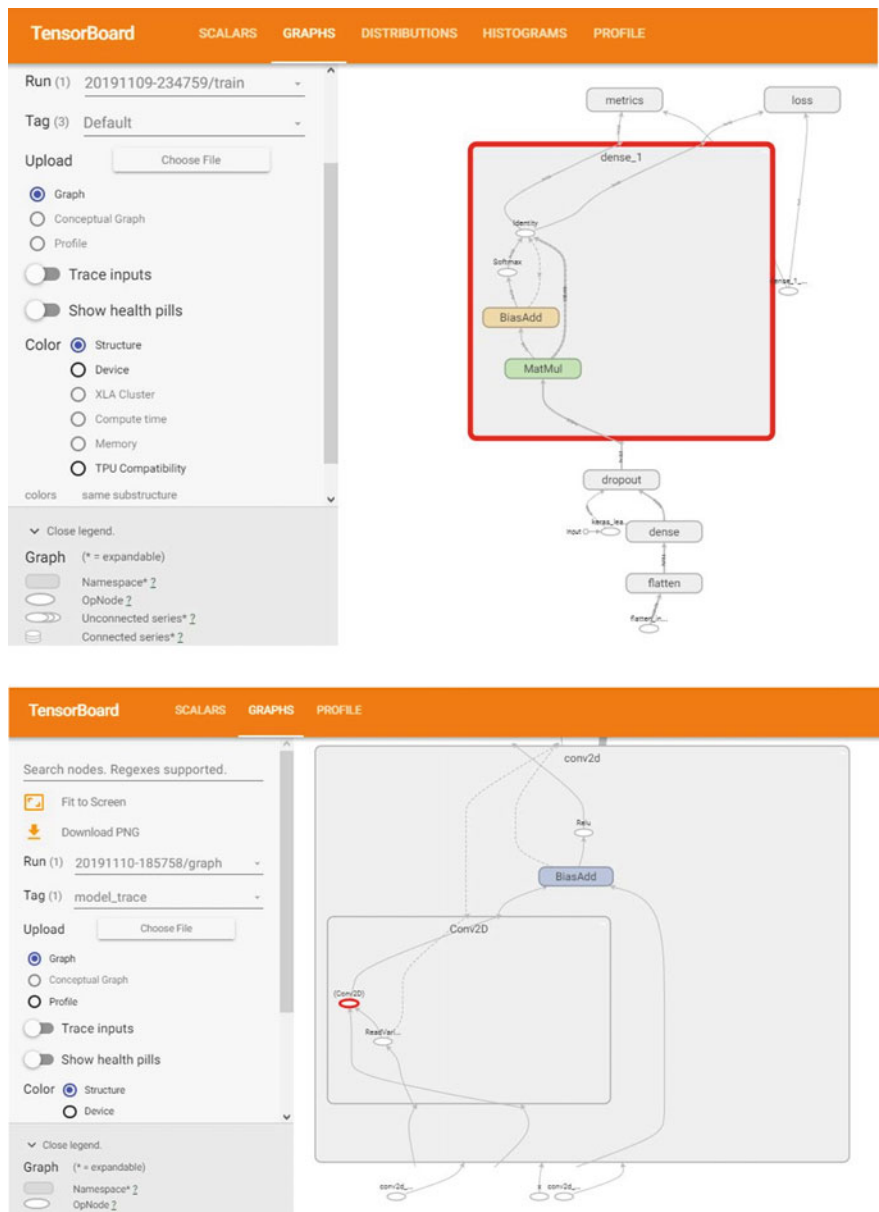
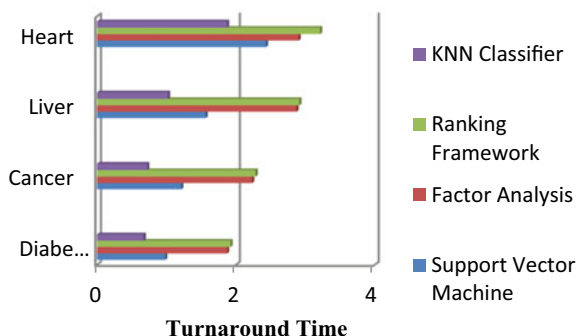


Fig. 3 TensorFlow result of deep belief network simulation

**Fig. 4** Comparison graph of different classifier



## 6 Conclusion

In this work, we proposed ensemble feature selection and classification method for predicting accuracy of medical dataset. The proposed KNN classifier swarm intelligence method gives better result for liver, heart, cancer and diabetes dataset. It is binary solution to select the feature and extract the feature based on attributes and instance factors. The large volume of dataset are diagnosed with various trained and test data using deep belief network. TensorFlow simulator is used to predict the features based on that accuracy and turnaround time obtained by using number iteration in each selected dataset. KNN classifier is applied for classifying dataset and features are selected to evaluate the accuracy index. In this case TensorFlow simulator is used to performance the deep belief network optimization. Here the proposed system evaluated to measure the performance indicators. Our proposed system has achieved 96% accuracy index and compared with existing methods KNN classifier provided good results. In future different modeled or higher volume of dataset can be used for simulating the results. Same situation we will use KNN classifier for measuring real time or multimedia dataset or data stream applications.

## References

1. Manikandan S, Dhanalakshmi P, Rajeswari KC, Delphin Carolina Rani A (2022) Deep sentiment learning for measuring similarity recommendations in twitter data. *Intell Autom Soft Comput* 34(1):183–192
2. Velmurugan P, Renukadevi M (2017) Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture based clustering features. *Artif Intell Syst Mach Learn* 9(1):8–10
3. Tuba E, Tuba M, Jovanovic R (2017) An algorithm for automated segmentation for bleeding detection in endoscopic images. In: *International joint conference on neural networks (IJCNN)*. IEEE, pp 4579–4586
4. Dolicanin E, Fetahovic I, Tuba E, Capor-Hrosik R, Tuba M (2018) Unmanned combat aerial vehicle path planning by brain storm optimization algorithm. *Stud Inform Control* 27(1):15–24

5. Emary E, Zawbaa HM, Hassanien AE (2016) Binary grey wolf optimization approaches for feature selection. *Neurocomputing* 172:371–381
6. Manikandan S, Radhika KSR, Thiruvenkatasuresh MP, Sivakumar G (2022) Deepq: residue analysis of localization images in large scale solid state physical environments. In: AIP conference proceedings 2393, 020078
7. Tubaa E, Strumbergera I, Bezdana T, Bacanina N, Tuba M (2019) Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine. In: 7th international conference on information technology and quantitative management (ITQM 2019). Elsevier Science Direct Computer Science Procedia
8. Palani Murugan P, Chinnadurai M, Manikandan S (2022) Tour planning design for mobile robots using pruned adaptive resonance theory networks. *CMC Comput Mater Continua* 70(1):181–194. <https://doi.org/10.32604/cmc.2022.016152>
9. Wang H, Khoshgoftaar TM, Van Hulse J (2010) A comparative study of threshold-based feature selection techniques. In: 2019 IEEE international conference on granular computing, pp 499–504
10. Tu MC, Shin D, Shin D (2019) Effective diagnosis of heart disease through bagging approach. *Biomed Eng Inform IEEE*
11. West D, Mangiameli P, Rampal R, West V (2013) Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. *Eur J Oper Res* 162:532–551
12. Wajs W, Wais P, Swiecicki M, Wojtowicz H Artificial immune system for medical data classification. In: Proceedings of 2015 international conference on machine learning, China
13. Cheng TH, Wei CP, Tseng VS (2016) Feature selection for medical data mining: omparisons of expert judgment and automatic approaches. In: Proceedings of the 29th IEEE symposium on computer based medical systems, pp 165–170
14. Kohavi R, John G (2017) Wrappers for feature selection. *Artif Intell* 97(1–2):273–324
15. Chuang LA, Chang H, Tu C, Yang C (2018) Improved binary PSO for feature selection using gene expression data. *Comput Biol Chem* 32(1):29–38
16. Yin Y, Yangqiu S, Zhang M (2017) NNEMBs at SemEval-2017 Task 4: neural twitter sentiment classification: a simple ensemble method with different embeddings. In: Proceedings of 11th international workshop on semantic evaluation, pp 621–625
17. Rouvier M, Favre B (2016) SENSEI-LIF at SemEval-2016 Task 4 : polarity embedding fusion for robust sentiment analysis. In: Proceedings of 10th international workshop on semantic evaluation, pp 207–213
18. Srivastava N et al (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15.1:1929–1958
19. Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of 2014 conference on empirical methods in natural language processing, pp 1532–1543
20. Ojha U, Adhikari U, Singh DK (2017) Image annotation using deep learning: a review. In: 2017 international conference on intelligent computing and control (I2C2'17). <https://doi.org/10.1109/CAIPT.2017.8320684>
21. Betul AY, Zeynep K, Mehmet D, Galip A (2019) A visual similarity recommendation system using generative adversarial networks. In: 19 international conference on deep learning and machine learning in emerging applications (Deep-ML), 978-1-7281-2914-3/19/\$31.00 ©2019. IEEE. <https://doi.org/10.1109/Deep-ML.2019.00017>
22. Young P et al (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist* 2:67–78
23. Pak M, Kim S (2017) A review of deep learning in image recognition. In: IEEE-2017 4th international conference on computer applications and information processing technology (CAIPT). <https://doi.org/10.1109/CAIPT.2017.8320684>
24. Szegedy C et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
25. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

26. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representation
27. Deng L, Yu D (2014) Deep learning: methods and applications, now publishers
28. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp 1097–1105

# Yet Another Parallelism Within the “Hobby Time Training”



Milen Loukantchevsky 

**Abstract** The Hobby Time Training (*HTT*) concept is part of the Developer’s point of view (*DPV*) learning approach. It is described as “perceive the very solution to the problem as a game,” that is the decision-making process itself becomes a game. Moreover, is assumed the usage of conventional development environments. Thus, following the constructivism, making transition to a next level of gamification in comparison to the more primitive perception of gamification as “make the hard stuff fun.” The *HTT* assumes solving of small, apparently simple problems, which encapsulates deeply hidden potential that could be find out only during the problem solving. The problem decision is made in the student’s free time, at his discretion. In the area of computer architectures, when viewed as an interface between the high-level languages (*HLL*) and the raw machine, the bitwise algorithms are a good choice of problem set. Bitwise operations contain the sought-after hidden creative potential, mainly due to the limited support both at the high and low levels. Besides that, bitwise algorithms suppose usage of special techniques as word-level parallelism, unrolling loops and branch elimination. As an illustration of the hidden deep inner content of the bitwise problems, we are going to focus on the computing of bit set problem, or namely the Population Count problem.

**Keywords** Bitwise · Constructivism · Gamification · Hobby time · Population count ·  $\times 86/\times 64$  · Word-level parallelism

## 1 Introduction

In recent Twitter post is claimed “The more you know about hardware the more utterly horrifying software becomes.” [2]. It seems ditto like the claim “The more you know about arithmetic, the more horrifying are your decisions of nonlinear

---

M. Loukantchevsky (✉)

University of Ruse, 8 Studentska Str., POB 7017, Ruse, Bulgaria

e-mail: [mil@ieee.org](mailto:mil@ieee.org)

systems of partial differential equations” or “The more you know the nature, the uglier are products you create.”

Within the *Hobby Time Training (HTT)* we consider an example of quite the opposite. The *HTT* concept is part of the Developer’s point of view (*DPV*) learning approach. It is described as “perceive the very solution to the problem as a game,” that is the decision-making process itself becomes a game. Moreover, is assumed the usage of conventional development environments. Thus, making transition to a next level of gamification in comparison to the more primitive perception of gamification as “make the hard stuff fun” [5, 10, 13].

The *HTT* assumes solving of small, apparently simple problems, which encapsulates deeply hidden potential that could be find out only during the problem solving. The problem decision is made in the student’s free time, at his discretion [11].

In the area of computer architectures, when viewed as an interface between the high-level languages (*HLL*) and the raw machine, the bitwise algorithms are a good choice of problem set. Bitwise operations contain the sought-after hidden creative potential, mainly due to the limited support both at the high and low levels [1, 8, 14]. Besides that, bitwise algorithms suppose usage of special techniques as word-level parallelism, unrolling loops and branch elimination [6].

The scope of bitwise algorithms is very broad. As an illustration of the hidden deep inner content of the bitwise problems, we are going to focus on the computing of bit set problem, or namely the *Population Count* problem [1, 14].

## 2 Population Count

### 2.1 Problem Definition and Naïve Solution

The problem is apparently simple by definition: to count the number of bits set in a byte, word, doubleword. The solution should be encapsulated in a function with prototype `BYTE_pop_count(BYTE x)`. The same for 16- and 32-bits cases but replacing the type of *BYTE* with *WORD* and *DWORD*, respectively.

The naïve approach requires one iteration per bit, until no more bits are set (Fig. 1).

The naïve solution seems well enough only at first glance. Well, it is compact and is universal for different size of the input argument. However, it suffers from several significant drawbacks:

- The estimated performance is  $O(n)$ , where  $n$  designates the size of the input argument.
- Short base block, containing only four machine instructions (three for *count* variable modification and one for shift of the *src* argument) is blocking the instruction reordering mechanism (*Out-Of-Order* Execution, *OOO* in Intel’s notation).
- Control stalls will appear because of the branch instruction.



**Fig. 1** Naïve algorithm

```

50  BYTE __pop_count(BYTE src)
    {
        BYTE count = 0;

        while(src != 0)
        {
            count += src & 0x01;
            src >>= 1;
        }

60  return count;
    }

```

Let us turn for precision to the translated version of the naïve algorithm code [9]. In our case it is *Windows 32-bit platform* oriented, produced by the *Embarcadero C++ Builder® 11.2* development environment and its classic *BCC32* compiler [3]. Then we could calculate the actual number of instructions as in Eq. 1.

$$\text{InstrCount} = 3 + 2 + (k \times 6) + 1, \quad (1)$$

where  $k_{\min} = 0$  and  $k_{\max} = n$ . The worst case gives 54, 102 and 198 instructions for byte, word and doubleword, respectively.

And while the first drawback is obvious, the other two are related to the local hidden parallelism and require knowledge of the modern pipelined superscalar computer architectures vs traditional scalar ones.

## 2.2 Word-Level Parallelism: Basic Idea

There are two main levels of parallelism: *local hidden parallelism* (Instruction Level Parallelism) and *global structural parallelism* (Multicore Parallelism, Multiprocessor Parallelism and Multimachine Parallelism). But there is yet another level of parallelism, the so-called *Word-Level Parallelism* (WLP). It is a special case of data-level parallelism on scalar units—bytes, words, double words, etc. The WLP is direct consequence of parallel internal structure of the Arithmetic Logic Unit (ALU).

As already noted above, bitwise operations and among them the Population Count problem contain hidden creative potential. One manifestation of it is the opportunity to use the Word-Level Parallelism.

To improve our current solution, we will need two elaborations: to find alternative way of counting bit set and a base unit to apply this alternative at. As a base unit we try the nibble (Eq. 2).

$$X = x_3x_2x_1x_0 \quad (2)$$

The alternative way of counting bit set in a nibble uses three shifts of the nibble to the right with appropriate floor and three subtractions [14, 15].

$$\text{PopCount} = X - \left\lfloor \frac{X}{2^1} \right\rfloor - \left\lfloor \frac{X}{2^2} \right\rfloor - \left\lfloor \frac{X}{2^3} \right\rfloor \quad (3)$$

From Eq. 4 one can follow the proof that this yields the desired sum.

$$\begin{aligned} X - \frac{X}{2^1} - \frac{X}{2^2} - \frac{X}{2^3} &= (x_3 \times 2^3 + x_2 \times 2^2 + x_1 \times 2^1 + x_0 \times 2^0) \\ &\quad - (x_3 \times 2^2 + x_2 \times 2^1 + x_1 \times 2^0) \\ &\quad - (x_3 \times 2^1 + x_2 \times 2^0) - (x_3 \times 2^0) \\ &= x_3 \times (2^3 - 2^2 - 2^1 - 2^0) + x_2 \times (2^2 - 2^1 - 2^0) \\ &\quad + x_1 \times (2^1 - 2^0) + x_0 \times (2^0) \\ &= x_3 + x_2 + x_1 + x_0 \end{aligned} \quad (4)$$

To reduce the above equation to the last its expression is used the next property (Eq. 5)

$$2^n - 2^{n-1} - \dots - 2^1 - 2^0 = 1. \quad (5)$$

What remains is to implicitly decompose the input argument into nibbles and utilize the *WLP* at maximum.

### 2.3 Developing the Solution

We are going to follow the evolution of the solution starting with argument of size nibble and ending with doubleword.

The direct application of the basic idea from Eq. 3 is shown at Fig. 2, no *WLP* here yet.

From the translated version [9] we calculate the machine instruction number needed as 27 (Eq. 6). If we replace in Eq. 1  $k$  with 4, we will get 30 machine instructions for the worst case of the naïve algorithm for a nibble.

$$\text{InstrCount} = 3 + 3 + 2 + (3 \times 6) + 1 = 27. \quad (6)$$

When working with a byte is triggered the first level of *WLP* and the operations of Eq. 3 are applied for the two nibbles in parallel (Fig. 3).

There is one additional step of summing the two sums.

From the translated version [9] we calculate the machine instruction number needed as 35 (Eq. 7). The final summing *C* operator requires additional 6 instructions.

```

40  BYTE __pop_count(BYTE src)
    {
        BYTE x = src & 0xF;
        BYTE t = x;

        // Count the bits in the one 4-bits field
        t = t >> 1;
        x = x - t;

        t = t >> 1;
        x = x - t;

50    t = t >> 1;
        x = x - t;

        return x;
    }

```

**Fig. 2** Basic algorithm for a nibble

```

50  BYTE __pop_count(BYTE src)
    {
        BYTE x = src;
        BYTE t = x;

        // [1] Count the bits in each of the two 4-bits fields in parallel
        t = (t >> 1) & 0x77;
        x = x - t;

        //
        t = (t >> 1) & 0x33;
60    x = x - t;

        //
        t = (t >> 1) & 0x11;
        x = x - t;

        // [2] Sum 1 pair of 4-bits sums
        x = (x + (x >> 4)) & 0x0F;

        return x;
    }

```

**Fig. 3** WLP algorithm, 8-bit case

$$\text{InstrCount} = 3 + 4 + ((3 \times 7) + (1 \times 6)) + 1 = 35 \quad (7)$$

Working with a word we reach next level of WLP—Eq. 3 is applied to four nibbles in parallel.

The summing of the four partial results is in two steps: at first two pairs of 4-bit sums are added up in parallel, then a pair of 8-bit sums is added up to form the result (Fig. 4).

And again, from the translated version [9] we calculate the machine instruction number needed as 36 (Eq. 8). The two final summing *C* operators requires additional 10 instructions.

```

* WORD __pop_count(WORD src)
* {
*     WORD x = src;
80  WORD t = x;
*
*     // [1] Count the bits in each of the four 4-bits fields in parallel
*     t = (t >> 1) & 0x7777;
*     x = x - t;
*     //
*     t = (t >> 1) & 0x3333;
*     x = x - t;
*     //
90  t = (t >> 1) & 0x1111;
*     x = x - t;
*     // [2] Sum the 2 pairs of 4-bits sums
*     x = (x + (x >> 4)) & 0x0F0F;
*     // [3] Sum the 1 pair of 8-bits sums
*     x = (x + (x >> 8)) & 0x00FF;
*
*     return x;
* }

```

Fig. 4 WLP algorithm, 16-bit case

$$\text{InstrCount} = 3 + 4 + ((3 \times 6) + (2 \times 5)) + 1 = 36. \quad (8)$$

Working with a doubleword we reach another level of WLP—Eq. 3 is applied to eight nibbles in parallel.

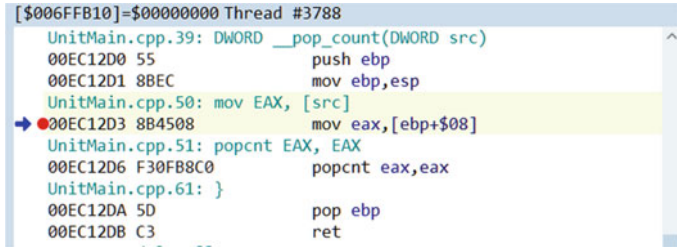
The summing of the eight partial results is in three steps in this case: at first four pairs of 4-bit sums are added up in parallel, then two pairs of 8-bit sums and last a pair of 16-bit sums is added up to form the result (Fig. 5).

```

* DWORD __pop_count(DWORD src)
* {
*     DWORD x = src;
*     DWORD t = x;
80
*     // [1] Count the bits in each of the four 4-bits fields in parallel
*     t = (t >> 1) & 0x77777777;
*     x = x - t;
*     //
*     t = (t >> 1) & 0x33333333;
*     x = x - t;
*     //
*     t = (t >> 1) & 0x11111111;
*     x = x - t;
90
*     // [2] Sum the 4 pairs of 4-bits sums
*     x = (x + (x >> 4)) & 0x0F0F0F0F;
*     // [3] Sum the 2 pairs of 8-bits sums
*     x = (x + (x >> 8)) & 0x00FF00FF;
*     // [4] Sum the 1 pair of 16-bits sums
*     x = (x + (x >> 16)) & 0x0000FFFF;
*
*     return x;
* }

```

Fig. 5 WLP algorithm, 32-bit case



```

[$006FFB10]=00000000 Thread #3788
UnitMain.cpp.39: DWORD __pop_count(DWORD src)
00EC12D0 55          push ebp
00EC12D1 8BEC        mov ebp,esp
UnitMain.cpp.50: mov EAX, [src]
→ 00EC12D3 8B4508      mov eax,[ebp+$08]
UnitMain.cpp.51: popcnt EAX, EAX
00EC12D6 F30FB8C0    popcnt eax,eax
UnitMain.cpp.61: }
00EC12DA 5D          pop ebp
00EC12DB C3          ret

```

**Fig. 6** Hardware supported (*HS*) algorithm, 32-bit case ( $\times 86$  translated code)

Again, from the translated version [9] we calculate the machine instruction number needed as 41 (Eq. 9). The three final summing *C* operators requires additional 15 instructions.

$$\text{InstrCount} = 3 + 4 + ((3 \times 6) + (3 \times 5)) + 1 = 41. \quad (9)$$

So, the set goal has been reached and we have got effective solution. However, we cannot help but ask ourselves what if there is hardware support of Population Count problem. And, surprisingly, there is. It is the machine instruction *POPCNT*, part of the *SSE4.2*.

As could be expected, the hardware supported (*HS*) algorithm of Population Count (Fig. 6) shows drastic reduction of the complexity to 4 machine instructions with comparison to the 198 instructions of the Naïve algorithm and even to the 41 instructions of the *WLP* algorithm. Also, a reference to the  $\times 86$  optimization guide shows a low latency of the *POPCNT* instruction, close to the conventional arithmetic logic instructions [4, 7, 12].

Probably, this is the best response to the Twitter post at the very beginning, if one should know the underlying machine or not, and if this knowledge leads to clear or to terrible construction.

As a conclusion, at Fig. 7 is presented the complexity estimation (in number of machine instructions) for all the three algorithms discussed. The *WLP* is times better than the Naïve algorithm. But the hardware supported (*HS*) version is beyond all competition.

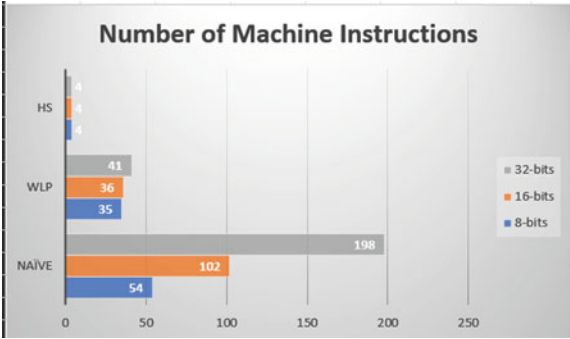


Fig. 7 Complexity estimation of the algorithms discussed

3 Conclusion

The professional software developer perceives the very solution of the problem as a game. To reach this highest level of gamification is proposed the concept of the *Hobby Time Training (HTT)* concept. The *HTT* is part of the *DPV* learning approach and assumes solving of small, apparently simple problems, which encapsulates deeply hidden potential.

In the area of computer architectures, when viewed as an interface between the high-level languages (*HLL*) and the raw machine, the bitwise algorithms are a good choice of problem set. Bitwise operations contain the sought-after hidden creative potential, mainly due to the limited support both at the high and low levels. Besides that, bitwise algorithms suppose usage of special techniques as word-level parallelism, unrolling loops and branch elimination.

As an illustration of the hidden deep inner content of the bitwise problems, we focused on the computing of bit set problem, or namely the *Population Count* problem. From the Naïve algorithm through application of the Word-Level Parallelism to the hardware supported solution.

From one side, you cannot be innovative in the big if you never even tried to be innovative in the small. And from another, we encounter the phenomenon of what is considered as obvious as a solution eventually turns out to be the most inappropriate one.

References

1. Anderson S (2022) Bit twiddling hacks. <http://graphics.stanford.edu/~seander/bithacks.html#ParityParallel>. Accessed 17 Sept 2022
2. Booch G (2022) Twitter post. [https://twitter.com/Grady\\_Booch/status/1573133592938659845?s=20&t=b51BkJQBshqNxx7mMYylzg](https://twitter.com/Grady_Booch/status/1573133592938659845?s=20&t=b51BkJQBshqNxx7mMYylzg). Accessed 28 Sept 2022

3. Embarcadero (2022) RAD studio Docwiki: C++ compilers. [https://docwiki.embarcadero.com/RADStudio/Alexandria/en/C%2B%2B\\_Compilers](https://docwiki.embarcadero.com/RADStudio/Alexandria/en/C%2B%2B_Compilers). Accessed 23 Sept 2022
4. Fog A (2022) Optimization manuals: lists of instruction latencies, throughputs and micro-operation breakdowns for Intel, AMD, and VIA CPUs. <https://www.agner.org/optimize/#manuals>. Accessed 28 Sept 2022
5. Hadjerrouit S (2005) Constructivism as guiding philosophy for software engineering education. ACM SIGCSE Bull 37(4). <https://doi.org/10.1145/1113847.1113875>
6. Intel (2022) Intel® 64 and IA-32 architectures optimization reference manual. Order Number: 248966-045
7. Intel (2022) Intel® 64 and IA-32 architectures software developer’s manual. Order number: 325462-077US
8. Knuth D (2009) The art of computer programming, vol 4, Fascicle 1: bitwise tricks & techniques; binary decision diagrams, 1st edn. Addison Wesley Professional. ISBN 978-0321580504
9. Loukantchevsky M (2022) GitHub repository: population count translated code. <https://github.com/milphaser/WLP.POPCNT>. Accessed 8 Oct 2022
10. Loukantchevsky M (2021) Solving classical problem in new context as constructive model of training: active memory array of concurrent processes concept. In: CompSysTech ’21. ACM, New York, NY, USA, pp 191–195. <https://doi.org/10.1145/3472410.3472430>
11. Loukantchevsky M (2022) The hobby time training approach. In: Proceedings of the University of Ruse—2022, vol 61, Book 3.2. ISSN 2603-4123. Preprint: [https://www.researchgate.net/publication/365027696\\_THE\\_HOBBY\\_TIME\\_TRAINING\\_APPROACH](https://www.researchgate.net/publication/365027696_THE_HOBBY_TIME_TRAINING_APPROACH)
12. Microsoft technical documentation (2022) x86 Instructions. <https://docs.microsoft.com/en-us/windows-hardware/drivers/debugger/x86-instructions>. Accessed 17 Sept 17
13. Severance C (2012) The art of teaching computer science: Niklaus Wirth. IEEE Comput 45(7):8–10. <https://doi.org/10.1109/MC.2012.245>
14. Warren H (2010) Hacker’s delight, 2nd edn. Addison-Wesley. ISBN 0-201-91465-4
15. Wolfram M (2022) Floor function. <https://mathworld.wolfram.com/FloorFunction.html>. Accessed 28 Sept 2022

# An Improved Apriori Algorithm for Interestingness of Association Rules: A Case Study on the Mushroom Dataset



Huynh Anh Duy, Bui Trong Vinh, and Phan Duy Hung

**Abstract** Currently, data mining is a field that is growing very strongly thanks to the trend of digital transformation globally. An important branch of data mining is Mining Association Rule, which tries to extract useful information from large amounts of data through finding frequent itemsets and association rules. The most popular and well-known algorithms of this domain is the Apriori algorithm. However, this algorithm still has weaknesses in terms of runtime as well as the quality of the generated rules is not high. There have been many studies aimed at reducing the running time of Apriori algorithm, but the quality of the generated rules has not been improved much. This drawback leads to the problem that users cannot use these rules or in worse case, get misleading information. This work focuses on proposing a new approach to improve the quality of association rules derived from the Apriori algorithm on Mushroom dataset. Results show that the new approach helps to improve the quality of rules and reduce the run time compared to the original Apriori.

**Keywords** Apriori · Association rules · Interestingness · Correlation · Kulc · Cosine · Mushroom dataset

## 1 Introduction

In the current rapidly developing technology context, data analysis has emerged as one of the key industries. The amount of data generated in the world is increasing dramatically every second [1]. This raw data becomes valuable only when we can

---

H. A. Duy · P. D. Hung (✉)  
FPT University, Hanoi, Vietnam  
e-mail: [hungpd2@fe.edu.vn](mailto:hungpd2@fe.edu.vn)

H. A. Duy  
e-mail: [duyhahe153764@fpt.edu.vn](mailto:duyhahe153764@fpt.edu.vn)

B. T. Vinh  
Hanoi Procuratorate University, Hanoi, Vietnam  
e-mail: [vinhbt@tks.edu.vn](mailto:vinhbt@tks.edu.vn)



mine and extract valuable information from it. This is a challenge because of the huge volume of data as well as the complexity of data mining techniques. Specifically, one of the business needs through transforming raw data into knowledge is finding the most frequent patterns in large datasets, then creating a generalizable and interpretable picture of reality. A branch of this problem is Market Basket Analysis or in other words, Mining Association Rule problem [2], which require the identification of the most frequent patterns along with association rules in transactional datasets.

The most known and simplest one for Mining Association Rule problem is Apriori algorithm which was first introduced in 1994 [3]. This algorithm allows us to extract important rules about customers' buying habits from the store's transactions data, thereby making more effective plans and strategies to develop the business. However, the Apriori algorithm still has a weakness in terms of computational performance because of its bottom-up and breadth-first approach. Besides that, the quality of the rules generated by the algorithm is not really appreciated. The quality of rule is evaluated through some interestingness measures including the confidence of rules and the correlation between itemsets of rule. According to the research results given in [4], the traditional approach of the Apriori algorithm has not considered the factor of correlation between itemsets in the rule, leading to some rules having low interestingness. In the worst case, these rules can lead to misleading information which in turn cause serious consequences in any area of its application. This study proposes a new approach by adding a new constraint to the joining step of the Apriori algorithm, the main goal is to eliminate rules with low correlation and reduce the running time of the algorithm. Two null-invariant measures including cosine and Kulc will be used to evaluate the correlation of the rules. Experimental results on the Mushroom dataset show that the proposed algorithm not only helps to retain the high percentage of the top highest correlation rules, but also significantly reduces the running time compared to the original Apriori algorithm.

## 2 Related Works

Many studies have been conducted in the field of Association Rule Mining. Some of these works are described below:

Paper [5] proposes an approach to avoid large computation when finding association rules, specifically, some items will be omitted and only some best sold items will be kept. The author applies this new approach to the French Retail Store dataset and Bakery Shop dataset, two algorithms used are Apriori and FP-Growth. The results show that the method of using only the best-selling products significantly reduces the running time for two algorithms. Besides that, whether or not the new method is applied, the execution time of the FP-Growth algorithm is still better than the Apriori algorithm.

The author in [6] improves the performance of Mining Association Rule with Apriori by allowing the algorithm to automatically calculate the appropriate min support value for each level. In other words, users will not need to manually define a

minimum support value, instead the min support value will be automatically calculated based on each different dataset and different level. Experimental result on some public datasets from UCI machine learning repository like: Mushroom, Flare1, Flare2... show that this change in the min support value reduces consuming time, reduces memory consumption as well as reduces high percentage of rules when compared to Apriori algorithm.

Another idea has been presented in paper [7], the author proposes to improve the Apriori algorithm by changing the way transactions are scanned for counting the number of each candidate  $k$ -itemsets. Instead of scanning through all transactions like the original Apriori algorithm, we will find the item with the smallest support value  $m$  in that itemset, and then just traverse through  $m$  transactions. This approach helps to reduce the number of transactions to be scanned, thereby reducing time-consuming by 67.38% compared to the original approach.

In [8], author points out the main problem of the Apriori algorithm is that users have to define the min support value themselves while they not having enough necessary information. From this motivation the author proposed a new version of the Apriori algorithm where the support threshold value is calculated using some mathematical function. More specifically, author suggest to perform database scanning only once, then place the itemsets and their support values in the respective addresses based on the analysis on them. Experiments were performed on 3 datasets: Chess, Mushroom and Tumor. From experimental results, the approach is proved to be better not only about run time aspect but also about memory usages due to properties of only scanning the database once.

Another improvement on this area was published in [9], the proposed algorithm from the author improves the Apriori algorithm through three aspects: reduce the number of time scanning database, limit the size of the candidate itemsets and improve the speed of both joining and pruning processes. Specifically, the method removes all itemsets with support less than  $k - 1$  in the set  $L_{k-1}$ , then construct the set  $C_k$  as usual. This strategy helps to reduce the amount of candidate items and decrease the execution time of the algorithm. Experimental results on the Mushroom dataset present that the new approach improve performance significantly through reducing the time execution by over 98% on average.

From all above studies, it can be seen that among those previous studies on the Apriori algorithm there are very few studies that focus on the interestingness of the generated association rules, instead they tend to focus on improving the computational performance. Therefore, a research on filtering out low-interestingness rules and make the algorithm more useful in real life problem is necessary.

### 3 Methods

#### 3.1 Association Rules

The general concept of association rule is presented as follows. Let  $I = \{i_1, i_2, \dots, i_m\}$  be the set of items, a set of items  $T$  satisfy  $T \subseteq I$  is called a transaction.  $D$  is the set containing all transactions in the dataset. The *association rule* consists of two parts: the antecedent part  $X$  and the consequent part  $Y$ , which were illustrated through formula as:

$$X \Rightarrow Y \text{ in which } X \subseteq I, Y \subseteq I, X \uparrow Y = \emptyset \quad (1)$$

An itemset is said to be a *frequent itemset* if its support value is greater than or equal to the given min support threshold. The support of rule  $X \Rightarrow Y$  is the total number of transactions in  $D$  contain  $X \cup Y$ . The confidence of rule  $X \Rightarrow Y$  is defined as the ratio of transactions containing  $X \cup Y$  out of the total number of transactions containing  $X$ . These definitions are expressed in detail by the following formulas:

(a) The support value for rule  $X \Rightarrow Y$ .

$$\text{Support}(X \Rightarrow Y) = \frac{|X \cup Y|}{|D|} \quad (2)$$

where  $|D|$  is the number of transactions in the dataset.

(b) The confidence value for rule  $X \Rightarrow Y$ .

$$\text{Confidence}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \quad (3)$$

The min support threshold and min confidence threshold values will be user defined. A rule is considered *strong* when support value and confidence value are either greater than or equal to the min support value and min confidence value.

#### 3.2 Apriori Algorithm

Apriori is a well-known algorithm that solves the mining frequent pattern problem. The main idea of Apriori is getting  $(k + 1)$ -itemsets from  $k$ -itemsets, which is called level-wise search strategy. Specifically, the algorithm first lists all the items with their support value, then keeps the items whose support meets the min support value. This set is called frequent 1-itemsets. Similarly, by traversing the frequent 1-itemset, we find the frequent 2-itemset set. The algorithm is repeated until there is no itemsets

that satisfy the given threshold. In general, the Apriori algorithm can be divided into the two following steps:

- *Joining step*: Finding set of candidate  $k$ -itemsets  $C_k$  through joining  $(L_{k-1})$  with itself.
- *Pruning step*: Any subset of a frequent itemset also is a frequent itemset.

*Pseudo code of APRIORI:*

```

 $C_k$ : Candidate itemset of size  $k$ 
 $L_k$ : Frequent itemset of size  $k$ 
 $L_1 = \{\text{frequent items of size } 1\}$ ;
for ( $k = 1$ ;  $L_k \neq \emptyset$ ;  $k++$ ) do begin
     $C_{k+1}$  = candidates generated from  $L_k$ ;
    for each transaction  $t$  in database do
        increment the count of all candidates in  $C_{k+1}$  that are contained in  $t$ 
     $L_{k+1}$  = candidates in  $C_{k+1}$  with min_support
end
end
return  $\cup_k L_k$ ;

```

### 3.3 Correlation Analysis

Association rules mining problem has received a lot of attention from researchers since its first publication. One of the big questions is whether the confidence-support framework has created the rules that users really need. A lot of research had been done on this domain and showing that using the confidence-support framework alone is not enough to create a really good set of rules, the paper of Sergey Brin in 1997 [10] is an example. Correlation measure has been added to address this issue. Specifically, the quality of the association rule is now evaluated through support, confidence and correlation as follows:

$$A \Rightarrow B(\text{supp}, \text{conf}, \text{corr}) \quad (4)$$

In general, the correlation measure represents the correlation between the antecedent part (itemset  $A$ ) and the consequent part (itemset  $B$ ). There are many measures used to calculate the correlation between itemsets in the mining association rule problem. Each measure has its own characteristic as described in [4]. However, based on the results and analysis from [11], *Kulczynski* measure (Kulc) is a null-invariant measure that gives relatively most reasonable results, especially when dealing with difficult cases of data including null-transactions and high imbalances. Therefore, in this paper, we prefer to use Kulc and come with another null-invariant

measure which is cosine to evaluate the correlation of the set of rules. Following are the definitions of these two measures:

- Consider itemset A and itemset B, the Kulc measure of A and B is calculated as:

$$\text{Kulc}(A, B) = \frac{1}{2}(P(A|B) + P(B|A)) \quad (5)$$

Briefly, this measure can be expressed as an average confidence measures of rule  $A \Rightarrow B$  and rule  $B \Rightarrow A$ .

- Consider itemset A and itemset B, the cosine measure of A and B is calculated as:

$$\begin{aligned} \text{cosine}(A, B) &= \frac{P(A \cup B)}{\sqrt{P(A)P(B)}} = \frac{\text{sup}(A \cup B)}{\sqrt{\text{sup}(A) \text{sup}(B)}} \\ &= \sqrt{P(A|B)P(B|A)} \end{aligned} \quad (6)$$

From the formula it can be seen that these two measures are not affected by the number of transactions that do not contain both considered itemsets (null-transaction). The range of values of the Kulc measure and the cosine measure are both between 0 and 1. Besides, we expect a high measure value because it means that the correlation between the itemsets will also be high.

### 3.4 Proposed Approach

As mentioned above, correlation is an indispensable factor when solving the Association Rules problem. Especially for the case of the Mushroom dataset, which is used to classify the edible mushroom and poisonous mushroom. The set of generated rules including those rules with low correlation or negative correlation will lead to extremely serious consequences.

Call  $\text{support}(A \cup B) = sab$ ,  $\text{support}(A) = sa$ ,  $\text{support}(B) = sb$ . From (1), we have:

$$\begin{aligned} \text{Kulc}(A, B) &= \frac{1}{2}(P(A|B) + P(B|A)) \\ &= \frac{1}{2} \left( \frac{\text{support}(A \cup B)}{\text{support}(A)} + \frac{\text{support}(A \cup B)}{\text{support}(B)} \right) \\ &= \frac{1}{2} \left( \frac{sab}{sa} + \frac{sab}{sb} \right) \end{aligned} \quad (7)$$

We need  $\text{Kulc}(A, B) > 0.5$  for positive correlation between 2 itemsets

$$\Rightarrow \frac{sab}{sa} + \frac{sab}{sb} > 1 \quad (8)$$

We choose those rules with  $\frac{sab}{sa} > 0.5$  and  $\frac{sab}{sb} > 0.5$  (\*) because this condition ensures that condition (3) is satisfied.

Since  $sab \leq \min(sa, sb)$ , if  $\min(sa, sb) \leq 0.5 \max(sa, sb)$  then (\*) cannot be obtained (\*\*).

Suppose  $P'$  is a superset of  $P$ , then  $\text{support}(P') \leq \text{support}(P)$  (\*\*\*).

From (\*\*), (\*\*\*) it can be seen that the joining between some itemsets is redundant because it does not guarantee that the generated rules have a high correlation. Therefore, find out a way to limit this redundant joining will help eliminate rules with low correlation and also help reduce computation time.

From this motivation, we propose a new approach by adding a constraint with a new parameter  $d$  to the *joining step* of the original Apriori algorithm. This approach helps to remove those rules with low correlation and keep those rules with high correlation, thereby improving the quality and reliability of the generated set of rules. The pseudo code of this approach is presented as follow:

```

 $C_k$ : Candidate itemset of size  $k$ 
 $L_k$ : Frequent itemset of size  $k$ 
 $L_1 = \{\text{frequent items}\};$ 
for ( $k = 1$ ;  $L_k \neq \emptyset$ ;  $k++$ ) do begin
     $C_{k+1}$  = candidates generated from  $L_k$  with the following constraint:
    Considering the general case that itemset  $A$ 
    and  $B$  belong to  $L_k$ , itemset  $A$  and  $B$  are joined together to form  $C_{k+1}$  if and
    only if:
         $\min(\text{support}(A), \text{support}(B)) >$ 
         $d \cdot \max(\text{support}(A), \text{support}(B)), d \in [0.5, 1)$ 
    for each transaction  $t$  in database do
        increment the count of all candidates in  $C_{k+1}$  that are contained in  $t$ 
         $L_{k+1}$  = candidates in  $C_{k+1}$  with min_support
    end
end
return  $\cup_k L_k$ ;

```

## 4 Experiments and Analyze

### 4.1 Data Preparation and Experimental Environment

This work uses a benchmark dataset for association rule mining problem which is the Mushroom dataset. This dataset can be downloaded through UCI Machine Learning Repository [12] or Frequent Itemset Mining Dataset Repository [13]. The Mushroom dataset contains 8124 records, these records represent 23 mushroom types of the

family Agaricus and Lepiota. The association rules will help to determine whether the attribute combinations belong to edible mushroom or poisonous mushroom. In other words, these rules provide essential information for humans in classifying poisonous and edible mushrooms.

All experiments are carried out in the environment with the following configuration: Intel(R) Core™ I3, 2.40 GHz, memory capacity: 4 GB, operating system: Microsoft Windows 10, programming language: Python.

We carry out the experiments by fixing value of *min support* and *min confidence* to 0.4, then varying the value of parameter  $d$  from 0.5 to 0.8. Besides the original measure of *confidence*, we will also use the two measures mentioned in Sect. 3 which are cosine and Kulc to evaluate the correlation of the generated rules.

## 4.2 Evaluated Through the Entire Set of Generated Rules

First, measures calculated based on the whole generated association rules are considered. Specifically, those measures used here are: number of frequent patterns, number of rules, average confidence of the rules, average cosine of the rules, average Kulc of the rules and running time. Experimental results are shown in Tables 1 and 2, in which Table 1 presents the results when using Apriori algorithm, Table 2 presents the results when using proposed approach.

As we can see, the proposed algorithm helps to reduce a significant number of frequent patterns and rules compared to the original Apriori algorithm. More specifically, from Table 2, it can be observed that the higher the value of parameter  $d$ , the lower the number of frequent patterns and rules. This trend is completely understandable because as shown in Sect. 4, increasing the parameter  $d$  will cause the constraint to be set tighter. Along with the reducing number of rules, those measures including average *confidence*, average cosine, average Kulc has also been improved, which shows that in terms of the entire set of rules, the quality of the rules has been improved. Another worth noting point is that limiting the number of rules as well as frequent patterns also helps to reduce the running time as we can see from Table 2.

## 4.3 Evaluated Through Top Highest Correlation Rules

The improvement on averaging of the measures in Sect. 5 only gives us an overview of the quality of the generated rules, which is not really specific and convincing. Therefore, here the proposed algorithm will be compared with the Apriori algorithm through the aspect of highest correlation rules. The higher the probability of preserving the top highest correlation rules, the lower the loss of information caused by the proposed algorithm. Table 3 shows the percentage of retained rules corresponding to each of the top- $k$  highest correlation rules (here correlation is calculated

**Table 1** Result on Mushroom dataset with Apriori algorithm

Min support	Min confidence	No. of frequent patterns	No. of rules	Average confident	Average Kulc	Average cosine	Runtime (second)
0.4	0.4	565	7020	0.7397658330937987	0.7397658330937991	0.7115090781055389	218.45919680595398



**Table 2** Result on mushroom dataset with proposed algorithm

Min support	Min confidence	$d$	No. of frequent patterns	No. of rules	Average confident	Average Kulc	Average cosine	Runtime (second)
0.4	0.4	0.5	494	6482	0.7419236341880902	0.7419236341880878	0.7154691641176625	188.14486646652222
0.4	0.4	0.6	303	4088	0.7503531850311842	0.7503531850311924	0.7283616264893529	110.15700602531433
0.4	0.4	0.7	86	398	0.8249568999015524	0.8249568999015531	0.816524263882622	12.543394565582275
0.4	0.4	0.8	61	216	0.8907946903669667	0.8880973376000193	0.8907946903669652	7.10567831993103

**Table 3** Ratio of retained rules in each top- $k$  highest correlation rules

$d$	No. of rules	Top 200 (%)	Top 300 (%)	Top 3000 (%)	Top 4000 (%)
0.5	6482	100	100	98.4	96
0.6	4088	97.8	98.7	71	62.11
0.7	398	97.8	70.68	9.91	8.27
0.8	216	95.8	61.33	6.75	5.2

based on the Kulc value). The value of  $k$  chosen is 200, 300, 3000, 4000, respectively, which is consistent with the amount of rules created by the proposed approach and also ensures to cover the range of typical  $k$  used mostly in many previous studies [14–16].

It is easy to see that the larger the value of  $d$ , the smaller the number of rules created and also the lower the percentage of retaining top correlation rules. However, based on the correlation between the number of rules created and the value of  $k$ , the retaining rate is still within an acceptable level. For example, in the case of  $d = 0.6$ , the number of rules created is 4088. It is not surprising that the retaining rate for the top 200 and top 300 is quite high because the number of rules created is relatively large compared to  $k$ , however the retaining rates for the top 3000 and top 4000 are 71% and 62.11% respectively, which are still acceptable ratios.

Moreover, the results also show that the proposed algorithm has the ability to preserve the top-rules (top- $k$  with small  $k$ ) very well. Specifically, top 200 is almost completely preserved with the lowest retaining rate being 95.8% when  $d = 0.8$ , a very high value of  $d$ . This is very necessary because in real life problems users often only need a sufficient number of best rules, not too much. Additionally, in the case of  $d = 0.5$ , almost all the top- $k$  are preserved with the smallest retaining rate of 96% for the top 4000. This shows that with the appropriate selection for value of  $d$ , the proposed algorithm can help us get the best rules in a short time while preserving most of the information obtained by the original Apriori algorithm.

## 5 Conclusion and Perspectives

This work presents a new approach based on setting new constraints at the joining step of the Apriori algorithm, the main contribution is to filter those rules with high correlation and improve the running time of the original algorithm. Experimental results on Mushroom dataset show that this approach not only minimizes the running time by eliminating those rules with low correlation, but also ensures the retaining rate of top-rules at an acceptable level.

In future study, this method can be further developed through combining with existing optimization techniques to improve the performance and increase the retaining rate of the top-rules. The applicability of this method in related problems such as high utility itemset mining will also be an interesting area to learn

in further research. The paper can be a good reference for many data mining and recommendation problems [17–20].

## References

1. Luna JM, Fournier-Viger P, Ventura S (2019) Frequent itemset mining: a 25 years review. *Wiley Interdisc Rev Data Min Knowl Discov*, 9(6)
2. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD international conference on management of data (ICDM '93)*, pp 207–216
3. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: *Proceedings of the 20th international conference on very large data bases*, vol 1215, pp 487–499
4. Tianyi W, Yuguo C, Jiawei H (2007) Association mining in large databases: a re-examination of its measures. In: *Proceedings of the international conference on principles and practice of knowledge discovery in databases (PKDD'07)*
5. Maliha H, Sattar AHMS, Paul MK (2019) Market basket analysis using apriori and FP growth algorithm. In: *Proceedings of the 22nd international conference on computer and information technology (ICCIT)*
6. Azzeddine D, Youssef B, Taoufiq G (2018) Using multiple minimum support to auto-adjust the threshold of support in apriori algorithm. In: *SoCPaR 2017, AISC 737*, pp 111–119
7. Mohammed A-M, Bassam A (2014) An improved apriori algorithm for association rules. *Int J Natural Lang Comput* 3(1):21–29
8. Sachin S, Shaveta B (2018) New approach for generating frequent item sets without using minimum support threshold. *Int J Comput Sci Inform*, 16(9):1–8
9. Xiuli Y (2017) An improved apriori algorithm for mining association rules. In: *AIP conference proceedings* 1820, 080005
10. Brin S, Motwani R, Ullman J, Tsur S (1997) Dynamic Itemset counting and implication rules for market basket data. *SIGMOD-97*, pp 255–264
11. Han J, Kamber M, Pei J (2011) *Data mining concepts and techniques* third edition. In: *The Morgan Kaufmann series in data management systems*
12. UCI Machine Learning Repository (2022) <https://archive.ics.uci.edu/ml/datasets/mushroom>. Accessed 10 Oct 2022
13. Frequent Itemset Mining Dataset Repository (2022) <http://fimi.uantwerpen.be/data/>. Accessed 10 Oct 2022
14. Jiawei H, Jianyong W, Ying L, Tzvetkov P (2002) Mining Top-K frequent closed patterns without minimum support. In: *Proceedings of the IEEE international conference on data mining*
15. Fournier-Viger P, Wu CW, Tseng VS (2012) Mining Top-K association rules. In: Kosseim L, Inkpen D (eds) *Advances in artificial intelligence. Canadian AI 2012. Lecture Notes in Computer Science*, vol 7310. Springer, Berlin, Heidelberg
16. Ahmed A, Mona N, Shaimaa S (2014) A comparative study of association rules algorithms on large databases. *Egypt Comput Sci J* 38(3)
17. Hung PD, Huynh LD (2019) E-commerce recommendation system using Mahout. In: *Proceedings of the IEEE 4th International conference on computer and communication systems (ICCCS)*, pp 86–90. <https://doi.org/10.1109/CCOMS.2019.8821663>
18. Phan DH, Do QD (2020) Analysing effects of customer clustering for customer's account balance forecasting. In: Nguyen NT, Hoang BH, Huynh CP, Hwang D, Trawiński B, Vossen G (eds) *Computational collective intelligence. ICCCI 2020. Lecture notes in computer science*, vol 12496. Springer, Cham (2020)

19. Quan VH, Hung PD (2022) Heterogeneous neural collaborative filtering for a business recommender system. In: Al-Emran M, Al-Sharafi MA, Al-Kabi MN, Shaalan K (eds) Proceedings of international conference on emerging technologies and intelligent systems. ICETIS 2021. Lecture notes in networks and systems, vol 322. Springer, Cham
20. Hung PD, Su NT, Diep VT (2019) Surface classification of damaged concrete using deep convolutional neural network. *Pattern Recogn Image Anal* 29:676–687

# Hybrid Network Anomaly Detection Based on Weighted Aggregation Using Endpoint Parameters



L. Y. Dobkacz , S. A. Sakulin , A. N. Alfimtsev , and Y. A. Kalgin 

**Abstract** The article considers a hybrid approach to anomaly detection taking into account the parameters of network endpoints. The transformation of the network endpoint parameters to the format of the session parameters is implemented using regrouping and dimensionality reduction. To detect anomalies, ensembles of classifiers are used, with the inputs receiving session parameters and network nodes. Ensembles use three methods of machine learning: logistic regression, stochastic gradient descent and decision trees. The fusing of classification results is based on weighted aggregation with anxiety threshold as a weighting factor. The anxiety threshold allows to regulate the influence of data from endpoint parameters on the classification by session format. The optimal value of the anxiety threshold was found experimentally to detect the most of anomalies on the testing set. An experimental evaluation of the developed approach proved its advantages in comparison with the detection of anomalies without taking into account the endpoint parameters.

**Keywords** Machine learning · Signature approach · Weighted aggregation · CICIDS2017 dataset · Corporate networks · Hybrid system · Ensemble classification · Traffic analyzer · Wazuh tool · Multi-agent systems

---

L. Y. Dobkacz (✉) · S. A. Sakulin · A. N. Alfimtsev · Y. A. Kalgin  
Bauman Moscow State Technical University, Moscow, Russia  
e-mail: [dobkachleo@mail.ru](mailto:dobkachleo@mail.ru)

S. A. Sakulin  
e-mail: [sakulin@bmstu.ru](mailto:sakulin@bmstu.ru)

A. N. Alfimtsev  
e-mail: [alfim@bmstu.ru](mailto:alfim@bmstu.ru)

Y. A. Kalgin  
e-mail: [kalgin@bmstu.ru](mailto:kalgin@bmstu.ru)

## 1 Introduction

In recent years, the bulk transfer of enterprises employees to a remote work from home, as well as informational confrontation, have led to the emergence of many problems related to data security in corporate networks [1]. Similar problems arise in networks serving multi-agent robotic systems [2]. Attacks on networks such as phishing, distribution of malicious code, DDoS attacks, threats of physical injection of malicious agents into the network are the most significant of these problems [1, 2].

Anomaly detection systems are traditionally based on the analysis of signatures corresponding to known anomalies, as well as on the basis of machine learning [3, 4]. The occurred event is assigned to a predetermined class according to the detected anomalies. The signature approach is limited because it cannot detect anomalies resulting from the attacks which are modifications of already known attacks, and machine learning-based approaches can lead to false positives and missed anomalies. The hybrid approach based on the signature analysis and an ensemble of classifiers shows better results compared to some other approaches, but it does not take into account information from standard anti-virus protection tools, firewalls, etc., presented in the form of parameters of network endpoints. Under these conditions, an urgent task is to develop systems that allow detecting and classifying network anomalies and failures using endpoint parameters. Such system is capable not only respond to information security incidents on network nodes more efficiently [5, 6], but also provide detailed information about detected anomalies.

The parameters of the network endpoints cannot be directly used to classify anomalies using classifiers which inputs are the session parameters. The approach to training classifiers on the endpoint parameters differs in the need of individual settings for each node, as well as the creating a labeled database of typical events.

Nevertheless, it seems possible to use these parameters for anomaly detection using the existing ensemble of classifiers if they are reduced to the session parameters format.

This article proposes a hybrid approach to anomaly detection taking into account the parameters of endpoints. The basis of the approach is the existing anomaly detection system [3]. The transformation of the network endpoint parameters to the format of the session parameters is implemented using a technique of regrouping and dimensionality reduction [7], with dimensionality reduction understood as bringing the dimension of the space of endpoint parameters to the dimension of the space of network session parameters suitable for ensemble classification. The choice of this transformation method is determined, firstly, by the compatibility of nodal and network event spaces due to the fact that nodes form a network, and their connections form network traffic; secondly, the need to synchronize the data of both levels for the full detection and study of anomalous activity [8].

To verify the accuracy and repeatability of the results, we carried out an experimental evaluation of the application of the proposed hybrid approach.

## 2 Network Endpoint Parameters and Ways to Convert Them to the Format of Session Parameters

The endpoint parameters  $y'_1, \dots, y'_{T'}$  are obtained from anti-virus protection tools, firewalls, additional endpoint detection and response (EDR) tools and represent the state of these nodes. Some of the parameters  $y'_1, \dots, y'_{T'}$  are also session parameters at the same time, e.g., discovery time, source and destination IP addresses, protocol, data packet length, etc. [9, 10]. Other parameters, such as a file name and its path, authorized user data, user rights data, etc., represent the state of the network nodes. There are two ways to convert endpoint parameters to the session parameters format: based on manual search and based on regrouping parameters by data type [11]. The first method involves a manual search for subsets of node parameters, from which it is possible to restore the session parameters corresponding to them by enumeration. Due to the exponential complexity, such enumeration is not efficient [12]. The second method is based on automated partitioning of the set of node parameters into subsets according to the parameter data type, as well as on the rules for converting parameters within these subsets into the session parameters format.

To convert the parameters of endpoints in a hybrid approach, a method based on regrouping by data type was chosen, since it allows the conversion in near real time mode when using standard software [11].

The procedure for converting endpoint parameters to the format of session parameters consists of the following steps:

**Step 1.** Obtain  $y'_1, \dots, y'_{T'}$  endpoint parameters and  $x_1, \dots, x_Q$  session parameters.

**Step 2.** Find a subset  $T_1 \subseteq T'$  of endpoint parameters that are simultaneously the parameters of session  $T_1 \subseteq Q$ . These session parameters will be equal to the endpoint parameters:  $x_q = y'_t, t \in T_1$ .

**Step 3.** Find such a subset  $T_2 \subseteq T'$  of endpoint parameters that  $T_2 \cap T_1 = \emptyset$  and can be used to restore the values of session parameters. Then, to make the  $Y(T') \rightarrow X(Q)$  transformation that can be used to restore the session parameters for all the endpoint parameters.

*Step 1* of the procedure is done by obtaining the available parameters of the network endpoints from anti-virus protection tools, firewalls, etc. The session parameters are taken from the traffic analyzer of the current system after dimensionality reduction [3]. On *step 2*, node parameters are identified, such as discovery time, source and destination IP addresses, which are both endpoint and session parameters. Primarily *step 3* reveals logical interconnections between node states and network parameter values. For example, deletion of a file on a node can be represented as a decreasing the size of the file, with information transmitting from the node's IP address in the form of the corresponding parameter values of this node. The deletion of a file occurs during the session when data is transferred from this node to another, that helps to be processed by the protection tools on the nodes and can be recognized as an unauthorized data deletion in the context of the corresponding session.

The result of the procedure is the  $Y(T') \rightarrow X(Q)$  transformation of network endpoint parameters to the format of the session parameters. The procedure is preliminary and single for a network under protection.

### 3 Hybrid Approach to Detecting Network Anomalies Using the Endpoint Parameters

The basis of the approach is the anomaly detection system [3]. The structure of this system is improved by the network endpoint parameters, a block for synchronizing the node parameters with the current session, a block for converting the network endpoint parameters to the format of the session parameters, a signature analyzer with the Wazuh rule base, and three classifiers: logistic regression [13, 14], stochastic gradient descent [15, 16], and decision tree [17, 18] identical to those already available (Fig. 1). In addition, due to the increase in the number of signature analyzers and classifiers in the ensemble, the terminal classifier is modified basing on the use of a weighted aggregation operator that takes into account these additional sources of information.

The endpoint parameters are obtained from the network protection tools and sent to the synchronization unit. The block of synchronization with the current session transfers the values of the parameters  $y'_1, \dots, y'_{T'}$  to the following blocks of hybrid system. The values of the parameters are only ones that are received in the interval from the beginning to the end of the current session. An additional signature analyzer is used to detect possible network anomalies by checking whether the node event parameter values match the specified regular expressions. These regular expressions correspond to previously known network anomalies and are included in the Wazuh signature set [10]. The choice of this set is conditioned by the fact that it is intended to detect anomalies based on the node parameters, including those representing the

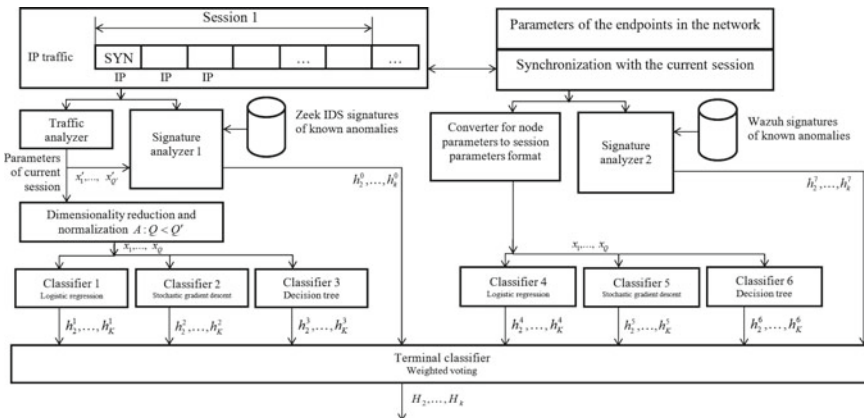


Fig. 1 Hybrid system structure



integrity of the file systems on the nodes and also because the corresponding analyzer is compatible with many other anomaly detection systems [11, 19].

The block for converting parameters of network endpoints to the session parameters format uses the conversion procedure described in the previous section and allows the conversion of all suitable endpoint parameters to the session parameters format. Classifiers 4–6 are exact copies of the classifiers of the implemented anomaly detection system [3]. These classifiers work in parallel with classifiers 1–3 and are designed to detect anomalies, with information extracted from the transformed parameters of network nodes. The outputs of classifiers 1–6 are discrete values  $h_2^1, \dots, h_K^1, \dots, h_2^6, \dots, h_K^6$  associated with anomaly classes  $S_2, \dots, S_K$ . Each of these discrete values takes values within the set  $\{-1, 1\}$ . Each classifier assigns the current session to the class for which the corresponding value is 1. Class  $S_1$  corresponds to a session without anomalies. In this case, the outputs of all classifiers must take the value  $-1$ . The outputs of signature analyzers 1 and 2 are discrete values  $h_2^0, \dots, h_K^0, h_2^7, \dots, h_K^7$ , respectively, with values within the set  $\{-1, 1\}$ . The current session is assigned to the class for which the corresponding output is equal to 1 by each of the signature analyzers. New signatures corresponding to newly detected anomalies can be added to the signature sets of both analyzers (traffic and nodal events) during system operation.

To assign the current session to one or another anomaly class  $S_2, \dots, S_K$ , it is sufficient that at least one of the values  $h_2^0, \dots, h_K^0, h_2^7, \dots, h_K^7$  equals to 1. In this case, one of the signature analyzers will reveal the match of the current values of the session or endpoint parameters with any signature.

The terminal classifier is a set of operators for aggregating the outputs of signature analyzers and classifiers 1–6 for each of the classes  $S_1, \dots, S_K$ . These operators represent the expert opinion on which class the current session should be assigned to for all possible combinations of the values of  $h_2^0, \dots, h_K^0, \dots, h_2^7, \dots, h_K^7$ . The weighted voting scheme is taken as the basis for building up these operators. The output value of the terminal classifier for class  $S_k$  is given by:

$$H_k = \max \left( \text{sign} \left( \sum_{i=1}^3 \alpha_i h_k^i + \beta \left[ \sum_{i=4}^6 \alpha_i h_k^i \right], h_k^0, \beta h_k^7 \right) \right) \quad (1)$$

where  $H_k$  may range within  $\{-1, 1\}$  and shows whether the current session is assigned to class  $S_k, k = 2, \dots, K$ ;  $\alpha_i$  are weighting factors matched to the classifiers based on the number of mistakes they made during the learning process. The authors in [3] describe the way they obtain the values of these factors. Since classifiers 4–6 are identical to classifiers 1–3, and the corresponding factors  $\alpha_i$  are initially equal. While data accumulating and training of classifiers, the factors will be corrected.

Factor  $\beta$  ranges within  $[0, 1]$  and adjusts the sensitivity of the terminal classifier to the parameters of nodes, with data enriching the information about the network session. The maximum sensitivity, or “anxiety threshold”, is set at  $\beta = 1$ . When  $\beta = 0$ , there is no adjustment at the node level and it reduces to the analysis of the network session only.

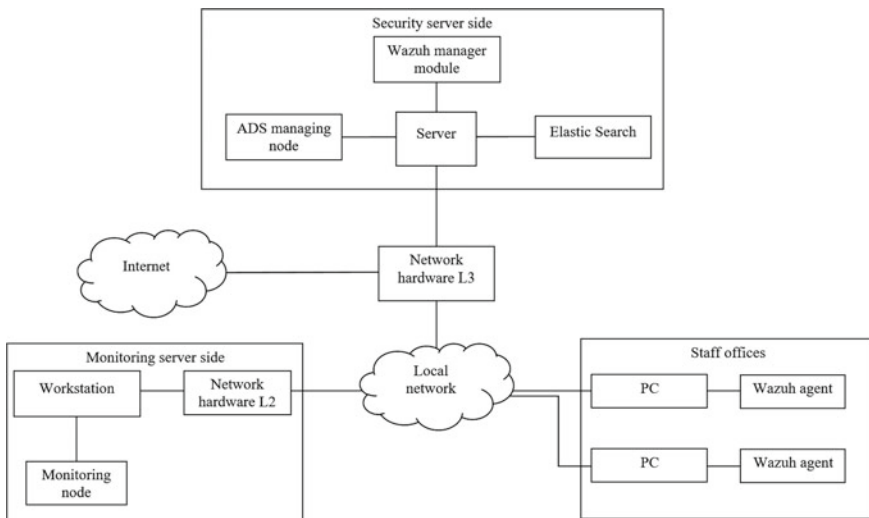
The anxiety threshold is set before the operation of the hybrid system, its value depends on the totality of external conditions that can affect the security of the system being protected from the outside. Such conditions can be abrupt changes in the distribution of the use of information resources, caused, e.g., by a bulk transfer to a remote work mode or information counteraction [19, 20].

## 4 Results

### 4.1 Gathering of Session and Endpoint Parameters

For the experiment, a virtual structure was used (see Fig. 2).

This structure emulates the work of several nodes of the local network. The security server side includes the server itself where Elastic Stack is installed, which will be accessed by the Wazuh manager module, as well as the managing node of the existing anomaly detection system (ADS). Elastic Stack is a data migration platform using the Wazuh agent. This platform receives data and notifications from endpoints, analyzes them and provides information about the state of the network in the format of session parameters, which makes it possible to detect anomalies using an ensemble of classifiers based on machine learning [21]. The monitoring node is located on a workstation on the monitoring server side to monitor all network traffic. In order to



**Fig. 2** Virtual structure

prevent failures in the operation of the ADS from disrupting the network, a workstation with an installed monitoring node is connected through the network equipment of the data link layer of the OSI model [22].

The Wazuh agents are installed on employees' workstations. The agents gather data in the form of parameters of endpoints, then they transfer them to the Wazuh manager module. To detect anomalies, rules are set in the configuration block of the Wazuh manager module. The procedure for converting the parameters of endpoints to the session parameters runs under the rules mentioned above. Transformation rules are determined before testing, but can be supplemented during the operation of the protective system. An example of such a transformation is the conversion of the "Deletion of a file on a node" event to the session format described in Sect. 4.2. Thus, the Wazuh manager module transfers the session parameters to the ADS managing node directly, and the parameters of the nodes after conversion in accordance with the procedure specified in the form of rules conversion of node parameters.

The values of network traffic parameters are checked for compliance with known signatures, brought to a normalized form (we mean both the normalization of values and the reduction in the dimension of the parameter vector itself [3]). Then the values are transferred to the ELK stack for final processing and classification described by these parameters network event [21]. File system parameter values, i.e., nodal events, on the Wazuh manager module, primary data are processed by step 3. In the case of the above example with deleting a file, data about changing the state of the file system (freeing memory where the file was stored) from the Wazuh agent are transmitted to the primary data processing unit by the Wazuh manager module, where they are divided into common ones with the network session parameters and the ones to be regrouped. The resulting new parameter vector corresponding to the parameter vector of the normalized session is sent to the ELK stack, where the final classification of the incoming event takes place.

## 4.2 Testing

The hybrid anomaly detection approach was tested using the tools of the scikit-learn library. The terminal classifier is a weighted voting method implemented using the VotingClassifier module; the GridSearch module is used to optimize hyperparameters. GridSearch implements an optimization algorithm that includes the choice of optimization parameter values. One of these parameters is "anxiety threshold"  $\beta$ . When  $\beta = 0$ , the system shows the best results for network attacks (98.66%) and the worst ones presented in column Msc of Table 1 (76.48%). Experiments to detect anomalies at different values of  $\beta$  showed that when  $\beta = 0.946$ , the largest number of anomalies was detected. Table 1 shows the detailed information for various types of network attacks and classifiers: logistic regression (LR), stochastic gradient descent (SGD), decision tree (DT), and their combinations.

These results allow us to conclude that the involvement of additional data in the form of endpoint parameters allows expanding the range of detected anomalies, as

**Table 1** Recognition accuracy for various types of network attacks by optimized classifiers, %

Classifier	Types of attacks				
	Normal session	DDoS	Brute force	Network attack	Msc
DT	90.32	96.66	82.17	94.29	82.31
SGD	88.47	93.58	80.11	90.63	74.88
LR	86.84	92.17	75.82	89.19	72.31
DT + SGD	93.37	97.92	85.83	95.41	84.17
DT + LR	92.88	97.21	84.37	94.91	83.93
SGD + LR	90.46	95.71	81.68	93.83	77.16
Network ensemble	86.93	95.73	85.53	98.66	76.48
Hybrid ensemble	96.15	98.95	91.33	99.86	86.74

well as increasing the accuracy of anomaly detection compared to anomaly detection without these data from 1 to 10% depending on the type of anomaly.

## 5 Conclusion

The article proposes a hybrid approach to anomaly detection based on a weighted voting classifier, taking into account the parameters from the endpoints. The experiment showed a higher efficiency in detection of various types of anomalies in comparison with ADS without taking into account the parameters from the endpoints. The use of an anxiety threshold allows us to control the performance of the system depending on changing conditions that makes it possible to free up computing power for other tasks.

The concept of Smart city involves the use of the Internet of things (IoT) to manage urban infrastructure. The IoT is a computer network where the nodes are the agents that interact with each other or with the outside world without human intervention. The full operation of the IoT systems requires well-coordinated interaction between all agents, for example, traffic lights can receive information about quantity of vehicles to regulate traffic. Thereby, it is very important to establish uninterrupted operation of the IoT. To do this, it is necessary to quickly identify and isolate malicious or failed agents. Malicious effects on agents or failures in their work are the causes of network anomalies. In the future, it is possible to introduce the developed approach into both current and future anomaly detection systems for Smart city.

In particular, such systems are designed to detect the attacks on agents, e.g., unmanned vehicles [23–25]. Malicious agents are detected based on reputation [24], trust [25], and data quality metrics [23]. In the future, we are planning to develop the proposed approach by setting the values of the anxiety factor based on these indicators.

The research (S. Sakulin and A. Alfimtsev) is done with the financial support: Russian Science Foundation # 22-21-00711.

## References

1. Borkovich DJ, Skovira RJ (2020) Working from home: cybersecurity in the age of COVID-19. *Issues Inform Syst* 21(4):234–246
2. Zakoldaev DA, Vorobeva AA (2021) Confidentiality assurance in multi-agent robotic system. *Turk J Comput Math Edu* 12(2):2659–2663
3. Sakulin S, Alfimtsev A, Kvitchenko K, Dobkacz L, Kalgin Y, Lychkov I (2022) Network anomalies detection approach based on weighted voting. *Int J Inform Sec Privacy* 16(1):82–99
4. Saranya T et al (2020) Performance analysis of machine learning algorithms in intrusion detection system: a review. *Proc Comput Sci* 171:1251–1260
5. Mokhtari S, Abbaspour A, Yen KK, Sargolzaei A (2021) A machine learning approach for anomaly detection in industrial control systems based on measurement data. *Electronics* 10(4):1–13
6. Habeeb RAA et al (2019) Real-time big data processing for anomaly detection: a survey. *Int J Inform Manag* 45:289–307
7. Acharya T et al (2021) Efficacy of machine learning-based classifiers for binary and multi-class network intrusion detection. In: 2021 IEEE international conference on automatic control & intelligent systems (I2CACIS). IEEE, pp 402–407
8. Heinz C, Zuppelli M, Caviglione L (2021) Covert channels in transport layer security: performance and security assessment. *J Wirel Mobile Netw Ubiqu Comput Depend Appl* 12(4):22–36
9. Abudalfa SI, Isleem ES, Khalil MJE, Dalloul RA, Iqtefan SM (2022) Evaluating performance of supervised learning techniques for developing real-time intrusion detection system. *Int J Eng Inform Syst* 6(2):103–119
10. Wurzenberger M et al (2022) Automatic attack pattern mining for generating actionable CTI applying alert aggregation. In: *Cybersecurity of digital service chains*. Springer, Cham, pp 136–161
11. Catch suspicious network traffic (2022) Learning Wazuh [Электронный ресурс]. <https://documentation.wazuh.com/current/learning-wazuh/suricata.html> (дата обращения: 08.04.2022)
12. Ma W et al (2021) A two-stage hybrid ant colony optimization for high-dimensional feature selection. *Pattern Recogn* 116:1–13
13. Ünal U et al (2021) Investigation of cyber situation awareness via SIEM tools: a constructive review. In: 2021 6th international conference on computer science and engineering (UBMK). IEEE, pp 676–681
14. Sworna ZT, Mousavi Z, Babar MA (2022) NLP methods in host-based intrusion detection systems: a systematic review and future directions, pp 1–35. arXiv preprint [arXiv:2201.08066](https://arxiv.org/abs/2201.08066)
15. Tabash M, Abd Allah M, Tawfik B (2020) Intrusion detection model using naive bayes and deep learning technique. *Int Arab J Inform Technol* 17(2):215–224
16. Zahras D, Rustam Z, Sarwinda D (2019) Soft tissue tumor classification using stochastic support vector machine. *IOP conference series. Mater Sci Eng* 546(5):1–6
17. Sokolov SA, Iliev TB, Stoyanov IS (2019) Analysis of cybersecurity threats in cloud applications using deep learning techniques. In: 2019 42nd international convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, pp 441–446
18. Elsaid SA et al (2019) Cybersecurity: design and implementation of an intrusion detection and prevention system. In: *International conference on computing*. Springer, Cham, pp 15–42
19. Landauer M, Skopik F, Wurzenberger M, Rauber A (2022) Dealing with security alert flooding: using machine learning for domain-independent alert aggregation. *ACM Trans Privacy Sec* 25(3):1–36

20. Mandal S, Khan DA, Jain S (2021) Cloud-based zero trust access control policy: an approach to support work-from-home driven by COVID-19 pandemic. *New Gen Comput* 39(3):599–622.32
21. Vadhil FA, Nanne MF, Salihi ML (2021) Importance of machine learning techniques to improve the open source intrusion detection systems. *Indonesian J Electr Eng Inform* 9(3):774–783
22. Hulič M, Baláž A, Štancel M (2019) Possibilities of methods for IDS testing. In: 2019 17th international conference on emerging elearning technologies and applications (ICETA). IEEE, pp 257–262
23. Hataba M et al (2022) Security and privacy issues in autonomous vehicles: a layer-based survey. *IEEE Open J Commun Soc* 3:811–829
24. Kianersi D et al (2022) Evaluation of a reputation management technique for autonomous vehicles. *Future Internet* 14(2):31–52
25. Chuprov S et al (2020) Reputation and trust models with data quality metrics for improving autonomous vehicles traffic security and safety. In: 2020 IEEE systems security symposium (SSS). IEEE, pp 1–8

# IoT Infrared Imaging of Livestock Tissues Using a One-Eyed Bandit Technique



Stefan Rizanov, Peter Yakimov, and Dimitar Todorov

**Abstract** Bio-monitoring of livestock has become a popular topic in scientific research over the past 15 years. With the introduction of the concept of smart farming a necessity has arisen for utilizing modern methods such as infrared thermography (IRT) for evaluating the State-of-Health of farm animals. Tissue imaging via IRT is a complex task, due to thermal pattern changes induced by variances in skin thickness, tissue structure and fat contents, fur coating color, hair length and thickness, skin emissivity, animal stress levels, animal movement, ambient temperature and humidity. The goal of our work is to present an automated novel IoT system, which can address these difficulties, allowing scalability and performing farm-level animal well-being assessment in a non-stressful manner. The developed system implements a ratiometric One-Eyed-Bandit Technique (OEBT), automated thermographic object recognition software and algorithms for applying dynamic corrections to the captured raw thermographic data.

**Keywords** Bio-monitoring · IoT · Livestock · Tissue imaging · Thermography · Smart farming

## 1 Introduction

The utilization of Infrared Thermography (IRT) as a biomedical diagnostic tool is gradually gaining popularity. Body extremity and tissue temperatures are related to the underlying blood flow dynamics within the peripheral blood vessels, with the

---

S. Rizanov (✉) · P. Yakimov · D. Todorov  
Faculty of Electronic Engineering and Technologies, Technical University of Sofia, 8, Kliment  
Ohridski Blvd., 1000 Sofia, Bulgaria  
e-mail: [srizanov@tu-sofia.bg](mailto:srizanov@tu-sofia.bg)

P. Yakimov  
e-mail: [pj@tu-sofia.bg](mailto:pj@tu-sofia.bg)

D. Todorov  
e-mail: [dgt@tu-sofia.bg](mailto:dgt@tu-sofia.bg)

localized tissue heat patterns being affected by vasodilation and vasoconstriction of near-surface vessels [1, 2]. The skin thickness is not fixed, in cattle it varies within the range of 3–5 mm—with areas of increased skin thickness appearing cooler [3, 4]. The observed skin temperature is lower than internal core body temperature by roughly 5 °C [5]. Localized fat thickness also affects the captured thermal patterns—skin patches with thicker fat layers appearing cooler. A bovine’s, equine’s and swine’s skin emissivity is roughly 0.98, but it varies in value depending on the body region, from 0.92 to 0.98, due to differences in tissue structure [1, 6, 7]. Hair presence additionally reduces the observed IR temperature by around 5 °C, with the hair’s density, thickness and length being contributing factors to this reduction—darker haired patches appear hotter, patches of denser hair coatings appear cooler. The observed temperatures of peripheral tissues are correlated with the ambient temperature and humidity. The increase in relative humidity leads to an elevation of captured thermographic temperatures, due to its inhibiting effect over the transmission of IR radiation and the IR radiation emission of water vapor [8]. Object-to-camera distance also affects captured IR data, making animal movements a source of measurement error. Handling, social isolation and interaction with stressors of sudden and unknown nature cause stress to the animals. The presence of stress induces changes in the skin temperature heat patterns, due to the activation of the hypothalamic–pituitary–adrenocortical axis (HPA), followed by an increase in catecholamine and plasma corticosteroid concentrations [9–11]. This fact necessitates that bio-monitoring of animals should be performed through a non-stressful and familiar to the animals method—typically this is only feasible via measurement automation and the mitigation of human intervention and interaction. These contributing factors—skin thickness variance, emissivity changes, hair coating effects, ambient environmental effects, object-to-camera fluctuation, stress-induced reactions to unfamiliar events, make the utilization of IRT difficult in practice. Our aim with this work was to develop an automated bio-monitoring system, which tries to tackle these sources of measurement errors and expand upon the utility of IRT as a diagnostic tool in smart farming. Chapter 2 presents structural details regarding the developed hardware system and the application of the OEBT. Chapter 3 discusses the embedded-level and high-level software, developed for our system, consisting of data filtering, IR object detection, object parameter evaluation and the application of dynamic corrections to the raw data. Chapters 4 and 5 presents our experimental setup and an analysis performed over a sample of our gathered experimental data.

## 2 Developed Hardware System and the One-Eyed Bandit Technique

The developed system consists of 3 PCBs, connected together—Master Board, Blinded Matrix Board, ToF Board. Figure 1 shows a block diagram of the whole system and the main modules, contained within the sub-boards. The developed device

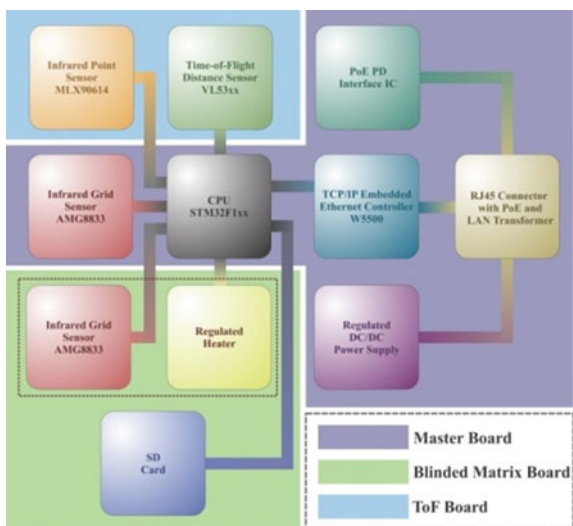


is an Ethernet-based PoE-powered PD Class 4 measurement end-node, part of a farm-level local network. The Ethernet connectivity and Power-over-Ethernet (PoE) allow for the transfer of data and power (up to 25.5 W) to the devices to be performed with limited requirements toward farm-level supporting infrastructure build-up, prior to the system's installation. The Master Board contains a  $8 \times 8$  microgrid IR sensor AMG8833 which captures the thermal pattern of the object of interest (OOI). The AMG8833 has a NETD parameter value of  $\pm 50$  mK, making in comparable in sensitivity even with expensive large-matrix IR cameras. The Blinded Matrix Board contains a second AMG8833 sensor, whose Field-of-View (FoV) is covered by a reference object, which is heated by a regulated heater. The contact temperature of the heater is measured via an on-board Pt1000 sensor circuit. Additionally the Blinded Matrix Board allows the connection of a SD card for local data storage (logging) purposes. The ToF Board contains a VL53xx Time-of-Flight (ToF) distance sensor and a MLX90614 point IR sensor, which expands the calibration capabilities of the system. Figure 2 shows photographs of the developed hardware system.

Within this work, we propose the utilization of the so-called One-Eyed Bandit Technique (OEBT). The principle behind this technique is illustrated in Fig. 3. When external factors affect a measurement they can be considered as either additive or multiplicative noise floor components, depending on whether or not the measured signal is modulated by them. When additive noise is present—a differential technique is required to remove it from the useful signal. This is performed by shielding the reference sensor from the stimulus of interest and subjecting it only to the noise component, then a subtraction is performed between the main sensor's and reference sensor's output values.

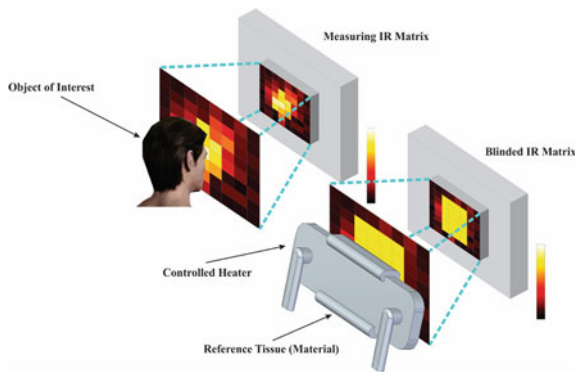
When multiplicative noise is present—a ratiometric technique is required to remove it from the useful signal. With the ratiometric technique the main sensor

**Fig. 1** Block diagram of the developed hardware system





**Fig. 2** Photographs of the developed system



**Fig. 3** One-Eyed Bandit technique principle

is subjected to the stimulus of interest and the reference sensor is subjected to a controlled known stimulus. With both sensors being subjected to the same work environment—performing a following division between the main sensor’s output value and that of the reference sensor’s output value eliminates the multiplicative noise [12]. Due to the fact that external environmental factors such as ambient humidity and temperature affect the IR radiation absorption and transmittance parameters—these environmental factors can be considered of multiplicative nature. Within our developed system, we implemented the One-Eyed Bandit technique by covering the reference microgrid IR sensor with a tissue of the same thermographic parameters as that of the tissue of the object of interest. Heating our reference tissue patch in a controlled and measurable manner allows us to observe a known stimulus with the reference sensors. Equation 1 describes the formula, through which we calculate the correction coefficient  $C$  and apply it over the raw IR data:

$$IR_{AVG_{REAL}} = IR_{AVG_{RAW}} * C = IR_{AVG_{RAW}} * \frac{T_{Contact_{REF_{OBJ}}}}{T_{IR_{REF_{OBJ}}}}, \quad (1)$$

where  $IR_{AVG_{RAW}}$ —IR main sensor measured average temperature of the detected object;  $C$ —correction coefficient;  $T_{Contact_{REF\_OBJ}}$ —Pt1000 contact measured temperature of the heated reference patch;  $T_{IR_{REF\_OBJ}}$ —IR reference sensor measured average temperature of the detected object. This One-Eyed-Bandit Technique creates essentially a ratiometric configuration between the two IR camera sensors. The developed hardware additionally allows us to switch to the application of a differential configuration, instead of a ratiometric one, by stopping the process of heating the reference tissue patch. Hence this module and its configuration allows us to compensate both the multiplicative and additive noise floor components. Having the reference tissue being of the same type as the observed tissue as that of the OOI—eliminates the necessity of performing additional computational steps in order to compensate for the tissue's emissivity and other thermographic parameters, such as thermal conductivity.

### 3 Developed Software Package

In order for the developed system to be operational we developed both a embedded-level C-based and high-level Python-based software package. On the embedded level we developed: a Client application for the Client–Server data transfer within the network, utilizing a custom Client–Server protocol; a time-switched state machine, controlling the processes of sensory data acquisition and transfer to the local Server. No exploratory data analysis (EDA) or digital signal processing is performed on the embedded level, in order to ease the end-node side's operational workload and make use of the larger computational resources of the Server. On the Server side we implemented a database module; an automated object recognition module and a coordinate table control module. The database module collects the raw sensory data from the end-node (Main sensor and Reference sensor IR images; ToF distance; Heater contact temperature; IR point temperature) and stores them in a data file, with every new file having a unique name, automatically generated each day. Each data entry header contains a timestamp; an ID number of the end-node device, which forwarded the data; detected object parameters. The detected object parameters are resultant from the automated object recognition module and are: `maximum_object_temp`; `minimum_object_temp`; `average_obj_temp`; `median_obj_temp`. The object recognition module computes an object mask for each captured thermogram and defines which pixels are part of an object. This is performed by two Python functions: **Image\_OBJ\_Mask (...)**; **Quick\_OBJ\_Mask (...)**. The calculation of the object parameters is performed by the function **Extract\_OBJ\_Parameters (...)**. A pseudo-code representation of these 3 functions is shown in Fig. 4. Object recognition is applied over both the main and reference sensors' IR data.

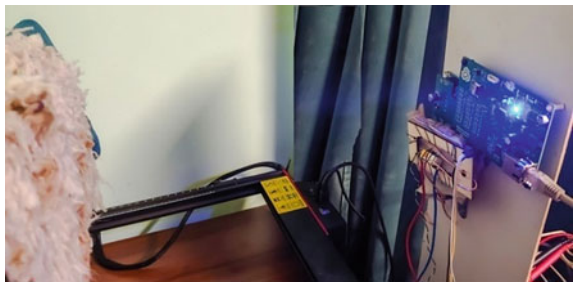
<b>Image_OBJ_Mask</b> ( <i>IR_Image</i> , <i>Threshold</i> , <i>Offset</i> ):  Create empty 2D array of size( <i>IR_Image</i> ) Extract corner ( <i>x,y</i> ) gradients Extract edge ( <i>x,y</i> ) gradients Extract internal ( <i>x,y</i> ) gradients <b>IF</b> ( <i>Extracted_Gradients</i> > <i>Threshold</i> ): Mark <i>IR_Image</i> Pixels as <i>OBJ_Contour_Pixels</i> Add <i>OBJ_Contour_Pixels</i> to <i>OBJ_Mask</i> <i>min_val</i> = min( <i>OBJ_Contour_Pixels</i> ) Scan <i>IR_Image_Pixels</i> <b>IF</b> ( <i>Image_Pixel</i> > ( <i>min_val</i> - <i>Offset</i> )): Add <i>IR_Image_Pixel</i> to <i>OBJ_Mask</i>  <b>return</b> <i>OBJ_Mask</i>	<b>Quick_OBJ_Mask</b> ( <i>IR_Image</i> ):  <b>Calculate:</b> min( <i>IR_Image</i> ) max( <i>IR_Image</i> ) avg( <i>IR_Image</i> ) <i>Dynamic_Threshold</i> = min( <i>IR_Image</i> ) + ( max( <i>IR_Image</i> ) - min( <i>IR_Image</i> ) ) / 2  <i>Dynamic_Offset</i> = avg( <i>IR_Image</i> ) * 0.02 <b>Image_OBJ_Mask</b> ( <i>IR_Image</i> , <i>Dynamic_Threshold</i> , <i>Dynamic_Offset</i> )  <b>return</b> <i>OBJ_Mask</i>	<b>Extract_OBJ_Parameters</b> ( <i>IR_Image</i> , <i>OBJ_Mask</i> ):  Scan <i>OBJ_Mask</i> elements <b>IF</b> <i>OBJ_Mask</i> element marks <i>OBJ_Pixel</i> : Add corresponding <i>IR_Image</i> Pixel to 1D List <b>Calculate:</b> min( <i>OBJ_Pixels</i> ) max( <i>OBJ_Pixels</i> ) avg( <i>OBJ_Pixels</i> ) median( <i>OBJ_Pixels</i> )  <b>return</b> min, max, avg, median
--	--	---

**Fig. 4** Pseudo-code representation of the object recognition software functions

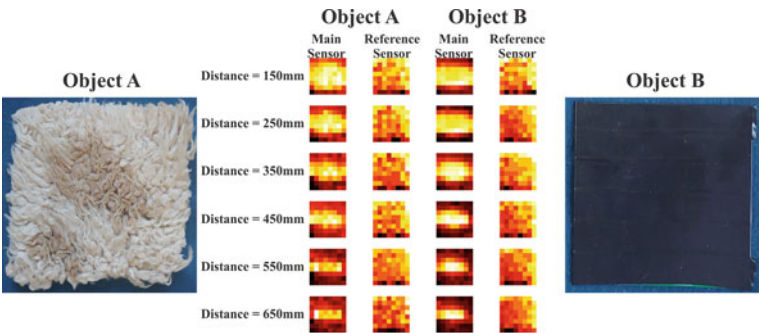
## 4 Experimental Setup and Gathered Data

In order for us to evaluate the functionality and validate the usefulness of the OEBT, we developed an experimental setup and gathered measurement data. Figure 5 shows photographs of the built experimental setup.

The experimental setup contains: the developed measurement system; a control-able 2-axial coordinate table with a 1 mm step and 500 mm movement range, driven by our developed software package; a heating bed, attached to the moving head of the coordinate table; a TENMA 72—10,505 power supply block; 2 calibrated Pt1000 sensors; a Fluke 867B calibrated multimeter; a PoE Injector; an operator PC. Two different objects were measured—a real sheep skin, with fur covering and different colored fur patches (Object A); an electrically insulating material, covered with a black adhesive tape, which has an emissivity of 0.985 (Object B). The Reference IR sensor of the developed system was blinded with a material patch of the same type as Object A and Object B. One of the Pt1000 sensors is mechanically attached between the frontal surface of the heating plate and on the back surface of the analyzed object. This Pt1000 sensor’s value is measured with the Fluke 867B high-precision multi-meter and this data represents the contact temperature of the object. Through our experimental setup we gathered around 500 thermographic images of Object A and Object B—collecting both temporal data (object temperature change as a function of time) and distance data (object temperature as a function of distance to the object).



**Fig. 5** Experimental setup



**Fig. 6** Thermographic image data sample for Object A and Object B at different distances

For the temporal data we performed measurement samples at fixed time intervals of 10 s. For the distance data we performed measurements with a 5 mm step size. Figure 6 shows Object A and Object B and a sample of the corresponding distance thermographic image data.

5 Data Analysis

Figure 7 displays the temporal data for Object B. As can be seen the contact sensor temperature fluctuates in a roughly sinusoidal form, due to the internal regulator of the used heater plate. This effect is useful, due to the fact that not only the relative difference between the IR-measured and contact measured values can be evaluated, but also the system’s capability and accuracy in monitoring temperature changes.

The raw IR measurement data fluctuates quite a lot, which lead us toward utilizing a moving average filter of `window_size = 5` samples over the raw data. This lead to smoothing of the curves and improved upon the measurement accuracy slightly. The raw and filtered data is shown on the two left sub-figures of Fig. 7. In order for us to evaluate which of the IR object parameters (min, max, avg or median) is correlated the most with the contact temperature. We calculated and plotted the differences between the curves—this is shown on the right 2 sub-figures of Fig. 7.

Our results showed that the average and median temperatures have a higher correlation with the contact measured temperature, than that of the minimum and maximum IR temperatures. Table 1 shows a table summarizing all of the results from this experiment.

In order to evaluate if the One-Eyed Bandit Technique improves upon the measurement accuracy, we performed a temporal data experiment with Object A at a fixed system-to-object distance of 250 mm. Figure 8 shows the time-series data for the average temperature. Figure 9 shows the time-series data for the corrected average temperature, using the One-Eyed Bandit Technique.

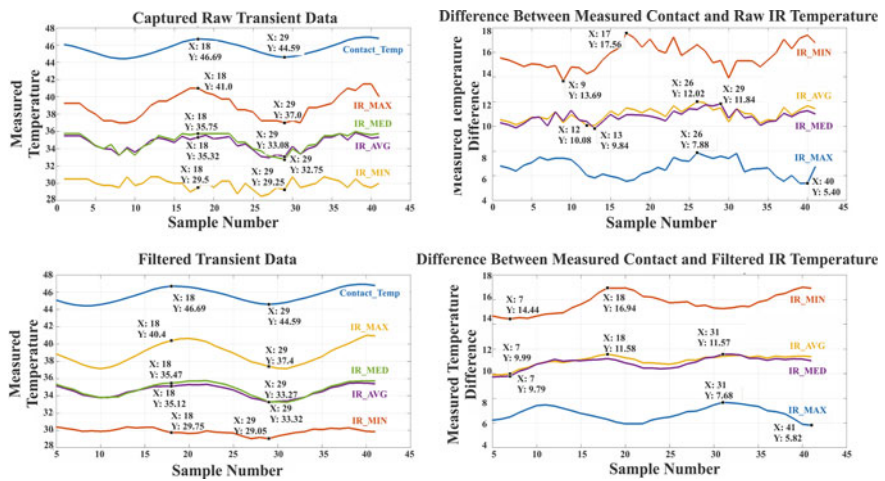


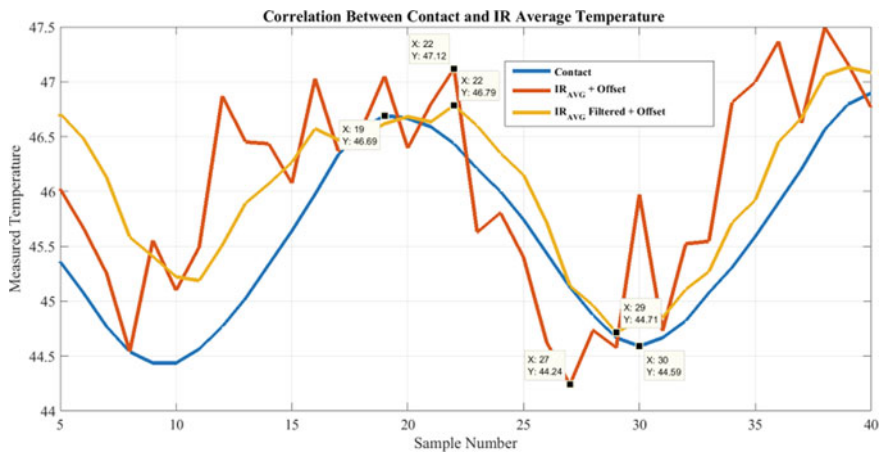
Fig. 7 Transient data for Object B

Table 1 Summary of the transient data from the experiment with Object B

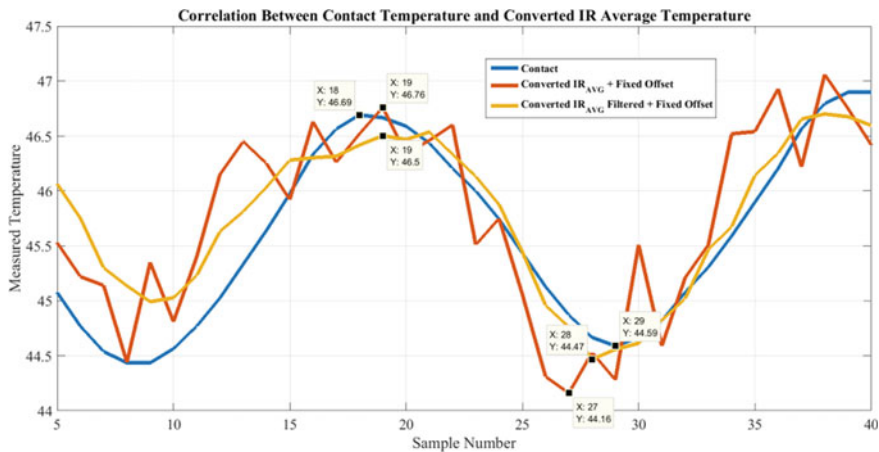
	Raw absolute temperature		Filtered absolute temperature		Difference between contact and raw IR temperatures		Difference between contact and filtered IR temperatures	
	Contact Max	Contact Min	Contact Max	Contact Min	Max	Min	Max	Min
Contact	46.69	44.59	46.69	44.59	0.00	0.00	0.00	0.00
IR_MAX	41.00	37.00	40.40	37.40	7.88	5.40	7.68	5.82
IR_MIN	29.50	29.25	29.75	29.05	17.56	13.69	16.94	14.42
IR_AVG	35.32	33.08	35.12	33.32	12.02	10.08	11.58	9.99
IR_MEDIAN	35.75	32.75	35.47	33.27	11.84	9.84	11.57	9.79

As can be seen from our data, applying the correction coefficients to the raw IR\_Average data improves upon the correlation between the measured IR temperature and the temperature fluctuations of the contact measured temperature.

When performing infrared imaging, the camera-to-object distance is of crucial importance, due to its attenuating effect over the received infrared electromagnetic signal power. This effect was monitored during our experiments. Figure 10 shows the data from two temporal data experiments performed with Object B at two different system-to-object distances—250 and 350 mm. As can be seen from our data, at a distance of 250 mm the difference between the contact and average IR temperature has an average value of 11.45 °C; at a distance of 350 mm the difference between the contact and average IR temperature has an average value of 12.77 °C.



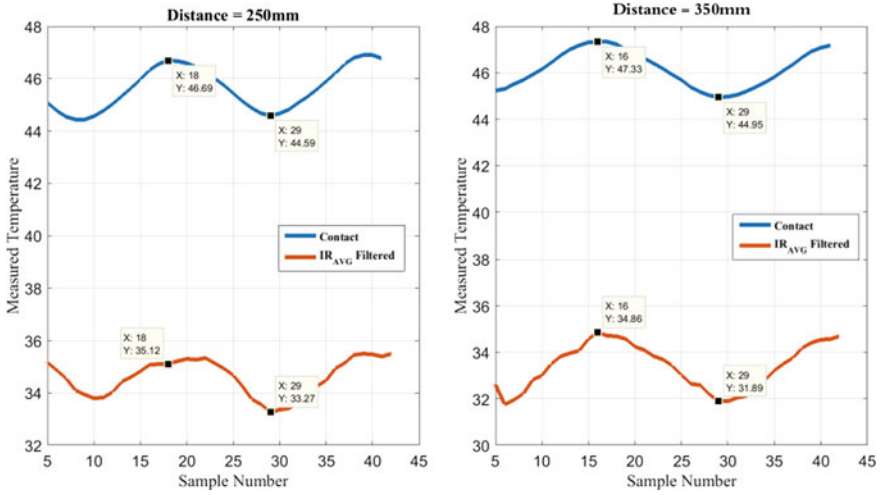
**Fig. 8** Time-series data for raw IR\_Average temperature of Object A



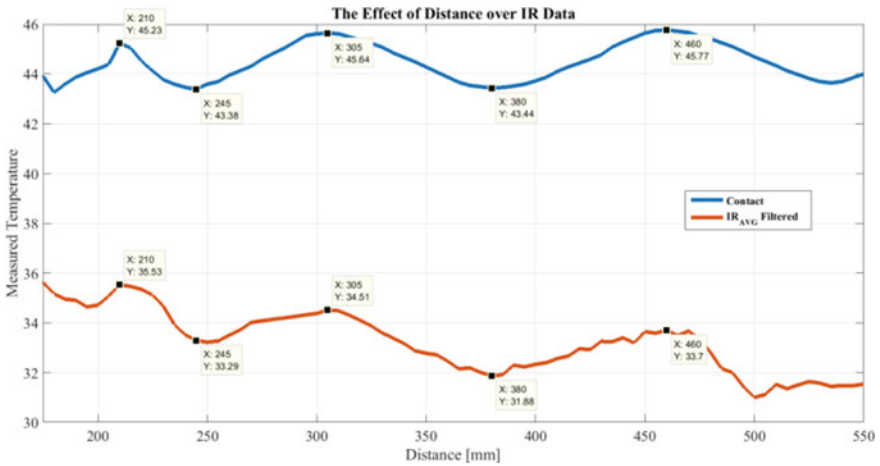
**Fig. 9** Time-series data for corrected IR\_Average temperature of Object A

In order for this system to be fully automated and to be able to apply corrections to the measured IR data (add a dynamically changing offset as a function of distance) this distance effect has to be quantified and modeled. In order for us to do this we performed a distance measurement of step size 5 mm with Object A. We captured IR data points, corresponding to maximums and minimums of the contact measured temperatures and calculated what the offset should be at this distance measurement point. Figure 11 shows a graph which plots the data from the performed experiment. After we captured these data points, we modeled a polynomial second degree regression model, defining what correction offset should be added to the raw IR data as a function of distance. Figure 12 shows the polynomial fit and its parameters.





**Fig. 10** Transient data for Object B at system-to-object distances 250 and 350 mm

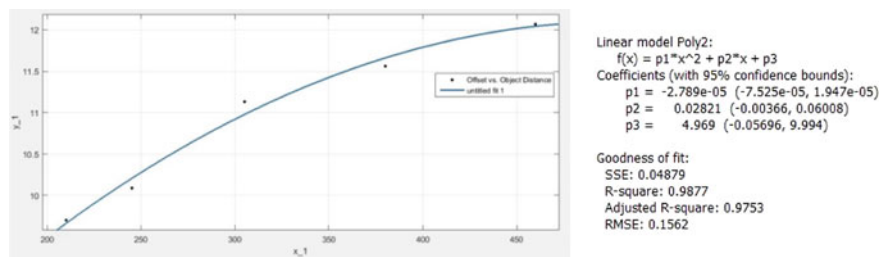


**Fig. 11** Assessing the effect of system-to-object distance with Object A

Our offset data was generalized well by a second degree polynomial, a possible alternative to this can be fitting an exponential model over the offset-to-distance data. Equation 2 shows the distance effect correction formula, which we apply over our raw IR data:

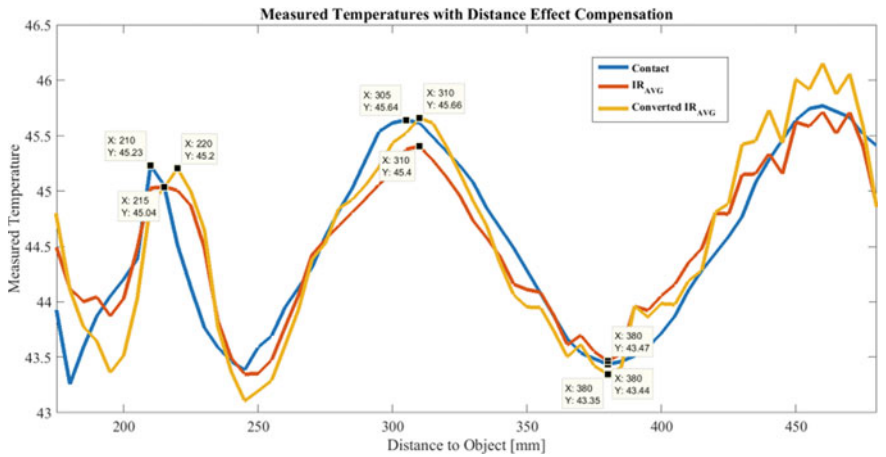
$$\begin{aligned} \text{IR}_{\text{Compensated}} = & \text{IR}_{\text{RAW}} + \text{Offset} = \text{IR}_{\text{RAW}} \\ & + (-2.789 * 10^{-5} * \text{ToF}_{\text{DIST}}^2 + 0.02821 * \text{ToF}_{\text{DIST}} + 4.969), \end{aligned} \quad (2)$$





**Fig. 12** Second degree polynomial fitted over the correction offset as a function of distance relationship

where  $IR_{RAW}$ —raw IR pixel value;  $ToF_{DIST}$ —measured system-to-object distance with the VL53xx ToF sensor;  $IR_{Compensated}$ —distance effect compensated IR pixel value. After applying the OEBT technique, data filtering and regression defined temperature offset as a function of distance, we get the data shown in Fig. 13. The plotted data shows that the application of the proposed methods over the raw IR imaging data allows for measuring variances in the contact measured temperatures through a non-contact manner with an accuracy of  $\pm 0.1 \text{ }^{\circ}\text{C}$ , independently from system-to-object distances and ambient conditions.



**Fig. 13** Distance data for IR\_Average and Corrected IR\_Average after applying the regression defined distance offset for Object A

## 6 Conclusion and Future Work

Within our work we presented a novel IoT thermographic platform for imaging live-stock tissues. We developed and utilized a ratiometric One-Eyed-Bandit Technique (OEBT), algorithms for object recognition, data filtering and dynamic distance-to-object corrections. The performed experiments and results can be summarized as follows:

- (1) the application of gradient-based automated object recognition and evaluating the minimum, maximum, average and median object temperatures is a feasible technique
- (2) the average and median IR object temperatures are more closely correlated to variances in contact measured temperatures than maximum and minimum IR data
- (3) applying a moving average filtering over the raw data improves accuracy and IR-to-contact temperature correlation
- (4) the application of the OEBT improves upon the IR-to-contact temperature correlation
- (5) measuring the system-to-object distance and applying a dynamic correction offset, based on a second degree polynomial regression model, makes the measured data less sensitive to distance effect attenuations
- (6) the application of the aforementioned methods and techniques can allow for the detection and measurement of contact temperature variations with an accuracy of  $\pm 0.1$  °C, using only non-contact IR data in a laboratory environment

The future goal of our work is to install such a bio-monitoring system in a real non-laboratory farm environment, capture data from different animals and evaluate their State-of-Health.

## References

1. Martello LS, da Luz e Silva S, da Costa Gomes R, da Silva Corte RRP, Leme PR (2016) Infrared thermography as a tool to evaluate body surface temperature and its relationship with feed efficiency in *Bos indicus* cattle in tropical conditions. *Int J Biometeorol* 60.1:173–181
2. Mota-Rojas D et al (2021) Clinical applications and factors involved in validating thermal windows used in infrared thermography in cattle and river buffalo to assess health and productivity. *Animals* 11(8):2247
3. Giro A et al (2019) Application of microchip and infrared thermography for monitoring body temperature of beef cattle kept on pasture. *J Therm Biol* 84:121–128
4. Montanholi YR, Odongo NE, Swanson KC, Schenkel FS, McBride BW, Miller SP (2008) Application of infrared thermography as an indicator of heat and methane production and its use in the study of skin temperature in response to physiological events in dairy cattle (*Bos taurus*). *J Therm Biol* 33(8):468–475
5. Salles MSV et al (2016) Mapping the body surface temperature of cattle by infrared thermography. *J Therm Biol* 62:63–69

6. Zhang K, Jiao L, Zhao X, Dong D (2016) An instantaneous approach for determining the infrared emissivity of swine surface and the influencing factors. *J Thermal Biol* 57:78–83. ISSN 0306-4565
7. Soroko M, Howell K (2018) Infrared thermography: current applications in equine medicine. *J Equine Veter Sci* 60:90–96. ISSN 0737-0806
8. Okada K, Takemura K, Sato S (2013) Investigation of various essential factors for optimum infrared thermography. *J Veter Med Sci* 13–0133
9. Schaefer AL et al (2007) The use of infrared thermography as an early indicator of bovine respiratory disease complex in calves. *Res Vet Sci* 83(3):376–384
10. McManus C et al (2016) Infrared thermography in animal production: an overview. *Comput Electron Agric* 123:10–16
11. Stewart M, Webster JR, Verkerk GA, Schaefer AL, Colyn JJ, Stafford KJ (2007) Non-invasive measurement of stress in dairy cows using infrared thermography. *Physiol Behav* 92(3):520–525
12. Fraden J (2016) *Handbook of modern sensors: physics, design and application*, 5th edn. Springer, pp 294–308

# The Digital Survival Game to Enhance the Digital Quotient of Lower Secondary Students



Amornphong Suksen and Nutteerat Pheeraphan

**Abstract** This research aims to study the results of a digital game with a student-centered approach to enhance the digital quotient of lower secondary school students. The sample group consisted of 133 lower secondary school students by multi-stage sampling from two classes of each level of the Assumption College English Program in the 2020 academic year. The research tools included (1) a digital game with a student-centered approach to enhance the digital quotient of lower secondary school students; and (2) a digital quotient test. The data were analyzed using the dependent sample's arithmetical mean, standard deviation, and a t-test. The research findings revealed the following: (1) the content and game design of a digital game with a student-centered approach had an Index of Item-Objective Congruence of 0.88 and 0.85 and an effectiveness Index of 0.28; and (2) the digital quotient achievement score was significantly higher than the pre-test scores at a level of 0.01.

**Keywords** Digital quotient · Digital intelligent · Digital game

## 1 Introduction

Stepping into the twenty-first century under globalization can significantly affect technology developed rapidly. Moreover, the advancement of technology has thoroughly affected the way of living in this society, applying technology in daily life, resulting in changes in lifestyle and society, communication, news presentation through various media, and electronic transactions.

Nowadays, learners inevitably assimilate into the digital world. Most children are online through various electronic devices such as computers, smartphones, and tablets, which causes the rapidly increasing number of digital citizens, to comply

---

A. Suksen · N. Pheeraphan (✉)  
Srinakharinwirot University, Bangkok, Thailand  
e-mail: [nutteerat@g.swu.ac.th](mailto:nutteerat@g.swu.ac.th)

A. Suksen  
e-mail: [amornphong.suksen@g.swu.ac.th](mailto:amornphong.suksen@g.swu.ac.th)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_23](https://doi.org/10.1007/978-981-99-3091-3_23)

293

with the Electronic Transactions Development Agency [1]. The results showed that Thais use social media such as Facebook, Instagram, Twitter, and Pantip for up to 3 h and 30 min per day. While watching streaming videos such as YouTube or Line TV, the average usage hour is 2 h 35 min per day. The chatting applications, such as Messenger and LINE, averaged 2 h. Playing Online games is 1 h 51 min per day and reading articles or books online is 1 h 31 min per day. Moreover, Internet users' usage behaviors showed that the users texting at 94.5%, hotel booking at 89.2%, booking or ticket purchases at 87.0%, made payments and services at 82.8%, and movie or music at 78.5%. Likewise, there are also many activities that Thai people are at risk of Invasion of privacy 45.34% of users do not change their password every three months, 45.04% shares a birthday through social media, 44.48% do transactions through Internet banking which do not notice whether it is safe or not, 43.36% open email and click to the unknown link and 35.70% upload photos or videos on social media immediately after shooting.

Thus, the trend of Thai students using electronic devices connecting to Internet networks has been increasing rapidly, which affects Thais stepping into the digital age world with unconscious adjustment to current social conditions. Teens apparently are addicted to social media, causing depression and attention deficit hyperactivity disorder [2] due to news reporting on television often deliver negative news, such as fraud, sexual offences, deception, money transfers for ordering products and not receiving products, fake news, impersonation of celebrity to deceive for taking property, and so on [3]. Additionally, social bullying on the Internet is also a widespread form of violence occurring in Thailand, such as blackening, using inappropriate words to others, and forwarding confidential information to damage others [4]. These are examples of a lack of Digital Literacy causing problems for oneself and others. If this problem occurs with children, it will result in children spending more time online than talking and interacting with people around and not being interested in studying. It is also a cause of a lack of communication skills, human relationships, ignoring the environment, affecting the physical condition like staying up late, bad eyesight, intense emotions, and depression.

Today, learners are inevitably integrated into the digital world. Most children are online through various electronic devices such as computers, smartphones, and tablets. They are causing the number of citizens in the digital world to increase rapidly from childhood to adulthood, changing people's daily lives today, entering the digital age society and becoming fully digital citizens. Without a digital quotient, it will have a negative impact on both you and society. To become a digital quotient (DQ), people should acquire an ability to live in the digital world consisting of 8 components, (1) Digital Identity, (2) Digital Use, (3) Digital Safety, (4) Digital Security, (5) Digital Emotional Intelligence, (6) Digital Communication, (7) Digital Literacy and (8) Digital Rights [5, 6]. So that users can use technology Digital media safely and efficiently. Moreover, they can survive and live in a digital world, including being responsible for one's actions and realizing the impact that will occur on others wisely.

Digital games are, therefore, suitable for presenting theory's content, skills, and behaviors. It encourages learners to learn by letting them play in situations where

there are a role, information, and rules of play. Besides, tackling with the situation game creators have defined as objectives will help learners apply their knowledge, skills, and attitudes from gameplay to solve problems and adapt into their life. In addition, the game has now changed its display using devices with various digital technologies such as computers, smartphones, and tablets and online games that use the Internet network to play, making the game format more exciting and attracting more players to play the game. Therefore, digital games have become modern media that can entertain, encourage learners to learn together, and equip them with a variety of experiences and knowledge to advance their critical thinking skills depending on the game's purpose specified by the creator. So, it encourages learners to learn by playing the game in situations will help learners to apply their knowledge, skills, and attitudes from gameplay to solve problems and adapt into their life.

In consonance with this information, researchers recognize the importance of cultivating ways of thinking, values, living, morality, ethics, and necessary skills to facilitate learners' lives in the digital world to boost immunity towards media and information literacy and to raise awareness of responsibility for one's actions in order to become a digital quotient person. Therefore, researchers aim to study a digital game with a student-centered approach to enhance the digital quotient of lower secondary school students to be able to live their life in the digital age world happily.

## **2 A Digital Game with a Student-Centered Approach**

Currently, the game has changed its display using devices such as computers, smartphones, tablets, and the Internet to play, which makes the game format more attractive and can attract more players, which can be called a "Digital Game". A definition of digital games entails their description as a wide variety of digital applications characteristics of the tools used to display, such as computer games. This game was created to be played on a computer. It can be divided according to the display characteristics into 2D games. Characters and Settings use flat images, and perspective can only be seen on the x-axis and y-axis, but the 3D game's perspective can be changed, [7]. An online Game is a video game that is either partially or primarily played through the Internet or any other computer network that offers online interactions with other players.

Therefore, a Digital game is a game that can be played through any technological device with rules that can arouse players' interest in acquiring all 4 Learning Readiness skills; physical, emotional and psychological, social, and intellectual.

## **2.1 *The Digital Game***

A classification of digital game elements is divided into two groups: gaming presentation design and user interface design.

### **2.1.1 The Gaming Presentation Design Has the Elements as Follows**

- (a) Goal is to determine what the player must do to win a game and the direction and purpose of the game by presenting it as a story or situation. If the player does not experience the game, the creator should have explained the situation to the player to understand the role.
- (b) Rules are a prescription guide for conduct, telling you how to proceed with your next course of action to achieve the goals.
- (c) Game interactions convey the player's actions to oneself. Also, there must be an obstacle or solution occurring within the game with the information the creator gives the player to make decisions.
- (d) Challenge, Obstacle, and competition are to encourage players or help them improve their playing ability, such as competing with themselves, competing with opponents, and competing with time.
- (e) Resolution enables players to understand the goals the creators want to communicate and examines whether the player understands the situation or their role.

### **2.1.2 The User Interface Design is Detailed as Follows**

- (a) The user interface is how a person interacts, presenting the situation with the story to the players and allowing players to enter the gameplay, such as the starting page and menu page.
- (b) Characters are the part that allows players to take on the role within the game to interact and play the game as specified by the creator. These include character movements that create the player's feelings to interact while playing. There should be the enemy, sectarian and stay alive to arouse interest and efforts of the players of the challenge in the game.
- (c) The setting is the environment that enhances a game by including the world atmosphere or situation that the creators want to convey and encouraging players to be a part of the game, such as background music and sound effects.
- (d) Rewards are things or keepsakes that players earn through gameplay. So, there should be a screen display showing the progress or players' development to motivate them to continue to develop their skills.

## 2.2 *Student-Centered Learning*

Facilitating student-centered teaching is a learning process that allows learners to learn through learning materials provided by the instructors to facilitate the learners. Therefore, teachers should have the knowledge, understanding, skills, and abilities to organize various activities to encourage students to have natural and full development under the learning principles by allowing them to take ownership of what they learn by the objectives. The researchers are interested in using games based on concepts that are essential to learners. As facilitating student-centered learning, teachers are required to use games as tools to make the lessons exciting and fun by allowing learners to experience, practice, make decisions, encourage learners to explore new knowledge by themselves, in addition to develop skills, understand apply lesson to real life in both the online and offline world.

## 3 Digital Quotient

Nowadays, the world has stepped into the digital world, and the world environment has rapidly changed. Most people spend time online using technology devices to facilitate, such as computers, smartphones, and tablets, to access the Internet. So, the digital has taken a vital role in society, affecting digital citizenship. However, using the technology has both advantages and disadvantages for users. Therefore, learners shall have the digital quotient to be prepared for living in the digital world happily.

Digital quotient has given the meaning that it is the ability to seek and use knowledge, including new skills in digital technology [8]. In accordance with Adams [9], Digital quotient is the ability to create and recognize information that affects the digital community. These capabilities form the basis of human values for respecting and caring for each other in the digital society and using technology wisely and responsibly. Damrongkiattisak [6], and Inthanon [10] mentioned that DQ is a comprehensive set of social abilities, emotions and perceptions that will enable us to face the challenges of digital life and be able to adapt to digital life. The digital quotient are the knowledge, skills, attitudes, and social value of the online world. In other words, it is media usage and social skills. To sum up, the digital quotient means the necessary ability to live and survive in the digital world.

Digital quotient is essential to enable learners to know how to manage their identity in the online and real world, their time, digital or computer evidence and personal information. In addition, it can also enable learners to deal with cyberbullying, protect themselves from danger on the digital world as well as acquire a well manner, Digital Literacy and Digital quotient. Therefore, the student can adopt the knowledge, skills, and attitudes to live their life in the digital age world happily.



In this study, the lesson level will be for lower secondary schools (Year 7–9). The learners have to learn to develop the digital quotient skill. The objectives of this study are as follows:

1. Digital Identity.
2. Digital Use.
3. Digital Safety.
4. Digital Security.
5. Digital Emotional Intelligent.
6. Digital Communication.
7. Digital Literacy.
8. Digital Rights.

## 4 The Digital Survival

The Digital Survival is a game using the student-centered approach to enhance the digital quotient concept by using the software package “RPG Maker MV” for building the RPG games of high quality and can be played online on the website; <http://the-digi.com>. Therefore, the students can access various devices such as smart-phones, tablets, and computers anywhere and at any time and use the subscription management system to keep the users’ information active.

There are eight levels that the researchers have defined as “The Digital Survival”, as detailed below.

1. Level 1 is called “Classroom”. This level has presented the idea of Digital Identity. The learning purpose of this level aims for the students to be able to tell and manage their online identity.
2. Level 2 is called the “Computer room”. This level presents the idea of Digital used. The learning purpose of this level is for the students to be able to talk about the use of digital devices and media quickly.
3. Level 3 is called the ‘Medical room’. This level presents the idea of Digital Safety. The learning purpose of this level aims for the students to be able to explain how to manage, prevent, and solve problems from accessing digital media content.
4. Level 4 is called the “Engineering room”. This level presents the idea of Digital Security. This level’s learning purpose informs the students on how to set and secure the password.
5. Level 5 is called the “Music room”. This level presents the idea of Digital Emotional Intelligent. The learning purpose of this level is to inform the students about good behaviors and politeness in communication via email chat and other online platforms.
6. Level 6 is called the “Dormitory”. This level presents the idea of Digital Communication. The learning purpose of this level is for students to know how to take action against inappropriate content.

7. Level 7 is called the “Science room”. This level presents the idea of Digital Literacy. The learning purpose of this level is for students to know how to select online resources correctly.
8. Level 8 is called “Canteen”. This level presents the idea of Digital Rights. The learning purpose of this level aims for students to tell the rights of digital media users.

The illustration shows the game structure using the student-centered approach to enhance the digital quotient concept, which consists of 2 parts.

### **Part 1: Gaming presentation**

1. Goals: Students will take the roles of characters in “The Digital Survival” to take responsibility for overcoming the enemies to rescue the school directors. The mission will be completed once you win all 8 levels.
2. Rules:
  - Players must register on the website to play the game; <http://the-digi.com>.
  - Players must overcome the enemy in each level.
  - The knowledge items the players will get after winning will be kept on the bookshelf.
  - The player stats will be increased from finding the knowledge items on the bookshelf.
  - Players must gain “STR” points (the strength stat) the flight.
  - Players must answer the question in each level correctly.
  - Players must answer correctly to reduce the enemy’s stat points.
  - If Players answer incorrectly, the player state will be deducted.
  - Players will loss when their stat points are decreased to “0”.
  - Players will win when the enemy stat points are decreased to “0”.
  - The game will be ended after winning all eight levels.
  - Players can start over by clicking “start new game” or “saving”.
3. Game interaction.
  - Players can create the character by selecting the gender and creating the character name.
  - Players can control the character’s movement.
  - Players can adjust the clothes, items, state, and talent of characters.
  - Players can explore various locations in the game.
  - Players will fight with the enemies by answering questions.
  - Players can “Save” and “Restart” the game.
4. Challenge, obstacle, and competition: The players will be completed it by themselves. The game’s challenge is that the player has to gain knowledge from reading to gain the stat. The obstacle of this game is that the Player has to fight with the enemies in each room to get the correct answer.
5. Resolution: After completing all eight events, the Player will receive an “Achievement of Digital Intelligence” trophy with their name.

## Part 2: Interface game design

1. The user interface is a game presentation available online on the website; <http://the-digi.com>. Therefore, students can access various devices such as smartphones, tablets, and computers.
2. Characters are for the Player to freely select their character.
3. Setting in the Digital Survival will be consisted of 2 parts.
  - Background will be changed in each level to change the Player's experience.
  - Background sound will be changed to change the players feeling, such as the rhythm of the sound.
  - Sound effects will be heard once a movement encourages players to be a part of the game.
  - Rewards will be given once the game end. The Player will receive a trophy of "Achievement of Digital Intelligence" with their name.

The results of developing "The Digital Survival" are as follows:

1. The index of Item–Objective Congruence (IOC) of The Digital Survival game design by five qualified is 0.85.
2. The index of Item–Objective Congruence (IOC) of The Digital Survival game contents by five qualified is 0.88.
3. The Effectiveness Index of The Digital Survival from 30 samples shows that the Effectiveness Index of the Digital Survival is 0.28.

## 5 Methodology

This study is a research and development, which aims to develop digital games based on concepts of the facilitating student-centered to improve the digital quotient of the lower secondary school (Year 7–9) and to study the results of using digital games by dividing into 2 phases. Phase 1; the development of digital games based on the facilitating student-centered to improve the digital quotient of students, and Phase 2; the study of digital games is based on facilitating student-centered to improve the digital quotient of students.

### ***5.1 The Development of Digital Games Based on the Facilitating Student-Centered to Improve the Digital Quotient of Students***

This research aims to study developing the digital quotient skills, which consist of 3 types of learned behaviors: knowledge, skill, and attitude. The DQ test has been designed to capture a knowledge and skills across all three types of learned behaviors. There are alternate tests; Multiple choice, Open-ended Questions, and Rating Scale.

Therefore, the researchers used the DQ test as a subjective and objective test in this research.

The researchers have developed the game called “The Digital Survival” by using games following the facilitating student-centered to advance the digital quotient concept. The program used is RPG Maker MV, the Software package for making RPF games of high quality and can be played online on the website; <http://the-digi.com>. Therefore, the students can access various devices such as smartphones, tablets, and computers anywhere and at any time. A subscription management system is used to keep the user’s information active.

The result of the lower secondary students (Year 7–9) using the game “The Digital Survival” five qualified people shows that the game had a content quality index of 0.88, a design index of 0.85, and a practical index of 0.28.

The researchers use “The Digital Survival” with the sample in this study is students at the lower secondary school (Year 7–9), two sections in each level, six sections in total. There are 113 samples as follows: the lower secondary school (Year 7) 42 people, the lower secondary school (Year 8) 39 people, and the lower secondary school (Year 9) 32 people by using the multi-stage sampling.

## ***5.2 The Study of Digital Games is Based on Facilitating Student-Centered to Improve the Digital Quotient of Students***

Researchers have collected data from the digital games based on digital games bringing a student-centered approach to enhance the digital quotient concept to experiment with target groups as the following steps;

1. The researcher coordinates with the teacher to inform the research purpose, procedures, and processes. To use digital games based on a student-centered approach to enhance the digital quotient in learning to measure and evaluate student participation in activities and the teacher’s role in the learning process, including consulting and guidance during the class.
2. The researchers conducted a trial of digital games based on a student-centered approach to enhance the digital quotient on the device before the experiment to check the game availability and internet connections.
3. The researchers organize student orientation in the classroom to clarify the students’ understanding of playing digital games based on a student-centered approach concept and clarify the role of the learners in the learning process.
4. The researchers evaluated the digital quotient of the learners before the experiment by allowing students to test their digital quotient before playing the game.
5. Students learn from digital games by themselves for one week. In this regard, the researcher provided orientation to students and instructors before playing the game by having the teacher observe and ask the learners periodically during the

class in a week with the contact channel asking questions that may arise during the game.

6. The researcher tested the digital quotient of the learners with a digital quotient test and interviewed students after using the game, and found that.

“The first student is a female playing the game on the Mac OS. The game can be played smoothly. The content of the game is fun and easy to continuously play on the computer with spending time for 1 hour”.

“The second student is a male playing a game on Windows 10 that can play smoothly. However, the game content is quite a lot to play. It must save and come back to play next time. On the other hand, the game’s story is fun. Moreover, the game characters should be able to customize their appearance, color, skin, and hairstyle”.

“The third student is a male playing a game on an iPad. The character can easily control. However, the rewards are too much, so the game is too easy. Therefore, the levels can be completed only once time by spending an hour playing”.

7. Using the results of student assessments with digital quotient before and after using the game to score and comparing the differences with an average score.
8. Conclusion and Discussion.

## 6 Findings

This study compares differences between digital quotient before and after using games following the facilitating student-centered by using the Digital Quotient Test (DQ test). There are 24 questions (24 points). The sample will be the lower secondary student (Year 7–9).

Table 1 shows the results of the lower secondary students (Year 7) after using games following the facilitating student-centered to advance the digital quotient is statistically significantly higher levels of digital quotient than before using the game, at level 0.1.

Table 2 shows the results of the lower secondary students (Year 8) after using games following the facilitating student-centered to advance the digital quotient is statistically significantly higher levels of digital quotient than before using the game, at level 0.1.

Table 3 shows the results of the lower secondary students (Year 9) after using games following the facilitating student-centered to advance the digital quotient is

**Table 1** Comparison of differences between digital quotient before and after using games at the lower secondary students level (Year 7)

Phrase	<i>N</i>	$\bar{x}$	SD	<i>t</i>	<i>p</i>
Before using games	42	10.60	3.269	25.863	0.00*
After using games	42	15.69	3.024	34.955	

\*  $p < 0.01$

**Table 2** Comparison of differences between digital quotient before and after using games at the lower secondary students level (Year 8)

Phrase	<i>N</i>	$\bar{x}$	SD	<i>t</i>	<i>p</i>
Before using games	39	13.41	2.807	29.834	0.00*
After using games	39	16.92	2.669	39.592	

\*  $p < 0.01$ **Table 3** Comparison of differences between digital quotient before and after using games at the lower secondary students level (Year 9)

Phrase	<i>N</i>	$\bar{x}$	SD	<i>t</i>	<i>p</i>
Before using games	32	13.88	3.035	25.863	0.00*
After using games	32	17.28	2.797	34.955	

\*  $p < 0.01$ 

statistically significantly higher levels of digital quotient than before using the game, at level 0.1.

In conclusion, after using the game “The Digital Survival”, the lower secondary students (Year 7–9) have statistically significantly higher levels of digital quotient than before, at level 0.1.

## 7 Conclusion

In this study,

1. The Digital Survival game was developed by following the facilitating student-centered for improving the digital quotient in lower secondary students (Year 7–9) was found that the game had a content quality index of 0.88, the design index of 0.85, and had an effectiveness index of 0.28.
2. The result of the lower secondary students (Year 7–9) using the game “The Digital Survival” shows that students have significantly higher levels of digital quotient than before, at level 0.1.

## 8 Discussion and Suggestion

### 8.1 Discussion

1. The Digital Survival Game delivers a good quality of content, design and practicality for the following reasons;
  - a. the researchers have selected the content about the digital quotient because the world recently has been stepping into the Digital age, which causes significant changes in the environment rapidly. The world has stepped into the digital world, and the world environment has rapidly changed. Most people spend time online using technology devices to facilitate, such as computers, smartphones, and tablets, to access the Internet. So, Digital has taken a role in society, affecting digital citizenship. However, using the technology has both advantages and disadvantages for users. Therefore, learners shall have the digital quotient to be prepared to live in the digital world happily. In this regard, the researchers applied and adapted the digital quotient concept within the game's content by using technology in learning and working to solve the problems efficiently, knowingly, and ethically. Hence, it shall be class learning material. Netwong [11] has been studied about developing digital citizenship and Factors Affecting Learning, the results show that the students' digital citizenship and academic achievement significantly improved using the students' e-Learning of the Dusit Rajabhat University technology program. In addition, bringing a student-centered approach to enhance the digital quotient concept into digital games causes various learning management methods. It allows students to find new knowledge themselves by providing a direct experience. So, students can improve their thinking skills and self-determination. Teachers will be the ones who facilitate students in the learning space, co-thinking, and sharing their creativity with building good relationships with students, which can be learned anywhere and all time.
  - b. The researchers selected the role-playing game through the student-centered approach concept. Students will learn by experiencing direct interaction while playing the game before designing the game. The RPG Maker MV was chosen to create the game because the program can create the digital game based on the student-centered approach concept. It is also allowed the Player to be a character and control the character in the game, which is the current trending game which can be played online on the website; <http://the-digi.com>. Therefore, the students can access various devices such as smartphones, tablets, and computers anywhere and at any time by following the student-centered approach concept.
2. Using the digital survival game helps advance higher digital quotient of students than prior to using the game following to research assumptions. The data from the research shows that students have improved the student-centered approach to enhance the digital quotient. This is because learning management by adopting

the student-centered concepts will allow students to use a learning management system in creative ways. Bringing a student-centered approach to enhance the digital quotient concept into digital games allows students to achieve the study purpose. The discussion will result from the Player's data, following the rules, presenting the content, behaviors, how players play the game and the result of players playing the game at the end. Digital games will help motivate learners to learn by allowing the Player to be a character, information, and rules with experiencing the environment according to study purpose, which is therefore suitable for presenting theory content, skills, and behaviors. This will help the student apply the knowledge to their lives. Making digital games by bringing a student-centered approach to enhance the digital quotient concept into digital games based on the concept will have to design and plan the story situation, the role, and rules to make it feel realistic once playing the character. Using the interface design, reward background, and sound will help the players continue focusing on playing the game to achieve the target and approach the players feeling with the game, which can be played everywhere and all time. The findings of this study are consistent with the results of Valerie [12]. He studied learning management to enhance digital citizenship. It found that by playing the Minecraft game, students could achieve the learning objectives of Digital Citizenship. Pansuwan [3] studied the performance with role-playing computer games, and story-line teaching of the secondary students grade 3 shows that the students have academic achievement, have a higher ability to analytical thinking and have a higher grade than before using the game.

The research shows that bringing a student-centered approach to enhance the digital quotient concept into digital games can genuinely improve the digital quotient of lower secondary students, creating learning innovation following the National Thailand Education Plan. The Ministry of Education in Thailand [13] has laid out the education management target framework in developing the capacity of Thais of all ages to reach their full potential, lifelong learning, Digital Literacy, responsible for their action and realize affecting others wisely.

## ***8.2 The Suggestion of the Use of Research Result***

1. The student-centered approach to enhance the digital quotient concept for lower secondary students developing process has factors involved in both the development process and the determination of quality and efficiency. From the study, there are additional suggestions for developing digital games based on the student-centered approach to enhance the digital quotient concept for better usability as follows: (1) Digital games based on the student-centered approach to enhance the digital quotient concept for lower secondary students have some limitations because the game is available online on the website <http://the-digi.com> which must always use on the network. Therefore, the game is not available for all target groups. In the future, to be able to use Offline, digital games



may be developed and available for download online. (2) Finding digital games based on the student-centered approach to enhance the digital quotient concept for lower secondary students by experts helps evaluate the quality of the game before developing and using it. The researchers have suggested that the evaluation results be used to improve digital games based on the student-centered approach. The researchers did not just look at the overall results; the experts' suggestions have been counted to improve and develop the game. For example, one expert suggests improving the fight scenes to a more colorful and exciting. However, the limitations of the RPG Maker program MV, which cannot develop the battle scenes to be more colorful and exciting. In addition, the limitation of the RPG Maker MV program is that it is a foreign program, therefore resulting in size and font style within the game, especially the Thai alphabet format, which appears to be some incorrect vowels and tonal displays. The researcher then chose to use the font format that can display the message correctly, which may affect the size of the characters in the text box that cannot be adjusted on different devices, as suggested by the experts. In this regard, the researchers consider the users. By finding that the learner can use, but the characters may be too small when used on devices with screen sizes less than 6.1 inches.

2. From the comparison result of the student-centered approach to enhance the digital quotient concept after using the game, the researchers found differences in test scores before and after studying. In testing, a group of students with a low pre-test score is more likely to develop digital quotient than students with a high pre-test score. From the study, a game base on the student-centered approach to enhance the digital quotient concept is a key to advancing the digital quotient for lower secondary students. The game can improve the digital quotient of all learners, but it will be seen in the group with a lower academic performance. Therefore, the researchers suggested ways to apply digital games based on student concepts as follows: (1) In the study, while playing games, learners contribute to learning; in the class, the learners interact, exchanging opinions and sharing experiences playing digital games. A good learner is a key to helping other learners to learn together, which helps to keep them entertained while playing digital games. (2) The teachers should be the one which encourages students to learn after the teachers have assigned students to learn by themselves from the student-centered approach to enhance the digital quotient concept for one week. Teachers may reinforce by scoring the learners, providing advice and assisting students who have limited in finding devices. Besides, teachers should inquire about the student's progress during the week to check the improvement and prevent over-screen time, completed at one time on the date specified by the instructor, and (3) The digital learning environment, including play devices and the Internet network, plays an essential role in using student-centered digital games. The game will not be available without devices, Internet networks and web browsers. Students are bringing their own devices, such as smartphones, tablets or computers, to play the game using Google Chrome and Safari. It reduces the time spent learning how to use the device because learners are already familiar with their own devices. In addition, the Internet network is another important

factor because without an Internet connection, students will not be able to access the game, and the Internet signal should be 1 Mbps or higher. Lower than this may affect the content download time.

### **8.3 The Suggestion for Future Research**

1. In this research, the researchers develop a digital game bringing a student-centered approach to enhance the digital quotient of lower secondary students. The study aims to advance the digital quotient. It should have other factors that influence the use of digital games based on a student-centered approach concept to enhance the digital quotient; gender, education level, time spent, the readiness of students and teachers, devices, and environment.
2. The experimentation of digital games bringing a student-centered approach to enhance the digital quotient concept of lower secondary students does not distinguish learners. Therefore, the result of digital games based on a student-centered concept should be studied to use with a learner who has a different learning ability to find ways of using digital games based on bringing a student-centered approach to enhance the digital quotient concept in other areas.
3. The digital game in this research brings a student-centered concept presented in inductive content by allowing students to study in various ways within games. It causes the game less challenging. Therefore, the content should be presented as deductive to make more significant efforts to overcome obstacles within digital games.
4. The study of the use of a digital game bringing a student-centered approach to enhance the digital quotient is the study and evaluation of the digital quotient of learners after the game was created. So, it has only one week to study. However, the results show that the learners have developed. In this regard, the persistence of digital learners' quotient should be evaluated or used to observe behavior. In the following research, observation and behavior should be added after the use of digital games by increasing the period to the first semester so that learners are digitally intelligence and able to apply effectively in their daily lives.

### **References**

1. Electronic Transactions Development Agency (EDTA) (2018) Internet user behavior in 2018, Thai people use the internet for 10 hours and 5 minutes per day. <https://www.eta.or.th/>. Accessed 01 Oct 2018
2. Suwanpho M (2018) Teenagers are addicted to social media. <https://new.camri.go.th/infographic/106>. Accessed 01 Oct 2018
3. Pansuwan N (2012) The effects of learning by using role-playing computer games and storyline teaching method of mathayom suksa 3 students towards learning achievement and analytical thinking ability. *Veridian E-J Silpakorn Univ* 5(2):538–553

4. Charoenvanich S (2017) Cyberbullying: impact and prevention in adolescents. *Sci Technol J* 25(4):639–648
5. Wannapiroon P (2018) Digital intelligence. *Tech Educ J King Mongkut's Univ Technol North Bangkok* 13–15
6. Dumrongkiattisak W (2018) Digital quotient. [http://cclickthailand.com/wp-content/uploads/2020/04/dq\\_FINAL.pdf](http://cclickthailand.com/wp-content/uploads/2020/04/dq_FINAL.pdf). Accessed 01 Oct 2018
7. Sa-ad SI (2014) Games and simulations in education. Ramkhamhaeng University Press, Bangkok
8. Waller S (2018) What is digital intelligent. <https://simonwaller.com.au/wp-content/uploads/2014/09/Digital-Intelligence-white-paper.pdf/>. Accessed 01 Oct 2018
9. Adams NB (2004) Digital intelligent fostered by technology. *J Technol Studi* 30(2):93–97
10. Inthanon S (2018) DQ: digital intelligence quotient. Foundation for the Promotion of Children and Youth Media, Bangkok
11. Netwong T (2014) Development of digital citizenship and learning achievement utilizing e-learning in information technology of students Suan Dusit Rajabhat University. *Tech Educ J King Mongkut's Univ Technol North Bangkok* 5(1):73–80
12. Valerie H (2015) Digital citizenship through game design in Minecraft. *New Library World* 116:369–382
13. The Ministry of Education in Thailand (2018) Education plan in 2017–2032. <http://www.lampang.go.th/public60/EducationPlan2.pdf/>. Accessed 01 Oct 2018

# An Experimental Analysis of Benchmarking Tools for Smart Contract-Based Blockchain Application



Deepa Kumari, Chirag Jain, Aman Saxena, Pranjal Gupta, Ashay Netke, and Subhrakanta Panda

**Abstract** Blockchain is an emerging technology that can merge with various business sectors. However, the blockchain is nevertheless in the early stage, and there are concerns regarding its acceptability. One of the primary bottlenecks is the low-performance issue. Thus, it is necessary to assess the multi-blockchain ecosystem's performance in different use case scenarios. This paper presents experimental insights on the smart contract-based blockchain using Hyperledger caliper and manual benchmarking. It monitors and simulates benchmarking tools for varying workloads. Thus, this work analyzes Hyperledger caliper and manual benchmarking on scalable and adaptable networks. It also presents a graphical analysis of the performance evaluation metrics such as transaction rate, throughput, and average read or write latency. It also gives a proper understanding of the relative parameters such as success rate, availability, consistency, and scalability. The proposed approaches will help researchers benchmark and make correct decisions in developing blockchain systems.

**Keywords** Benchmark · Blockchain · Hyperledger caliper · Performance evaluation · Smart contract

## 1 Introduction

Benchmarking [10] is not the measurement itself but a process of establishing gaps in performance and, as such, ensuring that an action plan is put in place to close identified gaps and verify continuous improvement [1]. These benchmarks can be the system's response time, latency, or measure-specific parameters such as the time to write a block to persistent storage [11]. However, a good process may have limitations [12] as the improvement may not be acceptable for setting high competitive standards. This paper makes the following contributions:

---

D. Kumari (✉) · C. Jain · A. Saxena · P. Gupta · A. Netke · S. Panda  
BITS-Pilani Hyderabad Campus, Secunderabad, India  
e-mail: [p20190020@hyderabad.bits-pilani.ac.in](mailto:p20190020@hyderabad.bits-pilani.ac.in)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_24](https://doi.org/10.1007/978-981-99-3091-3_24)

309

- It presents benchmarking on the developed blockchain application based on smart contract functionalities.
- To the best of our knowledge, there exists no literature on the evaluation of manual benchmarking. Manual benchmarking reduces the cost of development compared to existing tools by providing comprehensive, definitive, and unbiased results.
- It suggests the steps for installation and configuration for both benchmarking applications in a more user-friendly way.
- The experimental work presents the impact of varying transaction numbers, transaction rates, and transaction types on blockchain performance and gives insight into the use of both Hyperledger caliper and manual benchmarking.

The structure of the remaining sections of this paper is as follows: Section 2 gives a brief overview of the benchmark parameters. Section 3 shows the adopted methodology and discusses the experimental results. Section 4 concludes the work with insights to future work.

## 2 Blockchain Benchmarking Parameters

Blockchain is a data structure where transactions occur between nodes on a peer-to-peer (P2P) network [5, 6]. Fan et al. [3] highlight a blockchain system comprising five identified abstraction layers: application, execution, data, consensus, and network. Figure 1 shows every layer’s influence on performance evaluation. These layers are essential to preserving blockchain’s features, such as scalability and decentralization, and ensuring security. The topmost layer, the application layer, contains scripts, algorithms, and smart contracts. The smart contract can automatically trigger execution when the constraints are satisfied; otherwise, the contract is canceled. The primary function is to call the interface of the intelligent contract layer and deploy the blockchain applications such as Ethereum and Hyperledger fabric. The benchmarking process in this paper gives more weight to the application layer due to the trade-offs between scalability and decentralization.

---

### Algorithm 1 Server’s Process

---

**Require:** set up environment and load smart contract

**Ensure:** Transactions

```
1: connect to peers
2: elect leader = maintain secure connection
3: while server is up do
4:   Execute transactions
5:   if connection is interrupted then
6:     goto 2
7:   end if
8: end while
```

---

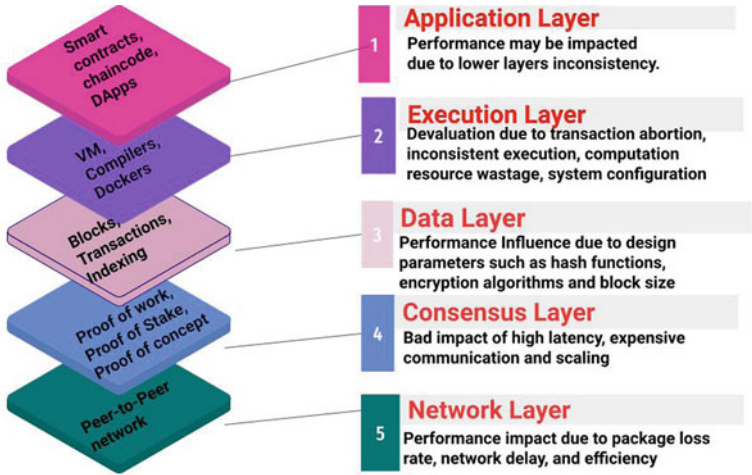


Fig. 1 Influence of layers of blockchain on performance evaluation

Benchmarking evaluates performance in terms of *transaction throughputs*, *latency*, *transaction response time*, *transaction success*, and *failure* between client and server [3]. There are two processes involved in benchmarking process server and client. Algorithm 1 gives the pseudo-code of the steps executed on the server side, and Algorithm 2 gives the steps on the client side [2]. Firstly, it is required to load the application on a deployed and configured network of interconnected servers. Secondly, the client side sends transactions at a higher rate to servers to measure their capacity for that period. Table 1 defines several standard metrics that apply to blockchains [10] and their mathematical formula wherever appropriate. Here, performance metric like throughput gets higher priority over bandwidth as a network performance benchmark. Throughput signifies the number of successfully committed transactions. Similarly, latency is an essential metric because it helps understand transaction rates and, correspondingly, any constant delays.

3 Methodology

This work uses a Python interface to interact with the local node by Web3.py and the brownie framework for the script. Web3.py is more interacting and readable with Ethereum. It is commonly found in decentralized apps (dApps) to help send transactions, interact with smart contracts, read block data, and various other use cases. Table 2 illustrates different use cases of the smart contract considered for testing performance bottlenecks.

The proposed approach implements Brownie, a Python integrated framework, for testing the smart contract functionality. Initially, it set ups the new task using Brownie

**Algorithm 2** Client's Process**Require:**  $n$  : number of TX,  $t$  : list < number of threads >,  $s$  : list < request size >**Ensure:**  $p$  : list < other parameter >, Output

```

1: Set up environment and load application
2: for each  $i$  in  $t$  do
3:   create  $i$  threads
4:   for each thread do
5:     for each size in  $s$  do
6:       for each  $i$  in  $[0..n/i]$  do
7:         start timer
8:         send TX to proxy to deliver to server's process
9:       end for
10:    end for
11:  end for
12:  Evaluate Performance Metrics and Benchmark
13:  Output= Benchmark Results
14: end for
15: Return Output

```

**Table 1** Benchmarking parameters

Metrics	Formulae	Description
Read latency	(Response received time) – (Submit time)	Measures time between when the read request is submitted and when the reply is received
Read throughput	(Total read operations)/(Total time in seconds)	Measures of how many read operations are completed in a defined time period, expressed as reads per second (RPS)
Transaction latency (write latency)	(Confirmation time at network threshold) – (Submit time)	Measures network-wide view of the amount of time taken for a transaction's effect to be usable across the network
Transaction throughput (write throughput)	(Total committed transactions)/(Total time at committed nodes)	Measures rate at which valid transactions are committed by the blockchain system under test in a defined time period

init. It adds smart contract code (Records\_Keeper.sol) to the contract folder. Table 2 depicts workload files for the Record\_Keeper smart contract, which is electronic healthcare record keeper, and its descriptions [4]. Then, the workload file is compiled by the brownie compiler that would be deployed as a script. In the main deploy script, def main is Brownie's entry point. So, Brownie defaults to a development ganache chain. As this work belongs to local ganache, it would set up the account as accounts[0], which is Brownie's default ganache account. It is also needed to import libraries from accounts, config, network, chain, Web3, and gas price. The compiled code is deployed to check all the smart contract functionalities for multiple

**Table 2** Record\_Keeper smart contract functionalities and its descriptions

Smart contract functions	Descriptions
getName(string memory)	To display the patient name
getAge(uint)	To display the age of the patient
getDocList(address[] memory)	To display list of the doctor
getImage(uint n)	To display medical image reports
setName(string memory n)	To enter patient name
setAge(uint n)	To enter patient age
Addtreatment(string memory disease, string memory treat, string memory med)	To enter patient disease, treatment, and medicine details

transactions. The record keeper checks whether the transaction has been accepted or not for every transaction. Finally, it includes the driver function to calculate various values observed in the process for each parameter.

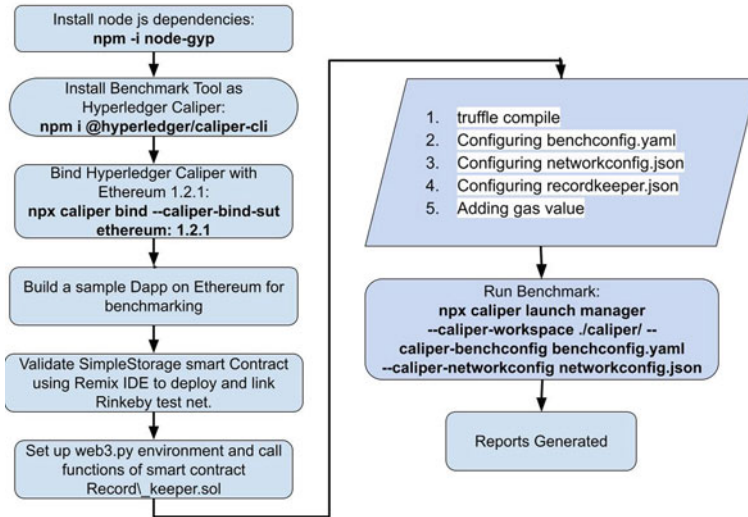
**3.1 Benchmarking with Hyperledger Caliper**

Hyperledger caliper [1] is a benchmark tool for the blockchain framework that generates a performance report for a blockchain with predefined use cases. Other existing tools, such as MixBytes Tank [9] and Whiteblock Genesis [8], also exist. Most of the existing tools resemble caliper in terms of functionality, such as setting numerous nodes, conducting the initial configuration process, executing the critical metrics, and closing the testing network. Hence, this paper chooses Hyperledger caliper for experimental analysis due to its wide acceptance and similar functionalities.

The experiments on benchmarking with caliper performed in system architecture consist of 8 GB RAM and 1 TB Hard disk, with its processor as Intel(R) Core i5-5200 CPU frequency of 2.20 GHz, 2195MHz, 2 cores, 4 logical processors. Figure 2 represents the installation and configuration steps for the Hyperledger caliper tool. This benchmarking tool, with some predefined use cases, generates a report that contains several performance indicators to serve as a reference when using the Ethereum blockchain solutions. It includes several performance indicators such as transactions per second and transaction latency. This paper aims to observe the caliper result generated from the user’s specific record keeper smart contract application.

Similarly, blockchain solution supports smart contract applications that measure their complexity by a performance indicator such as transaction per second (TPS). Hence, benchmarks can use different trial smart contracts that inadvertently affect the results. So, good use cases of smart contracts help understand the blockchain and performance indicators.





**Fig. 2** Installation and configuration steps for Hyperledger caliper

### 3.2 Manual Benchmarking

The flowchart in Fig. 3 explains the configuration steps of manual benchmarking. After selecting the parameters for the performance evaluation, it comprises the following steps to evaluate the parameter.

1. User Operation: A set of  $N$  (1, 50, 100, 200, 400, 600) signed transactions, i.e.,  $TX = TX_1, TX_2, TX_3, \dots, TX_N$ .
2. User Operation: The user node sends the created transactions to an Ethereum-based GANACHE blockchain and captures the current time point  $t$ .
3. Blockchain's Internal Process: The blockchain's P2P network distributes the transactions to miners. The mining process (PoW consensus algorithms) will confirm the valid transactions and broadcast them to all the nodes in Ethereum.
4. User Operation: The user node will continuously query the confirmation of each transaction through `web3.eth.wait_for_transaction_receipt()` functionality of the w3 module.
5. User Operation: When all the transactions  $TX_n \in TX$  are confirmed, the user node captures the current time  $T$  then calculates and records the transaction time of  $TX_n$ ,  $\Delta t_n = T - t$ . Here,  $TX_n$  is the transaction time difference between total transaction time  $T$  and current time  $t$ .
6. User Operation: If no transactions are confirmed, GOTO Step 4.

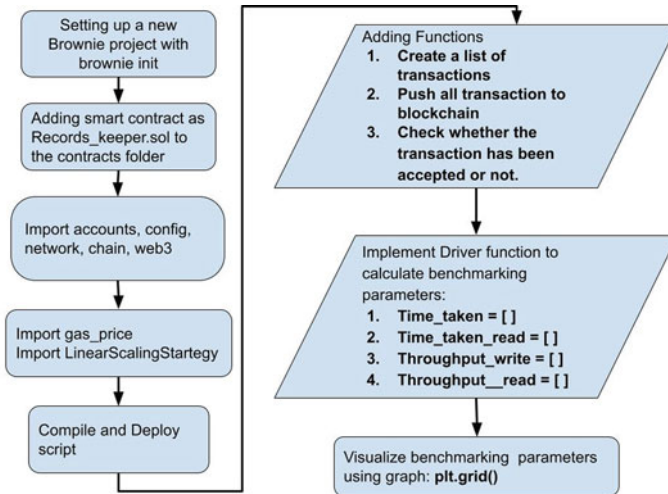


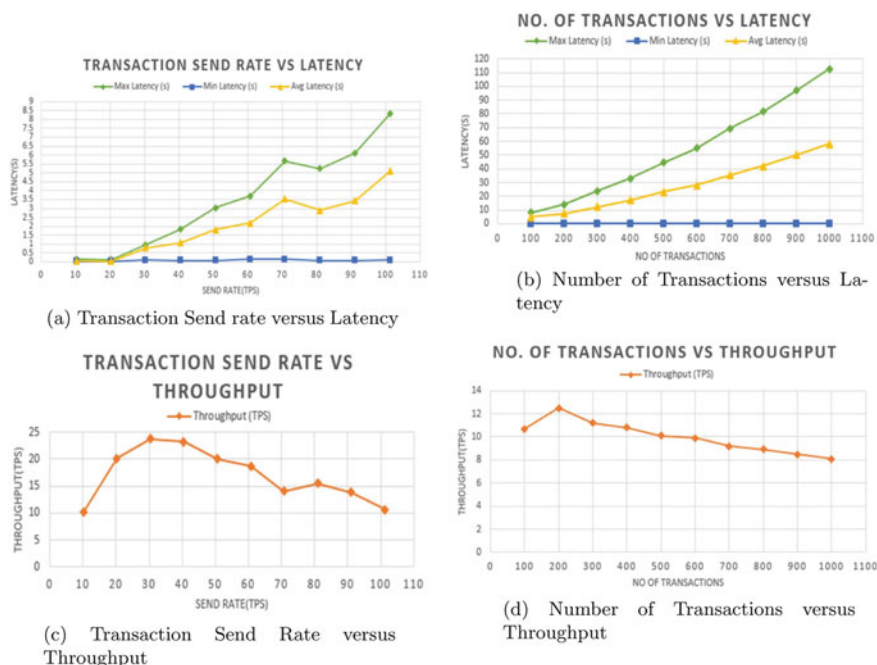
Fig. 3 Configuration steps for manual benchmarking

### 3.3 Benchmarking Results

This paper has evaluated parameters such as transaction send rate, read or write latency, and throughput. The Hyperledger caliper and manual benchmarking algorithm are applied on the same smart contract code and inferred similar behavior of graphic patterns from both benchmarking tools. Performance assessment summarizes the experimental benchmarking results on Hyperledger caliper and manual benchmarking applications.

#### *Performance Assessment of Hyperledger Caliper Tool*

- Hit rate: The experimental result indicates 100% success on 100–1000 simultaneous transactions on the same number of nodes (10 nodes taken). They analyze the execution of sent transactions dynamically over time. These results are pretty similar to other machine configurations.
- Average latency: Fig. 4a reflects that min, max, and average latency are at 0 till the transaction rate is 20. Thus, there was no delay with 20 transactions. But, there is uneven growth in latency after the transaction rate exceeds 20. Similarly, Fig. 4b depicts continuous growth in the average latency as the number of transactions increases. It means there will be growth in delay with more transactions.
- Throughput: Fig. 4c depicts the transaction throughput increases on 10–20 simultaneous transactions. It depicts the consistency and reliability of the given blockchain solution. But, after the transaction rate exceeds 20, throughput decreases due to delay (higher latency) in the transaction. Similarly, Fig. 4d represents that as the number of transactions increases to 200, there is a small growth in throughput. Still, throughput decreases after a continuous increment in the number of transactions.



**Fig. 4** Graphical analysis of benchmarking with Hyperledger caliper on smart contract-based blockchain application

- **Scalability:** The experiments performed on several nodes that can dynamically change over time. Figure 4 represents the analysis of transaction send rate concerning throughput and latency. It depicts the transfer of sending data in a specified amount of time to check the delivered throughput.

### **Performance Assessment of Manual Benchmarking**

- **Hit rate:** The experimental result indicates 100% success on 1–800 simultaneous transactions on the same number of nodes (10 taken). They analyze the execution of sent transactions dynamically over time. These results are pretty similar to other machine configurations also.
- **Average latency:** Fig. 5a and b depicts that there is a continuous growth in both read and write latency as the number of transactions increases. Also, both figures show a sudden read and write latency growth after 400 transactions. It means there will be somewhat growth in delay with more transactions.
- **Throughput:** Fig. 5c and d depicts the transaction throughput on 1–800 simultaneous transactions. Both figures show that as the number of transactions increases to 150–200, there is a big growth in throughput for both read and write processes. It depicts the consistency and reliability of the given blockchain solution.

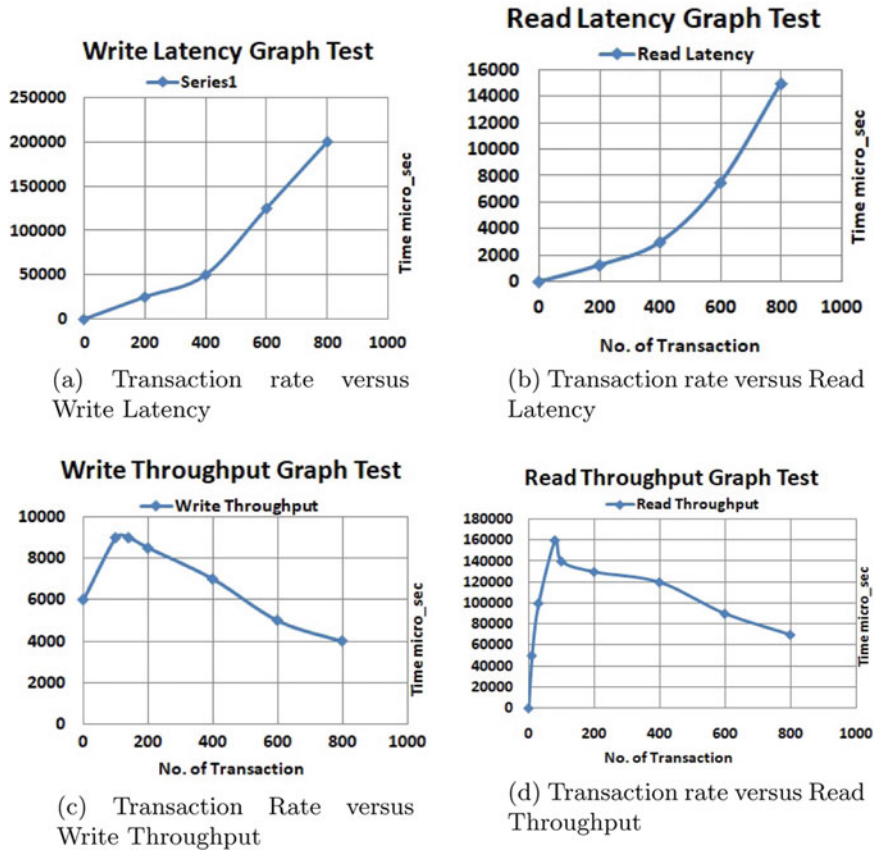


Fig. 5 Graphical analysis of manual benchmarking on smart contract-based blockchain application

- **Scalability:** The experimental results are not constrained to a fixed number of nodes. These dynamically add 200 new transactions over some time on 10 nodes, which can vary too. Figure 5 represents the analysis of transaction send rate concerning throughput and latency. It depicts the transfer of sending data in a specified amount of time to check the delivered throughput.

Hence, this paper can use the best Hyperledger caliper and manual benchmarking on the same smart contract functionalities. It shows that the benchmarking parameters give the ability to make effective decisions. In the case of Hyperledger caliper, the authors represent transaction throughput/latency instead of read or write throughput/latency. It is because reading operations are less engaging than write operations in Caliper [7] as its read operation is completed entirely on the host and does not involve any consensus mechanism. In contrast, the write operation is spread widely across the

**Table 3** Comparative analysis of Hyperledger caliper and manual benchmarking

Scenarios	Hyperledger caliper	Manual benchmarking
Platform	Modular blockchain platform	Generic blockchain platform
Ease of installation	No (software compatibility issues and lesser user-interface)	Yes (customized configurable compiled codes)
Smart contract code	Golang, Java on Linux foundations	Solidity on Ethereum virtual machine
Pros and cons	Adaptor-based framework, scalable (fabric adapter utilizing common connection profile (CCP) feature of the fabric SDK to provide compatibility); no predefined workload design; but support more workloads	Use of driver function in compiled code (software component that provides an interface for a physical or virtual device), scalable, carefully designed workloads but they are constant

blockchain nodes and thus requires consensus algorithms. At the same time, Manual benchmarking gives equal weightage to both read and write transactions.

Figure 4 also shows that transaction arrives at 10 s for Hyperledger caliper, whereas Fig. 5 depicts that transaction arrives at 0 s itself. Thus, in the case of manual benchmarking, systems stability is the transaction arrival rate, meaning the rate at which transactions reach the Blockchain but not necessarily the rate the system can process. Hence, it gives lesser timeouts and little more throughput than Caliper.

Also, Table 3 compares the performance of benchmarking applications in different scenarios. A blockchain system works with its initial scripts for making node accounts and prerequisites for the test (such as validating consensus errors may cause changing system parameters). These modifications are served differently in different blockchains, so manual benchmarking makes it easier for users to slowly adapt testing and benchmarking for their products to improve testing procedures. Whereas, the Hyperledger caliper is a little challenging to install due to its compatibility issues with the Ethereum version.

## 4 Conclusion and Future Work

Blockchain platforms are rapidly evolving and adapting by design and implementation; thus, reproducibility and usability have become complicated. Further, researchers are clumsy in establishing the network and deploying the system under tests. Thus, this paper addressed these issues in a user-friendly way. It also performed experimental analysis on Hyperledger caliper and manual benchmarking for performance measurement. It demonstrated the characteristics and operating principles behind these benchmarking frameworks. It presented a graphical analysis of

benchmark parameters such as latency, throughput, and transaction send rate. The benchmarking results conclude the performance assessment regarding success rate, availability, consistency, and scalability.

In future, authors will also work on other blockchain platforms with different numbers of transaction nodes. Further, the authors intend to observe the influence of varying consensus algorithms on variable workloads.

## References

1. Ahmad A, Saad M, Kim J, Nyang D, Mohaisen D (2021) Performance evaluation of consensus protocols in blockchain-based audit systems. In: 2021 international conference on information networking (ICOIN). IEEE, pp 654–656
2. Dong Z, Zheng E, Choon Y, Zomaya AY (2019) Dagbench: a performance evaluation framework for DAG distributed ledgers. In: 2019 IEEE 12th international conference on cloud computing (CLOUD). IEEE, pp 264–271
3. Fan C, Ghaemi S, Khazaei H, Musilek P (2020) Performance evaluation of blockchain systems: a systematic survey. IEEE Access 8:126927–126950
4. Kumari D, Rajita BSAS, Panda S (2020) Blockchain: a survey on healthcare perspective and its challenges. In: International conference on information and communication technology for intelligent systems. Springer, pp 111–119
5. Kumari D, Rajita BSAS, Sekhar MR, Garg R, Panda S (2021) Predictive modeling of anthropomorphic gamifying blockchain-enabled transitional healthcare system. In: Machine learning approach for cloud data analytics in IoT. Wiley Online Library, pp 461–490
6. Otta SP, Panda S (2021) Cloud identity and access management solution with blockchain. In: Blockchain technology: applications and challenges. Springer, pp 243–270
7. Pajoo HH, Rashid MA, Alam F, Demidenko S (2022) Experimental performance analysis of a scalable distributed Hyperledger fabric for a large-scale IoT testbed. Sensors 22(13):4868
8. Pan H, Duan X, Wu Y, Tseng L, Aloqaily M, Boukerche A (2020) Bbb: a lightweight approach to evaluate private blockchains in clouds. In: GLOBECOM 2020—2020 IEEE global communications conference. IEEE, pp 1–6
9. Pankov KN (2020) Testing, verification and validation of distributed ledger systems. In: 2020 systems of signals generating and processing in the field of on board communications. IEEE, pp 1–9
10. Wang R, Ye K, Xu C-Z (2019) Performance benchmarking and optimization for blockchain systems: a survey. In: International conference on blockchain. Springer, pp 171–185
11. Xu X, Sun G, Luo L, Cao H, Yu H, Vasilakos AV (2021) Latency performance modeling and analysis for Hyperledger fabric blockchain network. Inf Process Manag 58(1):102436–102448
12. Zhu L, Chen C, Su Z, Chen W, Li T, Yu Z. Bbs: micro-architecture benchmarking blockchain systems through machine learning and fuzzy set. In: 2020 IEEE international symposium on high performance computer architecture (HPCA). IEEE, pp 411–423

# Digital Twins in Agriculture as an Internet of Things Paradigm: The Case of Azerbaijan



Fuad Ibrahimov , Ulviyya Rzayeva , and Rasul Balayev

**Abstract** The article discusses the areas of functioning and features of implementing digital twins in agriculture as the Internet of Things technologies. It is shown that the use of digital twins and a virtual representation of a real farm, can significantly increase management efficiency, especially for suburban agricultural systems. The article presents an optimization problem, which allows the modeling of suburban agriculture that serves to provide the city with fresh products. Management of a virtual copy of the temporary, material, and human resources of the farm occurs through the capabilities of the Internet of Things, allowing to control of several assets that work together or perform the same function. The study examines the features of digital twins in the formation of suburban crop and livestock production using system-dynamic modeling. In conclusion of the paper, optimal values are obtained in determining the priorities for the selection of relevant data, which should be constantly collected and updated. The necessary steps have been identified to implement the Internet of Things in agriculture in developing countries.

**Keywords** Internet of Things · System-dynamic modeling · Digital twin of a farm · Optimization problem · Management efficiency of an agricultural facility

---

F. Ibrahimov

Public Association “Center for Socio-Economic and Environmental Research”, AZ1000 Baku, Azerbaijan

e-mail: [fuad.ibrahimov@fmg.az](mailto:fuad.ibrahimov@fmg.az)

U. Rzayeva (✉) · R. Balayev

Azerbaijan State University of Economics, Baku AZ1001, Azerbaijan

e-mail: [ulviyya.rzayeva@unec.edu.az](mailto:ulviyya.rzayeva@unec.edu.az)

R. Balayev

e-mail: [rasul.balayev@unec.edu.az](mailto:rasul.balayev@unec.edu.az)

## 1 Introduction

The Internet of Things (IoT) is gradually becoming the norm not only in high-tech production but also in the agricultural industry [1]. Smart farming will allow farmers and agricultural producers to increase yields and reduce costs [2]. In the long term, smart farming will allow farmers and other stakeholders to better organize the processes that affect the bottom line [3].

Although IoT technologies are still rarely used in farming in developing countries, it is already clear that the “connected farm” will combine real-time management of farm processes and a wealth of historical data, including the means to dynamically update or adapt the model to data, as well as expanding the scope of the topics currently being discussed [4].

In Europe, the movement toward smart farming is encouraged by various projects and programs funded by the state budget and private investment [5]. However, in Azerbaijan, unfortunately, these technologies do not have a mass distribution due to the high cost, complexity, and other reasons.

This article examines data-driven system dynamics methods such as machine learning, demonstrating the unique ability to capture and model the complex characteristics of a digital farm management system. The presented optimization problem, which describes the operation of a suburban farm connected to the Internet of Things, is the task of determining the best, in a sense, structure or values of object parameters, allowing the collection of direct and indirect indicators necessary for farm modeling.

System-dynamic modeling applied in the article is one of the approaches that support the decision-making process in data-driven agricultural production by extending the logic of decision-making, and while providing reliable and predictable working behavior.

Thus, this work presents an approach to system-dynamic data modeling as an area for future research, in terms of new applications of the Internet of Things and the study of shortcomings and strategies to remove or suppress them.

## 2 Methodology

System-dynamic modeling solutions make a valuable contribution to the reproduction of entities and their interactions in agriculture [6]. At the same time, the infrastructure and platform for automated data collection are the main ones for ensuring and adequately reflecting the state of resources in real-time [7]. As research on digital twins and IoT opportunities in agriculture in Azerbaijan is still at an early stage, modeling such farming systems can be challenging.

The system-dynamic modeling methodology was applied in the work, which involves comparing data to provide an understanding of the underlying interactions and the behavior of the parameters under consideration.



This approach was developed and proposed by Jay Forrester in the late 1950s as a study of information feedback in industrial activity to show how organizational structure, gains (in policies), and delays (in decisions and actions) interact affecting the success of the enterprise. The area of application of system dynamics also includes social, urban, and ecological systems [8]. The processes taking place in the real world, within the framework of this approach, are presented in terms of storage units (stocks), streams between these storage units (flows), and information that determines the magnitude of these flows.

The case considered in the article is only a theoretical basis for designing a farming digital twin without trying to develop a scale reference model.

The evaluation of the applicability of the presented scheme in the context of smart agriculture is being tested.

The system-dynamic model is being piloted on the basis of the current statistical indicators of real farms operating in Absheron (the peninsula in Azerbaijan), with the consideration of two options for predicting digital twins within the proposed optimization problem. The model is implemented with the support of the MATHCAD package of mathematical applications.

### 3 Suburban Farm Optimization Model

A characteristic feature of suburban farms in Absheron is economic and social heterogeneity. Large commodity farms coexist here with small peasant farms. However, these small economies are far from homogeneous. Group membership is determined not only by assets, but also by gender, ethnicity, and social status since all of these involve an unequal ability to use the same reserves and resources in response to opportunities.

To modernize the industry, the creation of digital twins (a mathematical model that accurately describes an object) of fields or farmland allows planning the time and type of sowing crops and when, the amount of fertilizer applied, control of the acid composition of soils, and so on [9].

The suburban area of Baku, taken as the object of study, is limited by the Absheron Peninsula.

We have studied the changes as a result of the application of system-dynamic modeling in the delivery of fresh food products to the Baku agglomeration. For this, the indicator “weighted average transportation distance” is proposed and the corresponding calculations are carried out [10].

In the problem of optimizing the structure of suburban agricultural production presented in the article, the objective function assumes the maximization of a marketable product:

$$\sum_{j \in J_1} p_j x_j + \sum_{j \in J_2} p_j x_j \rightarrow \max \quad (1)$$

Let us consider a number of conditions that take into account the characteristics of the proximity of an agricultural object to the city, describing the limitations of this function

- regarding the involved production resources of the city (labor resources, technical means, etc.):

$$\sum_{j \in J_1} a_{ij} x_j + \sum_{j \in K_k} a_{ij} \cdot \sum_{k=1}^3 x_k - x_i + \sum_{j \in J_2} a_{ij} x_j \leq b_i \quad (2)$$

- regarding the use of urban food waste as animal feed:

$$c_j x_j - x'_j = 0, \quad j \in J_2 \quad (3)$$

$$\sum_{j \in J_2} v_j - x_i - x_{ij} \leq 0, \quad i \in I_2 \quad (4)$$

$$d_j x'_j - x_{ij} \geq 0, \quad i \in I_3, \quad j \in J_2 \quad (5)$$

$$e_i x'_j - \sum x_i \leq 0, \quad i \in I_2, \quad j \in J_2 \quad (6)$$

Here  $j, J$  are the index and the set of suburban agricultural production areas, respectively. This set consists of the following disjoint subsets:  $J_1$  is a subset of sowing fields;  $J_2$  is a subset of livestock plots.  $K$  is a set of agricultural products by purpose and consists of the following disjoint subsets:  $K_1$  is a subset of personal consumption products and marketable products;  $K_2$  is a subset of products used for feed;  $K_3$  is a subset of seeds and products of on-farm non-productive consumption.  $i, I$  are the index and the set of production resources, respectively. The set  $I$  consists of the following disjoint subsets:  $I_1$  is a subset of the involved production resources of the city (labor resources, technical means, etc.);  $I_2$  is a subset of feed types;  $I_3$  is a subset of city food waste types.

Variables:  $x_j$  is the development intensity of  $j$ th plot of animal husbandry;  $x_i$  is the number of involved production resources of the city;  $x'_j$  is the total need for food products of the  $j$ th plot of animal husbandry;  $x_{ij}$  is an amount of the optimal additive of the  $i$ th type of feed for the  $j$ th plot of animal husbandry.

Fixed values:  $a_{ij}$  are special norms of resource consumption in agricultural fields. This does not include costs associated with the collection and initial processing of agricultural products;  $a'_{ij}$  are special rates of resource consumption for harvesting and initial processing of agricultural products;  $c_j$  is a coefficient of demand in units of feed per unit of measurement of the  $j$ th plot of animal husbandry;  $v_j$  is a minimum specific weight of food waste per feed ration per unit of measurement of the  $j$ th plot of animal husbandry;  $d_j$  is a difference between the maximum and minimum specific weight of food waste in the feed ration per unit of measurement of the  $j$ th plot of animal husbandry;  $e_i$  is a difference between the total need for feed (taken as a unit)

and the sum of the minimum specific weight of certain types of feed in the diet per unit of the  $j$ th plot of animal husbandry;  $b_i$  is an amount of the  $i$ th type of resource;  $p_j$  are marketable products in the  $j$ th plot of animal husbandry.

The simulation system is real, and in some cases, due to the complexity of information support, it was verified in a generally accepted and normative database [11].

Implementing the digital twin concept requires confidence in the simulation model, data, and model update algorithm [12]. The results of our survey were used to determine priorities in the selection of data. Therefore, it is considered appropriate to include the influence of resource use characteristics among the factors to be modeled. In general, in order to move from modeling sustainable agricultural production in the suburbs to digital twins, it is necessary to ensure continuous collection and updating of data on the following group of indicators: the level of satisfaction of demand for fresh food by the urban population; the state of the environment and the natural resource base and the dynamics of changes; resource efficiency; competitiveness of agricultural products; relevant indicators of the quality of life of producers and consumers [13].

## 4 Results Obtained

The data on which digital twins are based is divided into pre-collected data and data collected specifically for the creation of digital twins. Although these divisions are arbitrary, in practice it will be possible to gain new reasoning and knowledge thanks to the connections between them [14].

Taking into account the proximity to the city, according to the model for optimizing the structure of suburban production, options were considered to increase the production of vegetables in greenhouses by 15% and 30%, milk by 30% and 40%, eggs by 0% and 25%, respectively (Table 1).

In the suburban Absheron-Khizi economic region, calculations were made to increase the production of vegetables on the open ground by 20 and 35%, milk by 30 and 40%, egg production by 0 and 15% (Table 2).

It is expected that the volume of marketable output of farms in Baku will increase by 1.5% under the first option of calculations and by 6.5% under the second option. The volume of marketable output for farms in the suburban Absheron-Khizi economic region will increase by 1.5% according to the first calculation option and by 15.7% according to the second option. It is expected that the cost of fresh vegetables, fresh milk, and eggs in the farms of both districts will not change under the first option, and under the second option, this figure will decrease by 7–10%.

Digital twins will improve simulation results by supporting suburban farms at all stages of operation through continuous real-time monitoring.

By influencing the relationship between reproduction and resource consumption, digital twins can optimize performance. For this purpose, the proposed digital twin architecture is suitable for suburban agriculture.

**Table 1** Forecast for 2025 on commodity production in Baku farms

	Commodity product value, manat		Commodity product structure, %	
	I variant	II variant	I variant	II variant
Potato	14,476,2	14,476	—	—
Vegetables (open ground)	17,134,654	17,134,654	26.0	24.4
Vegetables (closed ground)***	976,109	1,060,988	1.4	1.5
Cereals and legumes	—	—	—	—
Melon plants	7326	7326	0.1	0.1
Fruits and berries	5,097,842.7	5,097,843	7.7	7.3
Grapes	7,005,828	7,005,828	10.6	10.0
<i>Crop production, total</i>	<i>30,236,236</i>	<i>30,321,115</i>	<i>45.2</i>	<i>43.3</i>
Meat (slaughter weight)	27,611,500	27,611,500	41.3	39.4
Milk	3 748 176	5 651 096	5.6	8.0
Eggs (thousands)	5,211,860	6,514,825	7.8	9.2
Wool (physical weight)	59,872.3	59,872	0.1	0.1
<i>Livestock, total</i>	<i>36,631,408.3</i>	<i>39,837,293</i>	<i>54.8</i>	<i>56.7</i>
Total	66,867,644	70,158,408	100	100

**Table 2** Forecast for 2025 on commodity production in the farms of suburban Absheron-Khizi economic district

	Commodity product value, manat		Commodity product value, manat	
	I variant	II variant	I variant	II variant
Potato	99,445.2	99,445	—	—
Vegetables (open ground)	27,748,580	31,217,153	2.9	2.9
Vegetables (closed ground)***	1,173,820	1,173,820	—	—
Cereals and legumes	3,858,030.4	3,858,030	0.4	0.4
Melon plants	244,588.5	244,589	—	—
• Fruits and berries	2,445,909	2,445,909	0.3	0.3
• Grapes	445,909	445,909	0.1	0.04
<i>Crop production, total</i>	<i>35,746,282</i>	<i>39,214,854</i>	<i>3.7</i>	<i>3.6</i>
• Meat (slaughter weight)	35,951,300	35,951,300	3.8	3.3
• Milk	42,826,403.75	46,120,743	4.5	4.2
• Eggs (thousands)	838,523,892.2	964,302,476	87.9	88.8
• Wool (physical weight)	540,311	540,311	0.1	0.1
<i>Livestock, total</i>	<i>917,841,907</i>	<i>1,046,914,830</i>	<i>96.3</i>	<i>96.4</i>
Total	953,588,189	1,086,129,683	100	100

## 5 Conclusion

One of the foundations of any “smart” production is the platform solutions of the Internet of Things, which allow the combination of all production units into a single system, as well as programs for collecting and exchanging information. At the same time, the implementation of software and hardware alone is not enough to talk about the implementation of a digital twin. The goal can be considered achieved when a qualitatively new asset management model is created based on analytical calculations performed on a digital twin.

Also, now we can quickly find losses in production and their causes, and monitor the implementation of tasks. But, perhaps, much more important is the fact that on the basis of digital twins it is possible to carry out system-dynamic modeling, that is, to make forecasts.

We have not only a beautiful picture but also the answer to the question: “What if?” With the help of simulation modeling, we will be able to introduce into our production already correct, well-established concepts that are based not only on intuition or accumulated experience but have a mathematical justification for certain decisions.

The main problem in creating a digital twin in Azerbaijan, an agricultural enterprise needs to solve a number of additional tasks. One of the key issues is related to data storage, especially since the digital twin costs quite a lot. However, this problem has its own solutions: from expanding the server to cloud storage or storing data on the provider’s servers. It is also necessary that communication channels transmit information without delays and distortions, comply with the criteria for compatibility, scalability, and cyber security with the existing IT system.

The article proposes a conceptual model for the implementation of the digital twins technology for suburban agriculture enterprises. Taking into account the specifics of agricultural facilities, the technology of digital twins, based on the implementation of a single integrated, constantly updated system-dynamic model, is considered as the basic one. The structure includes a mathematical control model based on the presented optimization problem, describing the object exactly, and a digital twin implementation model using statistical data. The operating model defines the functions and existing information flows to implement the concept of digital twins. The implementation model performs technical functions at the application level, a predictive model is proposed. The peculiarity of the proposed method is that it is an intermediary in the transition from system-dynamic modeling to digital twins. The proposed model can be considered as a foundation for the development of concepts, the formation of requirements and the development of solutions for the creation and development of information management systems for enterprises in the agricultural sector of the economy.

This framework is applied and tested in two predictive smart suburban farming scenarios.

## References

1. Verdouw C, Tekinerdogan B, Beulens A, Wolfert S (2021) Digital twins in smart farming. *Agric Syst* 189:103046. ISSN 0308-521X. <https://doi.org/10.1016/j.agry.2020.103046>
2. Pylaniadis C, Osinga SA, Athanasiadis IN (2021) Introducing digital twins to agriculture. *Comput Electron Agric* 184(4):105942. <https://doi.org/10.1016/j.compag.2020.105942>
3. Slob N, Hurst W (2022) Digital twins and industry 4.0 technologies for agricultural greenhouses. *Smart Cities* 5:1179–1192. <https://doi.org/10.3390/smartcities5030059>
4. Foldager FF, Thule C, Balling O, Larsen P (2021) Towards a digital twin—modelling an agricultural vehicle. In: *Leveraging applications of formal methods, verification and validation: tools and trends*. Springer International Publishing, Cham, Switzerland, pp 109–123
5. Purcell W, Neubauer T (2022) Digital twins in agriculture: a state-of-the-art review. *Smart Agric Technol* 3:100094. ISSN 2772-3755. <https://doi.org/10.1016/j.atech.2022.100094>
6. Turner BL, Menendez HM, Gates R, Tedeschi LO, Atzori AS (2016) System dynamics modeling for agricultural and natural resource management issues: review of some past cases and forecasting future roles. *Resources* 5:40. <https://doi.org/10.3390/resources5040040>
7. Li FJ, Dong SC, Li F (2012) A system dynamics model for analyzing the eco-agriculture system with policy recommendations. *Ecol Model* 227:34–45. ISSN 0304-3800. <https://doi.org/10.1016/j.ecolmodel.2011.12.005>
8. Forrester JW (1987) Lessons from system dynamics modeling. *Syst Dyn Rev* 3(2):136–149. <https://doi.org/10.1002/sdr.4260030205>
9. Bastan M, Ramazani Khorshid-Doust R, Delshad Sisi S, Ahmadvand A (2018) Sustainable development of agriculture: a system dynamics model. *Kybernetes* 47(1):142–162. <https://doi.org/10.1108/K-01-2017-0003>
10. Balayev R, Mirzayev N, Bayramov H (2021) Sustainability of urbanization processes in the digital environment: food security factors. *Acta Sci Pol Administratio Locorum* 20(4):283–294
11. Balayev RA (2007) *Urbanization: the urban economy and the food problem*. Baku, “Elm” Publishing House, pp 223–234; 241–266
12. Wright L, Davidson S (2020) How to tell the difference between a model and a digital twin. *Adv Model Simul Eng Sci* 7(13) <https://doi.org/10.1186/s40323-020-00147-4>
13. Walters JP, Archer DW, Sassenrath GF, Hendrickson JR, Hanson JD, Halloran JM, Vadas P, Alarcon VJ (2016) Exploring agricultural production systems and their fundamental components with system dynamics modelling. *Ecol Model* 333:51–65. <https://doi.org/10.1016/j.ecolmodel.2016.04.015>
14. Burg van der S, Kloppenburg S, Kok EJ, Voort van der M (2021) Digital twins in agri-food: societal and ethical themes and questions for further research. *NJAS: Impact Agri Life Sci* 93(1):98–125. <https://doi.org/10.1080/27685241.2021.1989269>

# Theoretical Fundamentals of Criteria for Evaluation of Efficiency, Quality and Optimization of Complex Informatiology Systems



Volodymyr Kulivnuk , Ivan Kuzmin, Oleksandr Hladkyi ,  
Alexander Gertsy , Tetiana Tkachenko , and Tetiana Shparaga

**Abstract** The new criterion that would allow synthesizing an optimal process and a complex system regarding the most important indicators of quality efficiency is explored. General requirements to criteria for evaluation of efficiency, quality and optimization of complex informatiology systems are investigated. The specific algorithm for criteria to evaluate efficiency, quality and optimization is defined. This algorithm is based on corresponding of algebraic logic theory, correlation theory, general theory of random functions, theory of statistical solutions, probability theory, game theory, information theory as well as on general theory of efficiency. Using this algorithm, we can solve different problems. There are: building mathematical model of complex system, building mathematical model of complex system functioning, development and optimization of the complex system functioning algorithm, complex system synthesis and analysis, selection of complex system implementation elements, alignment of complex system elements between each other as well as evaluation of complex system efficiency, quality and optimization. The generalized functional-statistical criterion for evaluation of efficiency, quality and optimization is derived. This derivation is based on command and control process as the information source as well as on automated command and control system operational algorithm. We suggest to use 3 devises in complex automated command and control system for

---

V. Kulivnuk (✉) · I. Kuzmin  
Vinnytsia National Pirogov Memorial Medical University, Vinnytsia, Ukraine  
e-mail: [vs.kulivnuk@ukr.net](mailto:vs.kulivnuk@ukr.net)

O. Hladkyi · T. Tkachenko  
Kyiv National University of Trade and Economics, Kyiv, Ukraine  
e-mail: [o.gladkey@knute.edu.ua](mailto:o.gladkey@knute.edu.ua)

T. Tkachenko  
e-mail: [t.tkachenko@knute.edu.ua](mailto:t.tkachenko@knute.edu.ua)

A. Gertsy  
State University of Infrastructure and Technologies, Kyiv, Ukraine

T. Shparaga  
Taras Shevchenko Kyiv National University, Kyiv, Ukraine

evaluation of efficiency, quality and optimization. Each of them operates according to its own algorithms, whose synthesis should be carried out regarding the state of the object described by the mathematical model, goals of the individual command and control stages.

**Keywords** Informatiology · Complex systems · Quality · Efficiency · Optimization · Evaluation algorithm

## 1 Introduction

Synthesis of any system, in particular a complex informatiology system, should begin with choosing and justifying the criteria for evaluation of efficiency, quality and optimization. At the same time, it is necessary to choose a criterion that would allow synthesizing an optimal process and a complex system regarding the most important indicators of quality efficiency. These indicators should in the first place include the following.

- Problem solving probability or system reliability;
- Information capacity;
- Speed;
- Volume and weight, complexity and cost;
- Accuracy of work and controllability of the system itself;
- Noise immunity.

In addition to these requirements, the criterion should possess certain constructability in order to easily evaluate its numerical value, which would allow calculating the efficiency of not only the process, device and the system itself in its proximity to potential perfection, but also comparatively similar devices, processes and systems in cumulation.

Propositions and consequences formulated in this research allows deriving a generalized functional-statistical efficiency evaluation criterion that satisfies all above-mentioned requirements. Main properties of the generalized criterion and particular criteria derived from the general, based on the maximum total utility at minimum cost are also described.

Presentation employs the elements of problem-based learning whose algorithm is reduced to the following main propositions [1].

- Formulation of the main problem, which is then subdivided into a number of smaller problems (sub-problems);
- Realization of the problem and sub-problems;
- Collecting alternatives and making hypotheses;
- Collecting criteria for evaluation of the optimal (the best) alternative and hypothesis;
- Proof of the optimality of the alternative and hypothesis;



- Solution verification;
- Repetition and analysis of the solution procedure;
- Summing up and conclusion.

## 2 General Requirements to Criteria

Criterion takes to mean a measure allowing quantitative and qualitative estimates for classification of a system or selection of the preferred option among the specified set of options.

Criteria can be expressed at various levels of abstraction.

- Linguistic and geographical;
- Set-theoretic and algebraic;
- Probabilistic and dynamic;
- Heuristic.

The main requirements for the criteria include [2–4] the following.

1. The criterion should be objective, i.e., reflect objective reality (ampere is the criterion for measuring current strength, volt for measuring voltage, Hurwitz-Rauss, Nyquist or Mikhailov equations for determining stability, etc.).
2. The criterion should characterize the efficiency, quality or optimality of a complex informatology system based on the functional purpose, for example, the efficiency criterion evaluates the degree of system's approximating to the global goal.
3. The criterion should be easy to interpret in physical terms and easy to calculate mathematically, best of all in numbers, using computer technology.
4. The criterion should be normalized, which allows obtaining zero dimension and evaluating the system's degree of approximation to the ideal (for example, the efficiency  $\eta$  varies from 0 to 1, while  $\eta = 1$  corresponds to an ideally useful system, and  $\eta = 0$  – to a useless one).
5. The extreme criterion values should characterize the extreme states or the efficiency of a potential and real complex informatology system.
6. The criterion should have a certain generality. It should be suitable for evaluating individual subsystems and the system as a whole at various periods of the system's life (development, operation, restoration, etc.).
7. The criterion should have an optimum, better analytical one, inside a certain area or on its border.
8. The criterion should be theoretical to be able to create a theory on its basis.
9. All criteria can be divided into deterministic and statistical particular (local) and generalized (global), additive and multiplicative.
10. As a rule, the criteria should satisfy all requirements. If the criterion fails to meet some requirements, then its quality will be lower, and sometimes this is confusing.

### 3 Selection Algorithm for Criteria to Evaluate Efficiency, Quality and Optimization

Algorithm is a description of the strict sequence of the operation in time and space.

Selection algorithm is summarized in Table 1, where columns 2–9 indicate criteria based on different theories. Logistic operator  $F(z)$  assuming value 1 or 0 can be used at the level of the algebraic logic. Unity in this case corresponds to the completion of the task by the system to obtain the specified effect, 0 – to the failure of the system to complete the task to obtain the specified effect.

The correlation theory may use mathematical expectation  $m_x$ , the standard deviation  $\sigma_x$ , the correlation coefficient  $\tau_x$  and the normal probability distribution law  $F_H(t, \tau)$ .

The general theory of random functions in addition to the previously indicated characteristics may use the kurtosis coefficient  $\gamma_x$ , the skewness coefficient  $\nu_x$  and the probability distribution law  $F(t, \tau)$ .

The theory of statistical solutions may use the likelihood ratio  $F(t, \tau)$ .

The probability theory may use the problem solution probability  $P(t, \tau)$ .

The game theory may use the risk  $L(P)$ .

The information theory may use the entropy  $H$ , information amount/bandwidth  $I_{\max}/T$ .

The general theory of efficiency may use a generalized functional-statistical criterion for evaluating the efficiency, quality and optimization  $E_{IC}(t, \tau)$ .

Column 1 (see Table 1) lists the main enlarged tasks of analysis and synthesis of complex informatology systems.

The accepted designations (see Table 1) are as follows: 1—the problem is solved completely, 1—the problem is not solved completely, 0—the problem is not solved at all.

Table 1 shows that a generalized functional-statistical criterion is the most convenient for the analysis and synthesis of the complex informatology systems.

### 4 Deriving the Generalized Functional-Statistical Criterion for Evaluation of Efficiency, Quality and Optimization

The criterion is derived by the induction method, i.e., from the particular to the general. At the same time, particular criteria can be considered as criteria for evaluating the quality of the complex informatology systems work. The complex informatology system is presented by the automated command and control system (ACCS) for the complex informatology object.

The criteria of the ACCS work quality comprise accuracy, the probability of completing the task, speed, cost, weight and volume, information capacity and total cost of the system production and operation.

**Table 1** Selection algorithm for criteria to evaluate efficiency, quality and optimization

Problems to solve	Algebraic logic, $F(z)$	Correlation theory, $m_x, \sigma_x, \tau_x$	General theory of random functions, $F(t, \tau)$ $m_x, \sigma_x, \tau_x, \gamma_x, \nu_x$	Theory of statistical solutions, $F(t, \tau)$	Probability theory, $P(t, \tau)$	Game theory, $L(P)$	Information theory, $H, I, I_{\max}/T$	General theory of efficiency, $E_{IC}(t, \tau)$
1. Building mathematical model of complex system (CS)	0	1 –	1 –	1	1	1	1	1
2. Building mathematical model of CS functioning	0	0	0	0	1 –	1 –	1	1
3. Development and optimization of the CS functioning algorithm	0	0	0	0	1 –	1 –	1	1
4. CS synthesis and analysis	0	0	0	0	1 –	1 –	1	1
5. Selection of CS implementation elements	0	0	0	0	0	0	1	1
6. Alignment of CS elements between each other	0	0	0	0	0	0	1	1
7. Evaluation of CS efficiency, quality and optimization	0	0	0	0	1 –	1 –	1	1
8. The number of estimates	1	3	5	1	4	3	1	

## 5 Command and Control Process as the Information Source

Earlier we have shown that entropy is an integral evaluation of the object and ACCS state. Entropy is also the main statistical characteristic of the command and control process.

It is known from the information theory that any phenomenon with uncertainty whose numerical measure is presented by entropy can be considered as a source of information. Therefore, the command and control process can also be considered as a source of information.

Potentially the command and control process has an infinite amount of information, since the systems of the object and the ACCS contain millions of particles with entropy. However, the numerical value of this entropy cannot be determined.

In practice we are usually interested not in the state of all particles, but in the state of enlarged complexes, whose statistical characteristics are necessary to evaluate the possibility of achieving certain goals and which can be determined with any accuracy. All this adds some subjective coloring to the characteristics of the command and control process without losing reality.

The real amount of information in the process is equal to the entropy of the object in ACCS.

$$V_n(t, \tau) = H_0(t, \tau) \quad (1)$$

The command and control process may produce the maximum information volume  $V_n(t, \tau) = V_{a \max}(t, \tau)$  with the greatest uncertainty of the object state. In other words, if to consider the state of the object while controlling  $i$ -system of the object as one event, then the greatest uncertainty is at  $P_{0i}(t, \tau) = 1/2$ .

The entropy of the state of  $i$ -system of the object under command and control may be determined by the formula.

$$H_0(t, \tau) = -\{P_{0i}(t, \tau) \log_2 P_{0i}(t, \tau) + [1 - P_{0i}(t, \tau)] \log_2 [1 - P_{0i}(t, \tau)]\}, \quad (2)$$

where  $P_{0i}(t, \tau)$ —probability of the task completion by the  $i$ -system of the object.

From this formula it follows that

$$V_{0i}(t, \tau) = H_{0i}(t, \tau) = \max \text{ at } P_{0i}(t, \tau) = 1/2$$

By substituting  $P_{0i}(t, \tau) = 1/2$  in Formula 2 we find

$$H_{0i}(t, \tau) = 1 \text{ bits}$$

if the object contains  $m$  systems:

$$H_0(t, \tau)_{\max} = \sum_{i=1}^m H_{oi}(t, \tau) = m \quad (3)$$

Naturally, the ACCS is the better, the more information about the state of the object it can accept and transmit.

The ideal ACCS is the system operating without information loss accepting and transmitting  $V_{\text{ACCS}}(t, \tau) = m$ .

In this case

$$V_0(t, \tau) - V_{\text{ACCS}}(t, \tau) = \Delta V(t, \tau) = 0 \quad (4)$$

In practice, an ideal system cannot be built, since the ACCS implements real conditions (usually  $P_{oi}(t, \tau) > 1/2$ ) with real equipment that has a finite operating accuracy.

In its turn the real equipment implements real command and control algorithms and real operating accuracy.

## 6 ACCS Operational Algorithm

The ACCS operational algorithm is a set of rules and instructions that determine the ACCS behavior in the process of the object command and control.

Usually, the rules and instructions are strictly specified in the process of the object command and control, i.e., the process is carried out according to a deterministic algorithm.

When optimizing the ACCS by the Monte Carlo method or another probabilistic method, the rules and instructions are given statistically, i.e., command and control are carried out according to a non-deterministic algorithm.

Complex ACCS as a rule, includes the following main functionally related devices [5].

- Devices for obtaining information directly from the object and converting it into a form convenient for further use, they are called primary information processing devices (PIPD);
- Devices regulating the entire process of preparation, which are secondary information processing devices (SIPD);
- Devices using information to change the state of the object and ACCS in the command and control process as well as devices for indicating and recording information, called final information processing devices (FIPD).

Each group of these devices operates according to its own algorithms, whose synthesis should be carried out regarding the state of the object described by the mathematical model, goals of the individual command and control stages whose

achievement is assessed by the relevant criteria for the quantitative characteristics of externalities and internalities effects, as well as the technical feasibility of algorithms.

Therefore, the ACCS operation algorithm consists of the systems of algorithms, whose timely and reliable implementation should be carried out by the ACCS.

Let's call a system of algorithms ideal if it is able to transmit the maximum amount of information from the command and control process into the amount of information received by the ACCS.

Ideality of the algorithm is only a necessary condition for transmitting the maximum amount of the process information into the amount of information received by the ACCS. A sufficient condition is the ideality of the command and control devices.

Achievement of the necessary and sufficient conditions produces the ideal system realizing the equality (4).

An ideal system of command and control algorithms gives a maximum amount of information equal to  $m$  bits.

Amount of information provided by the ACCS:

$$I_{\max}(t, \tau) = H_0(t, \tau)_{\max} - \Delta H_{\text{alg}}(t, \tau), \quad (5)$$

where  $\Delta H_{\text{alg}}(t, \tau)$ —entropy due to the algorithm imperfection.

It follows from equality (5) that the amount of the received information is maximum if  $\Delta H_{\text{alg}}(t, \tau) = 0$ , which is true for an ideal algorithm. Therefore, taking into account the maximum uncertainty of the object state evaluated by equality (3), we can write

$$I_{\max \max}(t, \tau) = H_0(t, \tau) = m \quad (6)$$

In practice approximation real algorithms are used as it is difficult and sometimes impossible to develop and implement ideal algorithms.

We call a system of algorithms real if it is chosen regarding the real possibilities of creating the algorithms structure and the real possibilities of these algorithms implementation.

The real system of algorithms allows obtaining the average amount of information

$$I_p(t, \tau) = I_{\max}(t, \tau) < I_{\max \max}(t, \tau) \quad (7)$$

Since in practice the inequality  $P(t, \tau) > 1/2$  is always true, then  $H_{0p}(t, \tau) < H_0(t, \tau)_{\max}$ , therefore the average amount of information can reach the level  $I_{\max}(t, \tau)$ , i.e.,

$$I_p(t, \tau) = I_{\max}(t, \tau) \quad (8)$$

The choice of a real system of algorithms results in some loss of information, which for ideal command and control devices is equal to.

$$\Delta V_{\text{alg}}(t, \tau) = V_0(t, \tau) - V_{\text{ACCS ap}}(t, \tau), \quad (9)$$

where  $V_{\text{ACCS ap}}(t, \tau)$ —amount of information received by the ACCS with real algorithms and ideal devices.

Naturally the smaller is the value  $\Delta V_{\text{alg}}(t, \tau)$ , the more perfect is the system of algorithms.

## 7 Conclusions

Algorithmization process is very complex and requires highly qualified specialists with extensive practical experience, since when developing a real system of algorithms, a specialist guided by knowledge is expected to neglect some of the ideal algorithms, excluding them from the command and control process, to simplify some of them some and to exclude some from the process after preliminary analysis.

Based on the above it is quite reasonable to introduce the concept of the algorithm accuracy. Under the algorithm accuracy we mean the accuracy with which it is possible to develop a real algorithm relative to the ideal one.

The finite accuracy of the real algorithm determines the loss of information in the command and control process, which is calculated by Formula (5).

The theory of real algorithms accuracy including the theory of criteria for evaluating the efficiency, quality and optimization of complex informatory systems are not yet developed and more to it—the clear formulation of the problem in question is still unavailable.

## References

1. Kuzmin IV, Trotsychyn IV, Kuzmin AI, Kedrus VO, Lubchik VR (2009) Fundamentals of Information Theory and encrypting. In: Kuzmin IV (eds) Teaching book. Khmelnytskyi National University, Khmelnytskyi
2. Kuzmin IV (1971) Estimation of effectiveness and optimization of ACCS. Soviet Radio, Moscow
3. Kuzmin IV (1981) Fundamentals of complex system modeling. Vyscha shkola, Kyiv
4. Kuzmin IV (1981) Synthesis of computational algorithms of management and control. Technika, Kyiv
5. Kuzmin IV.: Methodical fundamentals for problematic studies of basic technical tools and systems. vol. 1, 210 p. Vinnytsia Polytechnic Institute, Vinnytsia (1983).

# Detection of Structure Changes in Lightweight Concrete with Nanoparticles Using Computer Vision Methods in the Construction Industry



Roman Mysiuk , Volodymyr Yuzevych , Bohdan Koman ,  
Yuriy Tyrkalo , Oleksandra Farat , Iryna Mysiuk ,  
and Lyudmyla Harasym

**Abstract** The research object is the structure of concrete, which contains pores, hydrated calcium silicate gel, and other trace elements. A microscope is used for the visual analysis of microparticles. Analysis of images from a microscope can solve the problem of estimating the number of microstructural elements in a certain area. This affects the faster search for the optimal concrete structure to improve compressive strength. Automated recognition of the main elements of the concrete structure based on microscopic images using the basic methods of computer graphics and vision is proposed. Nanoparticles of CuO or SiO<sub>2</sub> are chosen for investigating the structure of the concrete due to their positive effect on concrete strengthening. The paper considers the influence of 1 and 2  $\mu\text{m}$  resolution of the microscope for comparing the microstructure of concrete with nano-SiO<sub>2</sub> particles. An analysis of images with CuO nanoparticles in concrete during curing is performed, and the percentage content of the main structural elements is determined. The described investigation can be used to improve methods of quality assessment of nanoconcrete, determination of resource of material, and calculation of parameter of strengthening on the basis of the investigated samples.

**Keywords** Computer vision · Image mining · Pixel · Lightweight concrete · Structure of concrete · Pore · CuO nanoparticle · SiO<sub>2</sub> nanoparticle

---

R. Mysiuk (✉) · B. Koman · I. Mysiuk  
Ivan Franko National University of Lviv, 1 Universytetska Str, Lviv 79000, Ukraine  
e-mail: [mysyukr@ukr.net](mailto:mysyukr@ukr.net)

V. Yuzevych  
Karpenko Physico-Mechanical Institute of the NAS of Ukraine, 5 Naukova Str, Lviv 79060, Ukraine

Y. Tyrkalo · O. Farat  
Lviv Polytechnic National University, 12 Stepana Bandera Str, Lviv 79013, Ukraine

L. Harasym  
Ukrainian National Forestry University, 103 Gen. Chuprynky Str, Lviv 79057, Ukraine



## 1 Introduction

Improving the strength, durability, and resource of elements of building structures made of nanoconcrete should be performed using modern methods of information technologies. This type of research contributes to increase the levels of reliability and quality of the relevant materials, in particular the structure, physical, and chemical properties. Important studies in this context are related to the ensemble of pores and microcracks of structural material and interfacial layers.

The correct combination of selected mixture components has an impact and important role in the strength of concrete. Usually, the durability of the architectural form or infrastructure elements depends on it. Therefore, such studies should be conducted, as there is a problem with the selection of optimal components and the formation of reliable structures from concrete, which require a lot of time and experiments.

Usually, nanoparticles of certain elements improve the properties of the concrete mixture [1, 2]. The optimal combination of concrete mixture components affects the curing period and density [3]. This makes it possible to establish changes in strength in the microstructure of concrete during this time.

Considering the size of the elements, scanning electron microscopes are used for research, which can bring the image closer to micrometers. The microstructure of a concrete mix often consists of pores, crystals, nanoparticles, and hydrated calcium silicate gel (C–H–S gel). Each of these elements has its own external differences, by which the object can be visually recognized.

The number and diameter of pores and C–H–S gel affect the strength of concrete. A positive effect on these structural trace elements of the mixture is observed when CuO or SiO<sub>2</sub> nanoparticles are added.

Concrete strength can be determined based on compressive, flexural, and split tensile strength tests. Otherwise, long-term use or selected components of cement paste may cause defects that can lead to the destruction of concrete structures.

Computer vision is designed to detect both moving and stationary objects. It is possible to determine the color and transparency of the image pixels using computer graphics methods. Since the microscopic images are in gray tones, certain objects can be identified by the shade of color. The results of such studies are relevant because, based on the realities of today's practice, they make it possible to dynamically detect and calculate the number of pores in the studied sections of concrete.

## 2 Literature Review and Problem Statement

The basic characteristics of concrete strength are mechanical tests of compressive, flexural, and split tensile strength. The work [1] is devoted to influence nanoparticles on the optimization of the structure of concrete with nanoparticles. This study contains the results of mechanical strength tests with the addition of SiO<sub>2</sub> nanoparticles to the concrete mixture. As a result of this work, a stronger structure of concrete

is obtained. Since the mixture of concrete with  $\text{SiO}_2$  nanoparticles is considered in one of our experiments, this information is useful for explaining the influence of nanoparticles on the structure of concrete. The disadvantage of this work can be considered insufficient analysis of microscopic images. This can be related to specific research, namely the strength of concrete. The work [2] is devoted to the combined discrete element-different approach for modeling mechanical reactions saturated with liquid porous materials. Pores are one of the microelements in the structure of concrete. The approach described in the work [2] allows modeling the dependence of concrete strength on pressure and load speed, mechanical properties of liquid, and sizes of sample concrete. This paper does not take into account the peculiarities of nanoparticles that can be added to the mixture. In the work [3], the properties of concrete with the addition of  $\text{CuO}$  nanoparticles, on the basis of which one of our experiments, are conducted. However, the disadvantage of the work is that the analysis of the structure of concrete is made not on the basis of microscopic images, but with the help of physical and mechanical experiments. Some data from these works are taken into account for the description of the relationship between steel and concrete in the theoretical part of the work.

In the work [4], research on the mechanical properties of concrete with the addition of  $\text{SiO}_2$  has been performed. As a result of this experiment, microscopic images are obtained, which are useful for the analysis of concrete structures. In the paper [5], the analysis of the level structure of the nanocomposite cement solution with titanium dioxide is considered. The analysis of the sample in this source is conducted without the use of computer vision, but it allows us to understand the way of separation of microelements in concrete.

Methods and means of application of computer vision are described in the work [6]. This information allows the application of the algorithm of pixel analysis and allocates the basic spheres of application of this information technology for data analysis.

In works [7, 8], the general overview of properties of concrete mixture with nanoparticles  $\text{SiO}_2$  is described. The works contain microscopic images with different details, but the impact of detail on the number of pores and C–H–S gel in concrete is not considered. These works are useful for the analysis of the dependence of the increase in microscope detail on the number of microelements of concrete in our research.

In the work [9], the study of the influence of  $\text{CuO}$  nanoparticles and boron wastes on the properties of cement solution is performed. The study contains a detailed description of the curing process with mechanical tests. The results of the research [9] help to understand the ratio of the number of basic microelements of concrete, and their influence on curing. Work [10] describes the process of hardening of concrete mixture with  $\text{CuO}$  nanoparticles with added slag. The microscopic images used in the work [10] are suitable for visual analysis of the shape and texture changes of the concrete mixture with  $\text{CuO}$  nanoparticles.

In most of the works, the process of data processing is not automated and is more connected with the investigation of the physical and chemical properties of the

mixture. The above works mostly did not use information technologies for detailed analysis of microscopic images.

From the above sources, it is possible to conclude that the important parameters in the structure of concrete are the number of pores and C–H–S gel.

The development of modern information technologies and the relevance of research in the field of urban development stimulate the development of the software market. Concrete is the main component for building bridges, buildings, and other structures, and durability depends on the strength of its components. Practical application in the future of such an approach will allow attracting more investment projects, using approaches described in works [11, 12]. However, for the implementation of the complex product with the described approaches, it is necessary to perform more detailed diagnostics of the investigated samples.

The topics of microstructure detection of concrete based on microscopic images are not given enough attention. The reason for this may be objective difficulties, due to the insufficient number of microscopic images of concrete in free access. It is the approach with the use of computer vision realized in work [13]. However, this analysis approach is used to identify defects, not to find structural microelements.

All this makes it possible to assert that it is expedient to perform research, devoted to analysis of the structure of concrete with the use of methods of computer vision.

### 3 The Purpose and Objectives of the Research

The purpose of the research is to reveal the regularities of changes in the structure of light concrete with nanoparticles taking into account the two-phase transfer of the substance and contact with the armature.

This allows dynamic detection of microstructural elements in concrete with nanoparticles and calculation of their quantity on the investigated area.

Achievement of the formulated goal includes the following tasks:

- determine the effect of microscope resolution on the number of pores in the concrete structure;
- perform microstructure diagnostics of concrete with added nanoparticles in the process of curing.

### 4 Materials and Methods of Research

The main object of the research is the microstructure of concrete with CuO and SiO<sub>2</sub> admixture. These nanoparticles contribute to rapid curing and increase in the strength of concrete. Thus, this will cause a reduction in the number of defects during long-term operation.

As is known, the addition of CuO or SiO<sub>2</sub> particles to the concrete mass has a positive effect on the concrete strength, because the pore structure shifts to harmless

or several harmless pores, and harmful pores are removed from the concrete structure [3, 4].

Images from microscopes are usually presented in shades of gray, and outlines of elements are displayed in contrast. The lightest elements are those located in the foreground, and the darkest are those with bends or voids.

As inclusion criteria for the research method, it is possible to consider the implementation of a program for image segmentation with concrete microstructure based on basic pixel data. The program for calculating the percentage of trace elements is written in the Java programming language using standard libraries. To work with graphic images, the Color class from the java.awt.\* library is used. The color of each pixel is divided into three elements red, green, and blue (RGB). The maximum value of each parameter is 255, and the minimum is 0. Accordingly, the combination of red = 0, green = 0, and blue = 0 is black, and red = 255, green = 255, and blue = 255 is white [6].

As for the algorithm of the program, firstly, each pixel is read, and the color of the pixel is checked. The next step is to change the color and count their number for the illuminated area. At the same time, the condition is checked that the color is less or more than a fixed value in the RGB scale. The percentage of certain elements can be calculated by taking the number of found pixels of the image to the size of the image.

## **5 The Results of Detection of Basic Microstructural Elements Based on the Image from the Microscope**

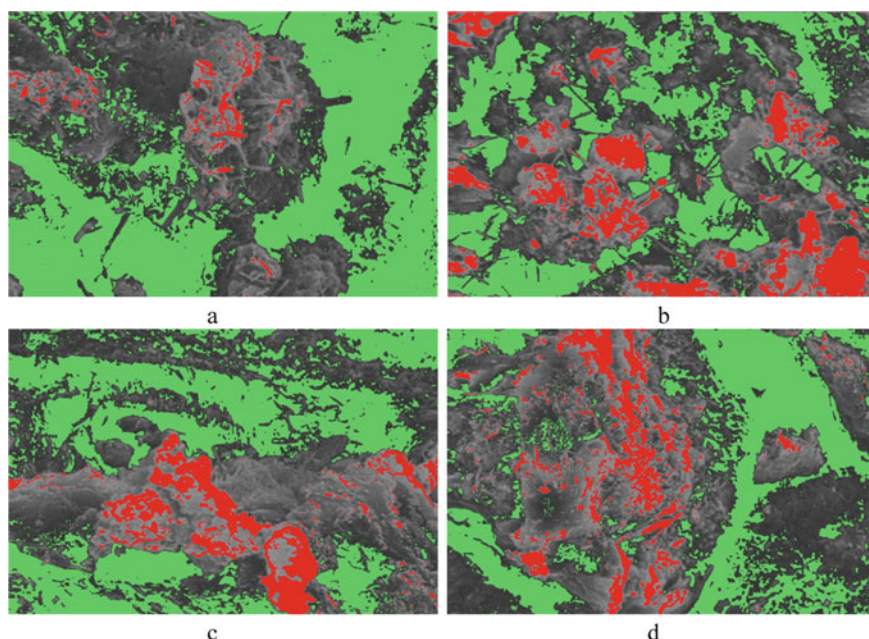
### ***5.1 Determining the Effect of Microscope Resolution on the Number of Pores in the Concrete Structure***

Images from microscopes are usually in shades of gray because the distances between the elements are small, and the illumination of the studied area is insufficient. However, it is known that the C–H–S gel is mostly formed near the pores and has a specific spherical structure. It is often the lightest element in an image because it is extreme. Instead, pores are unfilled dark and deep areas in the image.

As a result of superimposing the colors (see Fig. 1) green on the area of pores and red on the areas of C–H–S gel on the image made with a microscope, these microstructural elements can be visually distinguished.

The presented results of C–H–S gel and pore detection are performed on the basis of microscopic images from paper [7]. The number of pores and the gel can be analyzed. Such calculations make it possible to establish the percentage content of the main microelements at which the best ratio of strength can be observed.

The percentage of C–H–S gel and pores in these microscope images is determined by software. Table 1 allows us to establish the ratio of the number of pores to C–S–H gel in concrete equal to 29.18%, with nano-SiO<sub>2</sub>—3.4% at 1 μm. In the resolution



**Fig. 1** Areas of filling identified by pixel analysis: **a** Concrete mixture with magnification of the microscope 10,000. **b** Concrete mixture with magnification of the microscope 5000. **c** Concrete mixture with nano-SiO<sub>2</sub> particles with magnification of the microscope of the research area 10,000; **d** Concrete mixture with nano-SiO<sub>2</sub> particles with magnification of the microscope of the research area of 5000

of the microscope 2  $\mu\text{m}$ , these trace elements in concrete are 3.78% and in nano-SiO<sub>2</sub>—3.94%, respectively. It is also observed that the number of pores decreases with detail up to 10  $\mu\text{m}$  in a certain area and is equal to 4.39% of the image in the microscope. This trend is caused by working with images of different details and, accordingly, distances to microelements [4].

The number of formed pores and C–H–S gel in the microstructure of concrete depends on the detail of the microscope. The amount of C–H–S gel with the addition of SiO<sub>2</sub> nanoparticles increases, and the number of pores has smaller microscope resolution 1  $\mu\text{m}$  and almost does not change with the addition of these nanoparticles.

## 5.2 *Performing Microstructure Diagnostics of Concrete with Added Nanoparticles in the Process of Curing*

As you know, it is best to compare the results under the same conditions. Figure 2 shows three series of images of changes in the pore structure after 7, 28, and 90 days of curing. The first series is an ordinary cement sample, and the second is cement with

**Table 1** Distribution of the number of pores depending on the resolution of the microscope and materials

Material	Number of pores with microscope resolution 1 $\mu\text{m}$ , %	Number of pores with microscope resolution 2 $\mu\text{m}$ , %	Amount of C–S–H gel with microscope resolution 1 $\mu\text{m}$ , %	Amount of C–S–H gel with microscope resolution 2 $\mu\text{m}$ , %
Concrete without nano-SiO <sub>2</sub>	46.69	27.19	1.6	7.19
Concrete with nano-SiO <sub>2</sub>	34.68	27.23	10.2	6.91

admixtures of ground granulated blast furnace slag (GGBFS). This additional cementitious material is added to Portland cement concrete to help strengthen the concrete. The third series is a combination of cement with GGBFS and CuO nanoparticles.

In the process of curing, the structure of concrete can change.

The detected pores and C–H–S gel are indicated in green and red, respectively, from computer analyzes of the studied areas of concrete. In this case, the resolution of the microscope for all series is 50  $\mu\text{m}$ .

In these results, it is possible to observe changes in the structure of concrete with the addition of various additives and to determine the percentage of C–H–S gel and pores during the curing process.

From Table 2, it can be seen that the addition of GGBFS effectively reduces the size and number of pores compared to plain concrete.

The following Table 3 gives the percentage of C–H–S gels in a cement mixture with slag and CuO and in a mixture of each when these materials are added sequentially.

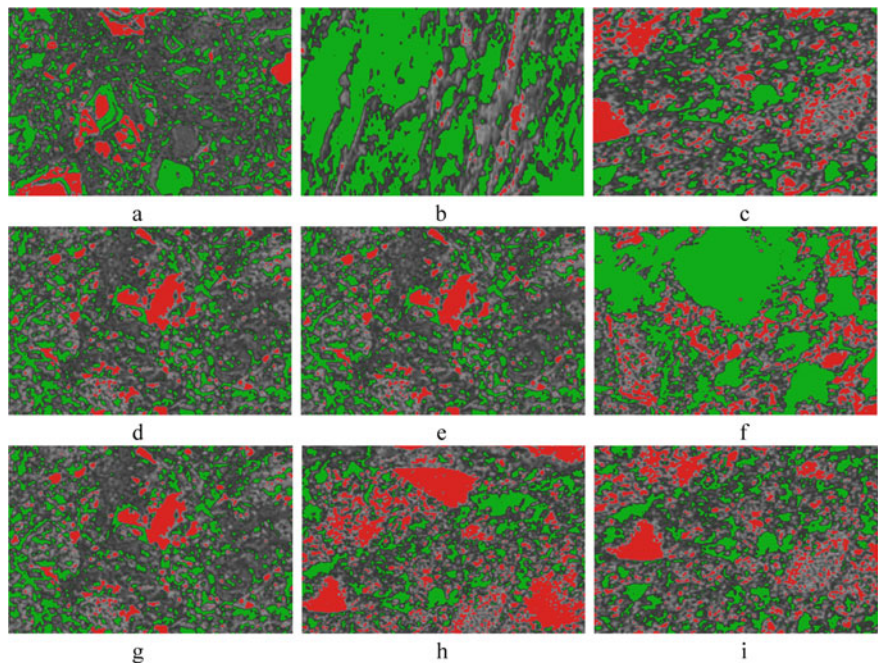
The amount of gel increases the fastest with a cement mixture with slag. In other cases, the values of the percentage amount of gel are almost unchanged. However, in the mixture with nanoparticles, the average value is 11.5. These values are greater than the same values in the cement mixture without the addition of slag and nanoparticles.

## 6 Discussion of the Results of the Study of the Microstructure of Concrete

The paper shows the possibility of conducting pore analysis based on images from microscopes. Since the images are taken from different distances, there are different resolutions of the microscope.

An image from a greater distance may have poorer detail, and the color contrast must be adjusted. The number changes due to the fact that CuO nanoparticles affect the optimization of the concrete mix.





**Fig. 2** Detection of the number of pores and C–H–S gels in: **a** A sample of ordinary cement after 7 days of curing. **b** Sample cement and cement with admixtures of ground granulated blast furnace slag after 7 days of curing. **c** Samples of ordinary cement, cement with admixtures of ground granulated blast furnace slag and CuO after 7 days of curing. **d** Cement sample after 28 days of curing. **e** Sample cement and cement with admixtures of ground granulated blast furnace slag after 28 days of curing. **f** Samples of ordinary cement, cement with admixtures of ground granulated blast furnace slag and CuO after 28 days of curing. **g** Cement sample after 90 days of curing. **h** Sample cement and cement with admixtures of ground granulated blast furnace slag after 90 days of curing. **i** Samples of ordinary cement, cement with admixtures of ground granulated blast furnace slag and CuO after 90 days of curing

**Table 2** Percentage ratio of the number of pores in different cement mixtures

Duration of curing, days	The number of pores in the cement mixture ( $450 \text{ kg}\cdot\text{m}^{-3}$ ), %	The number of pores in the cement mixture ( $450 \text{ kg}\cdot\text{m}^{-3}$ ) with slag ( $202.5 \text{ kg}\cdot\text{m}^{-3}$ ), %	The number of pores in the cement mixture ( $450 \text{ kg}\cdot\text{m}^{-3}$ ) with slag ( $202.5 \text{ kg}\cdot\text{m}^{-3}$ ) and CuO ( $13.5 \text{ kg}\cdot\text{m}^{-3}$ ), %
9	22.82	50.71	16.12
27	16.84	16.81	45.95
90	16.91	15.74	16.51

**Table 3** Percentage ratio of the number of C–H–S gels in different cement mixtures

Duration of curing, days	The number of pores in the cement mixture ( $450 \text{ kg} \cdot \text{m}^{-3}$ ), %	The number of pores in the cement mixture ( $450 \text{ kg} \cdot \text{m}^{-3}$ ) with slag ( $202.5 \text{ kg} \cdot \text{m}^{-3}$ ), %	The number of pores in the cement mixture ( $450 \text{ kg} \cdot \text{m}^{-3}$ ) with slag ( $202.5 \text{ kg} \cdot \text{m}^{-3}$ ) and CuO ( $13.5 \text{ kg} \cdot \text{m}^{-3}$ ), %
9	6.28	1.29	11.59
27	6.38	6.37	11.19
90	6.38	18.27	11.7

The number of pores decreases during curing with a cement mixture without impurities, and the amount of C–H–S gel remains almost unchanged. With the addition of slag, the number of pores also decreases, but the amount of C–H–S gel increases. The lowest indicators of the number of pores and C–H–S gel are observed in the first period of solidification with nanoparticles in a mixture of slag and cement. On the 27th day, there is an increase in the number of pores and a decrease in the amount of C–H–S gel.

It is possible to improve the detailing using the methods of highlighting image contours and more detailed analysis of image objects. The condition for using the described algorithm is the availability of images from the microscope of individual sections of concrete to assess its strength. The main effect of the implementation of the work results is expected to speed up the analysis process due to the use of computer vision.

## 7 Conclusions

The ability to determine the percentage of the main microelements of concrete decreases with an increase in the resolution of the microscope from 1 to 2  $\mu\text{m}$ . For the concrete mixture with the addition of  $\text{SiO}_2$  nanoparticles, the number of pores decreases from 34.68 to 27.23% and C–H–S gel from 10.2 to 6.91%. In the concrete mixture, on the contrary, the number of pores increases from 1.6 to 7.19%, but the amount of C–H–S gel decreases from 46.69 to 27.19%.

Diagnosing the number of pores depending on the distance  $L$  to the studied area of concrete from the microscopic image is performed. It has been established that computer graphics methods can be used to find the structure of nanoconcrete with nanoparticles. Based on information about the color of pixels, it is established that the structure of pores and their number on a certain area of concrete can be dynamically established. It is found that in order to correctly highlight certain areas, the depth of the pores in the image should be taken into account, and for the C–H–S gel, the illuminated surface. In the process of concrete curing, the presence of CuO nanoparticles in concrete leads to a change in the number of pores within the limits of the percentage of pores from 16.12% on the ninth day to 16.51% on the ninetieth day and C–H–S gel from 11.59% on the ninth day to 11.7% on the ninetieth day.



## References

1. Wang X (2017) Effects of nanoparticles on the properties of cement-based materials. Graduate Theses and Dissertations, pp 1–165. <https://lib.dr.iastate.edu/etd/16236>
2. Psakhie SG, Dimaki AV, Shilko EV, Astafurov SV (2015) A coupled discrete element-finite difference approach for modeling mechanical response of fluid-saturated porous materials. *Int J Numer Meth Eng* 106(8):623–643. <https://doi.org/10.1002/nme.5134>
3. Nazari A, Riahi S (2011) Retracted: optimizing mechanical properties of binary blended concrete utilizing CuO nanoparticles. *Int J Damage Mech* 21(1):81–96. <https://doi.org/10.1177/1056789510397074>
4. Tran H-B, Ley V-B, Phan VT-A (2021) Mechanical properties of high strength concrete containing nano SiO<sub>2</sub> made from rice husk ash in Southern Vietnam. *Crystals*. 11(8):932. <https://doi.org/10.3390/cryst11080932>
5. Shafaei D, Yang S, Berlouis L, Minto J (2020) Multiscale pore structure analysis of nano titanium dioxide cement mortar composite. *Mater Today Commun* 22. <https://doi.org/10.1016/j.mtcomm.2019.100779>
6. Al-Faris M, Chiverton J, Ndzi D, Ahmed AI (2020) A review on computer vision-based methods for human action recognition. *J Imag* 6(6):46. <https://doi.org/10.3390/jimaging6060046>
7. Zhuang C, Chen Y (2020) The effect of nano-SiO<sub>2</sub> on concrete properties: a review. *Nanotechnol Rev* 8(1):562–572. <https://doi.org/10.1515/ntrev-2019-0050>
8. Jo BW, Kim CH, Tae GH, Park JB (2007) Characteristics of cement mortar with nano-SiO<sub>2</sub> particles. *Constr Build Mater* 21(6):1351–1355
9. Yildirim M, Derun EM (2018) The influence of CuO nanoparticles and boronwastes on the properties of cement mortar. *Materiales de ConstruCCión* 68(331). ISSN-L: 0465-2746. <https://doi.org/10.3989/mc.2018.03617>
10. Nazari A, Rafeipour MH, Riahi S (2011) The effects of CuO nanoparticles on properties of self compacting concrete with GGBFS as binder department of materials science and engineering. Saveh Branch, Islamic Azad University, Saveh, Iran. <https://doi.org/10.1590/S1516-14392011005000061>
11. Skrynkovskiy RM (2011) Methodical approaches to economic estimation of investment attractiveness of machine-building enterprises for portfolio investors. *Actual Prob Econ* 118(4):177–186. <http://www.scopus.com/inward/record.url?eid=2-s2.0-77952681437&partnerID=MN8TOARS>
12. Skrynkovskiy R (2008) Investment attractiveness evaluation technique for machine-building enterprises. *Actual Prob Econ* 7(85):228–240. <http://www.scopus.com/inward/record.url?eid=2-s2.0-77952681437&partnerID=MN8TOARS>
13. Dinh H, Ha QP, La HM (2016) Computer vision-based method for concrete crack detection, 2016. In: 14th international conference on control, automation, robotics and vision (ICARCV), pp 1–6. <https://doi.org/10.1109/ICARCV.2016.7838682>

# Mild Cognitive Impairment Screening System by Multiple Daily Activity Information—A Method Based on Daily Conversation



Ayaka Yamanaka, Ikuma Sato, Shuichi Matsumoto, and Yuichi Fujino

**Abstract** In recent years, the number of patients with dementia has steadily increased. It is possible to recover from Mild Cognitive Impairment (MCI), which is a preliminary stage of dementia. Therefore, it is important to detect the signs of MCI at home as some characteristics of MCI will appear in daily life. In this study, we defined some features that can detect MCI in daily life when at home by focusing on daily conversations. The acoustic and linguistic features in conversation that were related to cognitive function were selected for analysis. In our experiment, features that were effective in detecting MCI were extracted using the corpus of conversations between healthy controls (HC) and elderly people with dementia, which includes MCI. Based on our findings, effective features to detect MCI were identified. We also used two different machine learning models to discriminate between the two groups using the features effective for MCI detection. We obtained more than 80% correct answers in both cases. In this report, we confirmed the elemental screening method for the TV-based screening system for MCI.

**Keywords** MCI · Screening · Multiple daily activity information · Daily conversation

---

A. Yamanaka (✉)

Graduate School of System Information Science, Future University Hakodate, Hakodate, Japan  
e-mail: [g2121058@fun.ac.jp](mailto:g2121058@fun.ac.jp)

I. Sato · Y. Fujino

Department of Media Architecture, Future University Hakodate, Hakodate, Japan  
e-mail: [ikuma-is@fun.ac.jp](mailto:ikuma-is@fun.ac.jp)

Y. Fujino

e-mail: [fujino@fun.ac.jp](mailto:fujino@fun.ac.jp)

S. Matsumoto

Japan Cable Laboratories, Tokyo, Japan  
e-mail: [s-matsumoto@jclabs.or.jp](mailto:s-matsumoto@jclabs.or.jp)

## 1 Introduction

In recent years, the number of patients with dementia is increasing in Japan, probably due to the increase in the aging population in the country [1]. Elderly people experience declining cognitive function and exhibit various changes in behaviors in daily life [2]. For example, verbal intelligence impairment can be observed in their daily conversations, and attentional-cognitive impairment can be observed in their gait. The symptoms of dementia gradually become more severe as the diseases progress. Although its progression can be slowed down, dementia cannot be cured completely [3]. Early treatment in Mild Cognitive Impairment (MCI), which is the preliminary stage of dementia, may delay the onset of dementia and restores the cognitive dysfunctions [4]. Therefore, the cognitive decline in early stage, i.e., MCI status, needs to be detected and treated.

Cognitive decline in an individual is generally detected when diagnosed by a physician and is treated accordingly at a hospital. In most cases, the elderly people will be aware of the decline in their cognitive function upon diagnosis by a physician rather than by family members, although some families may be aware of the decline in cognitive function. In some cases, the elderly may refuse to visit the hospital or may face difficulty in going to a hospital because of their advanced age even if a family member points out the issue in the elderly's cognitive function. In other cases, the elderly people living alone are likely to be unaware of the decline in their cognitive function because of no family members to point out the issue. In such cases, it is desirable to detect the signs of cognitive decline in the early stages of MCI for elderly people living alone at home.

We believe that it is necessary to develop a method to detect cognitive decline in the daily lives of elderly people who are living alone and to create awareness on the early detection of MCI.

## 2 Purpose

Based on the above background, this study aimed to detect MCI using the information available on multiple daily activities of elderly people when at home. In this paper, we focused on daily conversations as an MCI screening method and examined its applicability to the detection of MCI.

### **3 Related Research**

#### ***3.1 Conversational Assessment of Neurocognitive Dysfunction***

Ohba et al. developed the “Conversational Assessment of Neurocognitive Dysfunction (CANDy)” to assess cognitive function from daily conversations [5]. The CANDy is a normal conversation test that evaluates the frequency of 15 conversational features found in people with dementia. For example, repetitions or vocabulary losses in conversation are a conversational feature in people with dementia. The sensitivity and specificity of the CANDy test are 86.2% and 94.5%, respectively, and a score of 6 or more is considered as a possibility of dementia. CANDy evaluates cognitive function through daily conversations and through correct/incorrect questions, thus reducing the psychological burden on the participants. However, CANDy is a method to detect dementia and cannot detect MCI, which is our purpose of this research.

#### ***3.2 Dementia Screening Test Using Speech Analysis***

In their study on dementia screening using speech analysis, Winterlight Labs asked the subjects to describe the pictures presented on a tablet and then recorded their speech [6]. They extracted and analyzed the linguistic and acoustic features from the collected speech data to identify subjects with dementia. The linguistic features included repetitions of words and parts of speech, while the acoustic features included pitch, tone, and pauses. Based on their analysis results, they could identify people having dementia from healthy controls (HC) with 80% accuracy. However, this method does not classify MCI and is applicable only to English-speaking countries.

#### ***3.3 Dementia Detection by Computer Avatar***

In their study, Tanaka et al. detected dementia based on multiple pieces of information obtained during a conversation with a computer avatar [7]. Features such as acoustic, language, and facial expressions were extracted from the audio and video aspects of the conversation with the avatar. The content of the conversation with the avatar was created with reference to Mini-Mental State Examination (MMSE) and other dementia tests. Based on the logistic regression analysis of the three features, 94% of subjects with dementia were classified correctly. However, this study obtained information by asking subjects to answer questions pertaining to an existing dementia test, involved electronic version of the existing tests, and did not detect MCI through daily conversation, which is the purpose of this research.

4 Methods

Based on the above background and previous studies, we propose a screening system for MCI that considers daily conversations at home.

4.1 The MCI Screening System

In most cases, when a patient develops dementia, he or she refuses to see a doctor because he/she cannot accept the fact that they have dementia. The Captology, proposed by B. J. Fogg to persuade people using a computer technology [8], is one of the methods used to alleviate this problem. We have already proposed a TV-based life support system for elderly people who are living alone. This proposed system incorporates elements of the Captology, presenting a virtual pet (VP) on the screen and facilitating communication with the VP to manage health or prevent dementia [9].

This system displays the VP on a TV that is used daily in the home of elderly people who live alone, inducing conversation and notifying the possibility of MCI via the VP to encourage the patient to go to the hospital. We hypothesize that this system of using VP will reduce the psychological burden for the elderly people and induce them to visit a hospital. In the present research work, we define and extract the effective conversational features for detecting MCI and verify the possibility of discriminating between MCI and HC. We try to extract and analyze the features that indicate cognitive decline from conversations of elderly people living alone at home. By modularizing this extraction and analysis method and equipping it as a function of the Android TV, we intend to detect the decline in cognitive function and promote early medical examination. Thus, we aim to develop the MCI screening system using a home TV that detects MCI at an early stage.

4.2 Two Features for Screening

We defined acoustic and linguistic features related to cognitive function that are required for MCI detection. Table 1 shows the features.

Table 1 Table captions should be placed above the tables

Acoustic features	Linguistic features	
	Quantitative features	Content features
F0, zero-cross, MFCC	Changes in total number of words, repetition of the same word, proportion of the five parts of speech	Co-occurrence relation, recall process

Acoustic features, such as fundamental frequency (F0) and MFCC, were selected based on related studies to supplement changes in speech with changes in linguistic features.

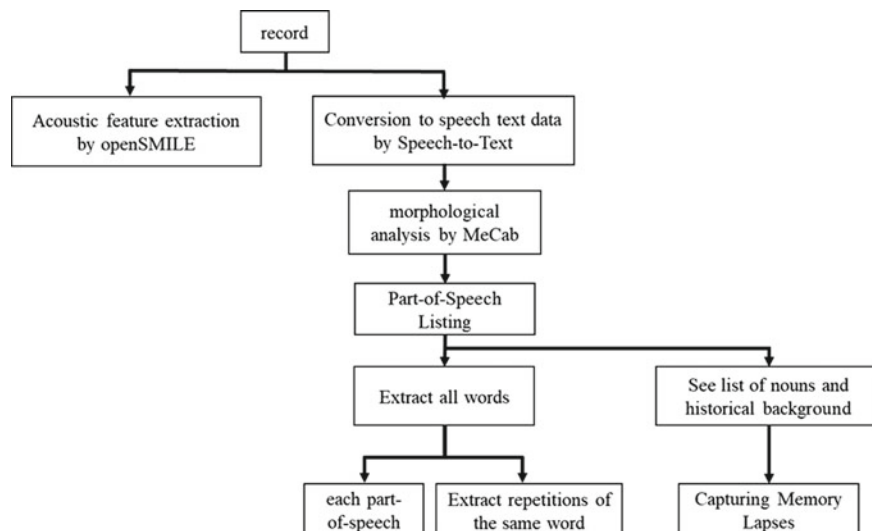
Linguistic features were selected to capture changes in conversational content, emergence of long-term memory, and absence of short-term memory due to cognitive decline. In this study, we classified these features into two types, quantitative and content features. Quantitative features include changes in the total number of words, repetition of the same word, and the proportion of the four parts of speech, i.e., nouns, fillers, adverbs, and adjectives. The content features show the historical background of the nouns, the co-occurrence relation, and the period in which the nouns are frequently used. Co-occurrence relations measured vocabulary loss and connection of conversational content, while the recall process measured whether the participants were conversing using long-term memory.

Co-occurrence refers to the frequent occurrence of a limited number of words in a sentence [10]. We applied these features because we hypothesized that the number of words associated with a certain keyword may decrease due to the decline in vocabulary decline owing to the decline in cognitive function. Recall is one of the basic processes of memory [11] that involves the feeling of having experienced something once, including when and where it was experienced. Episodic memory, which is also called long-term memory, is related to one's past experiences and events and is a part of cognitive function. Impaired short-term memory and episodic memory are the early symptoms of dementia. In the case of impaired short-term memory, the elderly people may fail to recall recent events. In the case of impaired episodic memory, the elderly people fail to remember the key parts of their recent experiences. In the present study, we focused on one of the processes of episodic memory, i.e., recall, in which the details of a past experience, including its content, are recalled. The recall process is susceptible to the effects of aging. Therefore, we focused on the association of words and phrases by co-occurrence and the process of recall.

### ***4.3 Feature Extraction Flow***

The flow of feature extraction is shown in Fig. 1.

Preliminary experiment showed that 6373 features were extracted from speech data using the ComParE2016 set of openSMILE software [12] in order to capture the acoustic features. Google Cloud's Speech-to-Text [13] and morphological analysis using MeCab [14] were used to shape the text data used for extraction in order to transcribe linguistic features. The morphological analysis data were used to calculate the quantitative features. Co-occurrence relations of qualitative features were extracted using a co-occurrence network, which illustrates the distance between words and the number of occurrences of each word. Recall relations were extracted by asking questions using the VP, which expands the conversation based on the historical background. For example, the VP will first start the conversation by asking about the



**Fig. 1** Extraction method for each feature

recent Tokyo Olympic games held in Japan. Next, the VP will ask what the elderly people are doing at that time. We believe that the above results can be used to detect short-term memory deficits by categorizing the contents of the answers according to chronological age and episodic memory deficits depending on the specificity of the content because the Tokyo Olympic games were held two times, one in 1964 and the other in 2021.

## 5 Experiments

In this study, we conducted two types of experiments. The first experiment involved the use of a *t*-test to classify MCI and HC, while the other test used machine learning to show significant differences between the classifications.

### 5.1 Methods of the First Experiment

In this experiment, MCI and HC were discretely classified into binary categories and judged by a *t*-test to clarify the features effective in detecting MCI. We examined both acoustic features and quantitative linguistic features. For the speech data, we used a corpus of elderly people with target groups [15], which was created by Shibata et al. These data consisted of 60 participants, among whom 15 had MCI and 45 were HC. A total of 12 questions were asked to each participant: 10 natural sentence tasks, 1

illustration description task, and 1 animation description task. In the natural sentence tasks, the participants responded to a recent event, while in the illustration task, they responded to a description of the content of an image. In the animation task, the participants responded to a description of the content of an animation. Both audio files and transcriptions of the interviews were used in our analysis.

### 5.1.1 Acoustic Features

With respect to acoustic features, all corpus tasks included 540 data for the HC and 180 data for the MCI groups. The significance levels were set at 5, 3, and 1%

### 5.1.2 Linguistic Features

With respect to linguistic features, a task-specific feature analysis and a task-general analysis were conducted because the linguistic features of the responses differed depending on the task in the corpus. The significance level was set at 1%. The number of data used for the natural sentence task included 450 data for the HC and 150 data for the MCI group, while the number of data used for the illustration description task and the animation description task included 45 data for the HC and 15 data for the MCI group.

## 5.2 Results of the First Experiment

The results for each feature in the first experiment are shown below.

### 5.2.1 Results of Acoustic Features

Table 2 shows the order of the acoustic features that showed the most significant differences between MCI and HC. A total of 9 features showed significant differences, including F0, MFCC, Jitter, etc. With respect to the preliminary experiment, 2871 features were significantly different at the 5% significance level, 2565 features at the 3% level, and 2045 features at the 1% level.

### 5.2.2 Results of Linguistic Features

The results of the linguistic features are shown in Tables 3.

The vertical columns in Table 3 show the kinds of features, while the horizontal rows show the t-value and p-value for each task. With respect to the illustration description task, a significant difference was observed in the “change in total number



**Table 2** Features with significant differences

Rank	Feature's name
1	MFCC
2	pcm-RMSenergy_sma
3	audspec
4	FFT
5	Zcr
6	F0
7	log-HNR
8	Shimmer
9	jitterDDP

**Table 3** Results of *t*-tests of linguistic features

Kinds of features	All tasks		Natural sentence tasks		Illustration description task		Animation description task	
	<i>t</i> -value	<i>p</i> -value	<i>t</i> -value	<i>p</i> -value	<i>t</i> -value	<i>p</i> -value	<i>t</i> -value	<i>p</i> -value
Change in total number of words	— 0.933	0.346	— 1.532	0.127	<b>1.669</b>	<b>0.011</b>	1.896	0.068
Repetition of the same word	0.982	0.327	0.917	0.360	0.703	0.489	0.012	0.990
Proportion of nouns	0.910	0.364	0.278	0.781	1.926	0.063	1.454	0.153
Proportion of filler	1.382	0.168	1.487	0.138	— 1.085	0.289	1.075	0.291
Proportion of adverbs	— 0.122	0.903	0.169	0.982	0.169	0.867	— 0.571	0.573
Proportion of adjectives	— 1.259	0.209	— 1.843	0.067	1.284	0.213	1.157	0.255

of words”. This indicates that the decrease in the number of total words in the MCI group was more than that of the HC.

**5.3 Methods of the Second Experiment**

In this experiment, a machine learning model, i.e., random forest (RF) and support vector machine (SVM), was used to discriminate the MCI and HC groups using the features that showed significant differences in the previous experiments in items of 5.2.1. The generalization performance of the model was evaluated by tenfold cross-validation. The used data were a corpus of elderly people as target groups, similar to that in the analysis method. The explanatory variables included 2045 acoustic

**Table 4** Model classification results (%)

Model	Positive solution rate	Fitness rate	Recurrence rate	<i>F</i> value
Random forest	83.33	84.71	95.68	89.86
Support vector machine	88.89	88.59	97.78	92.96

features that showed significant differences in the feature analysis. The cognitive level was the objective variable in the experiment, and it was set to 0 for MCI and 1 for HC based on the MMSE score in the corpus data. Due to the lack of their sufficient number, the linguistic features were not used in this analysis.

**5.4 Results of the second experiment**

Table 4 shows the experimental results.

The RF and SVM showed 83.33 and 88.89% of correct responses, respectively. The reproducibility of both methods exceeded 90%, and the *F* value was 89.86% for RF and 92.96% for SVM, indicating high classification results.

**6 Discussions**

The results of the above two experiments are discussed as follows. In addition, we would like to describe the furthermore experiment.

**6.1 Discussions of the First Experiment**

A discussion of each feature in the first experiment is as follows.

**6.1.1 Acoustic Features**

With respect to acoustic features, significant differences were found in F0, Zcr, and MFCC features that have been associated with the lack of cognitive functions in the previous studies. Significant differences were also observed in jitterDDP, Shimmer, and other features that were associated with hoarseness. This suggests that cognitive decline may affect voice pitch, silent intervals, and vocal fold change. However, the large number of extracted features may also include features unrelated to MCI detection. So, suitable features need to be extracted from the features that showed significant differences.

### 6.1.2 Linguistic Features

With respect to linguistic features, the significant differences found in the change in total number of words in some tasks suggest that the number of utterances decreases with the decline in cognitive function. However, no significant differences were observed in other features. Thus, new linguistic features, such as frequently occurring words and co-occurrence relationships between utterances, need to be investigated using different extraction methods or analysis methods.

## 6.2 *Discussions of the Second Experiment*

The above results related to the discrimination between MCI and HC confirmed that these classification features were capable of classifying MCI, indicating their effectiveness for MCI screening. No background noise was observed as the conversation recording environment for the corpus was tidy. The present study considers MCI screening from daily conversations at home. Preprocessing, such as noise cancellation, is required when using daily conversations. In the future, classification experiments need to be conducted using daily conversational speech, and preprocessing and hyper-parameters need to be set according to the results of classification experiments.

## 6.3 *Furthermore Experiment*

In the above experiments, we used only quantitative features of conversational speech for classification and showed the results of the two-group discrimination using acoustic features only. The results showed that significant differences in some of the quantitative linguistic features could not be confirmed. Therefore, we have tried to add a quantitative feature, “the proportion of pronouns”, and conducted an additional analysis using the corpus data. As the results showed that ko-so-a-do words were used more frequently in MCI than in HC, it may suggest that MCI causes a decline in vocabulary. “Ko-so-a-do words” mean series of Japanese words that can be used to mention to things, people, and locations. We think this result showed that the “pronoun ratio” was an effective feature for detecting MCI.

Furthermore, we are trying to extract “meanings of words” related to the contents, and at present, we have conducted an additional analysis of “co-occurrence relations” and “contents feature” using the corpus data. As a result, the frequency of occurrence of a variety of words in a conversation was higher in HC than in MCI, indicating that the vocabulary of HC is larger than that of MCI. We think “co-occurrence relation” is also an effective feature for detecting MCI. However, “recall process” is still in the development stage. Currently, we are considering to extract some nouns from which the period can be read from the answers to questions and classifying them according to the period. To further improve the accuracy of MCI detection, the features should

be extracted and analyzed from two aspects of linguistic features, adding the content features. Therefore, we think that these feature extraction and analysis methods need to be integrated in the future classification experiments.

## 7 Conclusion

This paper described the results of a two-group discrimination between MCI and HC after clarifying the feature extraction and effective analysis methods for detecting MCI, with the aim to examine the applicability of MCI detection from conversations at home. Our test results confirmed that 2045 acoustic features showed significant differences. We discriminated MCI and HC from the corpus of speech data using significantly different acoustic features and obtained more than 80% correct responses for both methods. However, since these experiments were used to detect MCI using only the corpus, the classification accuracy from everyday speech needs to be calculated in the future. In addition, features from two planes, acoustic and linguistic features, need to be extracted and analyzed to achieve more accurate MCI detection.

With respect to linguistic features, significant differences were observed in the change in the number of words but not in other features. Therefore, it is necessary to investigate new linguistic features using different extraction or analysis methods. Additional analysis at the present stage revealed that "the proportion of pronouns" is an effective feature for detecting MCI. In particular, the extraction method for content features needs to be clarified, e.g., some nouns related to historical background. For "co-occurrence relation", HC was found to be effective for MCI detection because HC had a larger vocabulary than MCI. However, an extraction method for "recall process" of content features is under development. The method extracts nouns from the responses to questions that can be used to identify the age of the respondents or questions about recent topics. Then, a list of events in each genre was created and labeled by the period, followed by matching the nouns to the period vector. Finally, the responses are categorized by the period based on the list. In this way, short-term memory lapses can be determined.

In the future, we will develop a TV-based in-home MCI screening system that combines the above two feature extraction and analysis methods to detect MCI at an early stage from daily conversations. The system will detect cognitive decline by communicating with the VP on the TV set in daily life. In addition, if the lack of cognitive functions for elderly people living alone is presented and explained via the VP using the Captology, the elderly people can be made to be aware of their cognitive decline and encouraged to see a doctor. Such an approach can result in the early detection and early treatment of MCI.

## References

1. Cabinet office, annual report on the ageing society. [https://www8.cao.go.jp/kourei/whitepaper/w-2020/html/zenbun/s1\\_1\\_1.html](https://www8.cao.go.jp/kourei/whitepaper/w-2020/html/zenbun/s1_1_1.html), (in Japanese). Last accessed 34 Jun 2022
2. Ninomiya T, et al (2016) Study on the future estimation of the elderly population with dementia in Japan, health and labour sciences research grants, special research project for health and labour sciences
3. Japanese Society of Neurology (2017) Dementia guidelines 2017(in Japanese)
4. Takeda, et al (2008) Early detection and early treatment of dementia. *J Jpn Soc Mibyou Syst* 14(1):64–67
5. Oba H, Sato S, Kazui H, Nitta Y, Nashitani T, Kamiyama A (2018) Conversational assessment of cognitive dysfunction among residents living in long-term care facilities. *Int Psychogeriatr* 30(1):87–94
6. Imai E (2019) The latest medical treatment for dementia (in Japanese). *Fuji Med Publish* 9(1):11–16
7. Tanaka H, Adachi H, Ukita N, Ikeda M, Kazui H, Kudo T, Nakamura S (2017) Detecting dementia through interactive computer avatars. *IEEE J Transl Eng Health Med* 5:1–11
8. Fogg BJ (2005) Technology that moves people: what experimental psychology teaches us (in Japanese). Nikkei Business Publications, Tokyo
9. Ohnishi M et al (2019) The life support system for elderly living alone using the home TV. *IEICE Tech Committee Submission Syst* 118(485):133–138
10. Ishida M (2017) Introduction to text mining with R, 2nd ed. (in Japanese), Morikita Publishing Co., Ltd.
11. Brain Science Dictionary, souki gosouki (memory). [https://bsd.neuroinf.jp/wiki/souki\\_gosouki](https://bsd.neuroinf.jp/wiki/souki_gosouki) (memory) (in Japanese). Last accessed 28 Sep 2022
12. Eyben I, Wöllmer M, Schuller B (2010) OpenSMILE - the munich versatile and fast OpenSource audio feature extractor. In: *Proceedings of the ACM Multimedia(MM)*, vol 25. ACM, Florence, Italy, pp 1459–1462. ISBN 978-1-60558-933-6
13. Google cloud, speech-to-text. <https://cloud.google.com/speech-to-text>. Last accessed 07 Sep 2022
14. Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to Japanese morphological analysis. In: *Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP-2004)*, pp 230–237
15. Shibata D, Ito K, Wakamiya S, Aramaki E (2019) Detecting early stage dementia based on natural language processing (in Japanese). *Trans Jpn Soc Artif Intell* 34(4):B-J11\_1-9

# System Models of a Single Information Space of Digital Twins



Mykola Korablyov  and Sergey Lutskey

**Abstract** The paper considers a systematic approach to modeling a single system-information space of virtual digital production. The terms system information, information norm, and information measure introduced in the work correspond to the generally accepted classical definition of the concept of “information”. An analysis of system-information models of processes and systems was carried out, and a definition of the term system information was proposed as a proportional ratio of a general attribute to its particular value, that is, information about the proportion of an information measure is subject to transmission. A methodology for system-information modeling of the information space is proposed. An example of calculating equivalent system information is given and an analysis of methods for solving production problems using system-information models is carried out. The analysis of methods for solving production problems, based on system-information models, showed additional possibilities for processing information in single information space of virtual digital industries, which are based on the developed terms of system information, information norm, and information measures.

**Keywords** System-information models · System information · Information norm · Information measure · Information space · Digital production

## 1 Introduction

The creation of a single information space is an urgent problem for modern digital production. Digital manufacturing involves not only the use of technology to increase productivity, but we are talking about creating a “digital twin” for a product or process, or even the entire enterprise. The Digital Twin concept is designed to help businesses detect physical problems faster, more accurately predict their outcomes, and produce better products.

---

M. Korablyov (✉) · S. Lutskey  
Kharkiv National University of Radio Electronics, Kharkiv 61166, Ukraine  
e-mail: [mykola.korablyov@nure.ua](mailto:mykola.korablyov@nure.ua)

Digital production models (“digital twins”) are multi-level layouts of both technological and production processes, as well as individual technological operations, operating with a huge number of production facilities (equipment, employee workplaces, service departments, and so on). The functioning of such models requires accounting and analysis of a huge amount of heterogeneous data. This is one of the reasons why digital manufacturing required significant development of information technology.

One of the biggest challenges of digital manufacturing is getting the concept of virtual mapping of manufacturing information processes right. In production, separate information modules can be introduced in the form of the Internet of things, 3D modeling of several production lines, and so on. But it still will not be a full-fledged digital enterprise, if information technologies are introduced “separately”, and are separated from each other if there is no “core”—information modeling, based on which functional, logical, and informational links are organized between all technologies: modeling, internet of things, robots, big data processing, etc.

The methodology of system-information modeling is based on the concept of system information [1–3], which is characterized by a quantitative indicator of the communication ability of an object to exchange information with the environment. Product system information is the primary attribute of production, which is laid down at the early stages of the product life cycle and does not change during the entire production cycle. System information on related technologies used, technological processes, and production systems, which may vary depending on the specifics of a particular enterprise, is a secondary attribute. The concept of a unified system-information space of digital production, which is based on the methodology of system-information modeling of production facilities and information laws for the transformation of system information, is a promising basis—the “core” of the digital twin of production [4].

The proposed concept of a single information space for digital products based on system-information models does not require further revision of the orientation of the developed software for enterprises and allows the use of digitalization of numerical data on the parameters of production facilities for system-information models. The use of system-information models in digital twins of production makes it possible to solve problems of analysis and synthesis, management, and forecasting on a virtual level based on the quantity, quality, and value of system information, which characterizes the parameters of production processes and systems.

## 2 System-Information Methods and Models for Solving Production Problems Based on Virtual Digital Twins

### 2.1 System-Information Models of the Systems State

The principles of developing system-information models of processes and systems are based on scientific provisions in the field of information theory, set out in [5]. The information object of study of the system-information approach is the system—a set of elements  $X$  that is in relationships and information links with each other, which forms certain integrity and unity. The system information possessed by the elements of the  $X$  set is characterized by the interval between the upper  $X_{\max}$  and lower  $X_{\min}$  boundaries of its manifestation, as well as the sensitivity threshold  $\Delta x = (x - X_{\min}/n)$ , where  $x = (n\Delta x + X_{\min})$  is a discrete variable value from the set  $X$ ,  $\Delta x$  is the sensitivity threshold of the discrete variable value  $x$  on the interval  $(X_{\max} - X_{\min})$ .

The system information of an object is characterized by an information measure and an information norm [6]. The information measure is a function of the absolute value of the qualitative and/or quantitative proportion of the ratio

$$|I(X)| = f\left(\frac{X_{\max} - X_{\min}}{\Delta x}\right), \quad (1)$$

the information norm is equal to the value of the ratio of the general value of the attribute to its private value

$$\|(X)\| = f\left(\frac{X_{\max} - X_{\min}}{x}\right) = f\left(\frac{X_{\max} - X_{\min}}{n\Delta x + X_{\min}}\right), \quad (2)$$

where  $x = (n\Delta x + X_{\min})$ .

Based on the information measure, we determine the amount of information. Amount of first-kind information

$$\log_2 |I(X)| = \log_2 \left( \frac{X_{\max} - X_{\min}}{\Delta x} \right), \quad (3)$$

where  $\Delta x = X_{\min}/k$ ,  $k = 1 \dots m$ ,  $m \neq 0$ ,  $m \neq \infty$ ,  $X_{\min} \neq 0$ .

Amount of second kind information

$$\log_2 \|(X)\| = \log_2 \left( \frac{X_{\max} - X_{\min}}{x} \right) = \log_2 \left( \frac{X_{\max} - X_{\min}}{n\Delta x + X_{\min}} \right), \quad (4)$$

where  $x = (n\Delta x + X_{\min})$ .

To identify the physical property of an object, the principle is used as an information measure, while fixing the value of the sensitivity threshold of a physical property in the form of a standard of a physical quantity  $\Delta x = 1$  standard, presenting the value



$X_{\min} = 0$ , while  $X_{\max} = x$  is a discrete variable from the set  $X$ . Then the amount of information on the first kind of physical quantity has a view

$$\log_2 |I(X)| = \log_2 \left( \frac{x}{\Delta x} \right). \quad (5)$$

All-natural sciences are built on this principle, in which the ratio of the general to the particular is used as a generally accepted concept of information.

## 2.2 Methodology of System-Information Modeling of the Information Space

Interacting systems,  $A$  and  $B$  can be given by sets of stable and changing characteristics in them. The first set its certainty on the considered (finite) time interval, and the second—is the behavior. The unstable characteristics of systems include:

1. A finite set of actual perceived (directly or indirectly) properties of the acting system at a certain time interval  $E^A = \{e_n^A\}$ ,  $E^B = \{e_m^B\}$ ,  $n = 1 \dots k$ ,  $m = 1 \dots g$ .
2. Upper and lower bounds for perceiving the intensity of each graded property  $\Lambda(e_n^A)$ ,  $\lambda(e_n^A)$ ,  $\Lambda(e_m^B)$ ,  $\lambda(e_m^B)$  (sensitivity threshold).
3. The maximum scales of the duration  $\Gamma(T)$  and the length  $\Gamma(R^n)$ , perceived by the systems as “now” and “here”.
4. Minimum scales of duration  $\gamma(T)$  and extent  $\gamma(R^n)$  distinguishable by systems (distinction limit).

The first of these characteristics is associated with the perception of the qualitative certainty of systems, and the next three—are their quantitative certainty. The stable characteristics of systems include the limit values of non-stable characteristics:

1. A finite set of potentially perceived by systems  $A$  and  $B$  properties of influencing systems

$$E^{*A} = \{e_n^A\}, \quad E^{*B} = \{e_m^B\}. \quad (6)$$

2. Limit upper and lower limits of perception of the intensity of each property with gradation

$$\Lambda^*(e_n^A) = \sup\{\Lambda(e_n^A)\}, \quad \lambda^*(e_n^A) = \inf\{\lambda(e_n^A)\}, \quad (7)$$

$$\Lambda^*(e_m^B) = \sup\{\Lambda(e_m^B)\}, \quad \lambda^*(e_m^B) = \inf\{\lambda(e_m^B)\}. \quad (8)$$

3. Limit maximum scales of duration and extent

$$\Gamma * (T) = \sup\{\Gamma(T)\}; \quad \Gamma * (R^n) = \{\Gamma(R^n)\}. \quad (9)$$

#### 4. Limit minimum scales of duration and length

$$\gamma^*(T) = \{\gamma(T)\}, \quad \gamma^*(R^n) = \{\gamma(R^n)\}. \quad (10)$$

The above characteristics are structured depending on the type of system: inorganic, biological, or artificial. This structuring determines the value of the system. We define the universe  $U$  by a finite set of properties

$$E^{*A} = \{e^A\}, \quad E^{*B} = \{e^B\}. \quad (11)$$

Then it is possible to order all the boundaries of the ranges of duration and length  $\Gamma(T)$ ,  $\gamma(T)$ ,  $\Gamma(R^n)$ ,  $\gamma(R^n)$ , in which these properties are manifested. The universe  $U$  can be considered as a set of layers of duration  $U = \{S_i(T)\}$ , or as a set of layers of extent  $U = \{S_j(R^n)\}$ , each of the scales  $T_i$  and  $R_j^n$  is the boundary between the corresponding layers  $S_i(T)$ ,  $S_{i+1}(T)$ , or  $S_j(R^n)$ ,  $S_{j+1}(R^n)$ .

By definition, when moving from one layer to another, at least one property either appears or disappears. Any pair of layers of duration and extent may either have no common properties at all, or have a larger or smaller number of them. In the latter case, one can speak of a greater or lesser correlation between these layers. With a significant correlation, it is possible to combine the corresponding pair of layers into one layer of duration-length  $S_{ij}(T, R^n)$ , characterized by many properties  $\{e^*(ij)\}$ ,  $\{e^B(ij)\}$ .

A reflection by the system of one or other property from  $\{e_n^A\}$ ,  $\{e_m^B\}$  and determination of its intensity is possible if there is a property in the reservoir of duration-extent  $S_{ij}(T, R^n)$  within one minimum scale of duration  $\lambda(T_i)$  or extent  $\gamma(R_j^n)$ . Within the scales  $\Gamma(T_i)$  and  $\gamma(T_i)$  properties  $\{e^A\}$  and  $\{e^B\}$  can show qualitative certainty in the form of variability (constancy, variability). Quantitative certainty will be expressed by spectra of such scales  $T_k(e_{ij}^A)$   $dT_k(e_{ij}^B)$ , upon passing through which the intensity of properties will change. The introduction of such secondary characteristics of properties means the transfer of time and spatial descriptions of systems  $A$  and  $B$  from purely quantitative to qualitative-quantitative.

Based on the foregoing, each range of duration scales  $(T_{k-1} \div T_k)$  or extensions  $(\frac{R_{n1}}{S_{-10}} \div R_s^n)$  can be associated with a layer of duration  $L_k(T)$  or extension  $L_s(R^n)$ , in which any properties  $\{e_i^A\}$ ,  $\{e_j^B\}$  or  $\{e_i^A\}$ ,  $\{e_j^B\}$  can be constant or variable, homogeneous or inhomogeneous. When passing from the layer  $L_k(T)$  to the layer  $L_{k+1}(T)$ , or from the layer  $L_s(R^n)$  to the layer  $L_{s+1}(R^n)$ , at least one of the constants  $\{e_{ij}^A\}$ ,  $\{e_{ij}^B\}$  become variable, or at least one of the homogeneous properties becomes heterogeneous (or vice versa).

Let us introduce the concept  $b_i$  of a layer of duration  $B_k(T)$  and extent  $B_s(R^n)$  as a set of adjacent layers  $L_k(T)$ ,  $L_{k+1}(T)$  and  $L_s(R^n)$ ,  $L_{s+1}(R^n)$ . Its characteristic feature is that in the corresponding range of scales of duration  $T_i$  or extent  $R_j^n$ , at

last one property of  $\{e_{ij}^A\}$ ,  $\{e_{ij}^B\}$  acts a constant or variable, or as homogeneous or inhomogeneous.

With this approach to the analysis of systems, duration and extent are inherent characteristics of systems  $A$  and  $B$  and are variable, heterogeneous, discrete, and finite. Their definiteness requires an indication of the ranges of duration and extent, in contrast to homogeneous continuous and infinite in time and space, the definiteness of which requires an indication of a reference system.

The use of duration structuring and length makes it possible to consider a different level of “objectivity” of the interaction of two systems. If systems are considered in terms of duration and extent, covering two boundaries between layers, that is

$$\gamma(T) < T_{k-1} < T_k < \Gamma(T), \quad \gamma(R^n) < R_{s-1}^n < R_s^n < \Gamma(R^n), \quad (12)$$

then in this case the influencing system can form changes in the state of the system on which it acts, as well as its structure. This becomes possible if, within the limits of the duration and extent of the system allocated to the block  $B_{ks}$ , variable and inhomogeneous properties appear that are not fixed beyond these limits. Then they can act as variable properties of the given system. The use of the concept of the state of the system in solving system problems assumes that the values of the output (reaction)  $y$  through the inverse mapping  $\eta^{-1}$  uniquely characterize the state of the system in the implemented process of the system functioning. Thus, any parameter of the system (external, internal) characterizes its property—the qualitative side, and the value of the parameter—quantitative.

The amount of information that carries the qualitative value of a finite set of properties potentially perceived by the system is equal to

$$I_E = \sum_{n=1}^d \log_2(d - n), \quad n = 1 \dots d, \quad E = \{e_n\}. \quad (13)$$

The amount of information that is physically (technically) carried by the parameters of the system properties is equal to

$$I_X = \sum_{k=1}^g \log_2(g - k), \quad g = \frac{\Lambda(e) - \lambda(e)}{\lambda(e)}, \quad k = 1 \dots g. \quad (14)$$

The amount of information that carries the duration during which the properties of the system manifest is equal to

$$I_T = \sum_{g=1}^e \log_2(e - g), \quad e = \frac{\Gamma(T) - \lambda(T)}{\gamma(T)}, \quad g = 1 \dots e. \quad (15)$$

The amount of information that carries the extension, in the space of which the properties of the system are manifested, is equal to

$$I_R = \sum_{b=1}^A \log(t - b), \quad f = \frac{\Gamma(R^n) - \gamma(R^n)}{\gamma(R^n)}, \quad b = 1 \dots f. \quad (16)$$

The system-information equation is a function of a complex information indicator

$$S_I = f(I_E, I_X, I_T, I_R), \quad (17)$$

where  $I_E$  is the set  $E$  of properties that have system information;  $I_X$  is the amount of information possessed by the intensity  $X$  of the system properties,  $I_T$  is the amount of information possessed by the duration  $T$  of the system properties,  $I_R$  is the amount of information possessed by the extent  $R$  of the system properties.

The above expressions for calculating the amount of information allow us to evaluate the quantitative ( $I_X, I_T, I_R$ ) and qualitative ( $I_E$ ) properties of the system.

### 3 Analysis of Production Tasks and Methods for Their Solution Using System-Information Models

The methods used to solve production problems based on a digital twin depend on the type and subspecies of the system-information model of processes and systems used. System-information models are divided into four types:

1. Absolute system-information model (ASIM).
2. Relative system-information model (RSIM).
3. Equivalent system-information model (ESIM).
4. Mixed system-information model (MSIM).

Subtypes of system-information models differ from each other by the sensitivity threshold function, which is used in a certain form of the system-information model of processes and systems. System-information models are divided into subspecies.

1. When the sensitivity threshold of the physical property of an object is a function of a single standard of a physical quantity  $\Delta x = f(1Et)$  [a unit of a standard of a physical quantity (PHQ)]. In this case, the system information characterizes the numerical value of the values of the physical property in units of the standard.
2. When  $\Delta x = f(IT)$  is a function of the tolerance on the accuracy of the parameter. The system-information model acquires the characteristics of parameter information. The higher the IT tolerance for the accuracy of a parameter, the more complex the technology of its reproduction. In this case, system information characterizes the technological costs of the product.
3. When  $\Delta x = f(PL)$  is a function of the Planck unit, and  $X_{\max}$  is the desired variable. Since the values of the Planck units are derived from the fundamental physical constants, in this case, the system information characterizes the optimal numerical value of the properties of objects and their equilibrium in the system.

4. When  $\Delta x = f(U)$  is a function of the extended uncertainty interval of the value of the physical quantity. In this case, system information characterizes the probability of the quality of the system under conditions of uncertainty.
5. When  $\Delta x = f(\mu A(x))$ ,  $x \in X$ , where  $\mu A(x)$  is the membership function, then the system-information model acquires the characteristics of fuzzy information.
6. When  $\Delta x = f(p)$  is a probability function. The proportion of the ratio of the general ( $m$  is the number of observations) to the particular ( $n$  is the number of occurrences of an event) was used by K. Shannon in his formula for calculating the amount of information  $I(p) = \log_2 1/p$ ,  $p = n/m$ , the information measure is  $|I(p)| = 1/p = m/n$ ,  $m$  is general,  $n$  is particular.

Thus, the type and subtype of the system-information model requires the development of various methods for solving the production problems of digital production, which are determined by specific requirements and production conditions.

The main production tasks using system-information models of processes and systems in a single information space of digital production can be attributed.

1. Forecasting. Forecast of technical and economic indicators (KPI) and the amount of resources required to launch a new product into production in the early stages of the life cycle based on system information of the design documentation (DD) parameters of a new and old product.
2. Optimization of new production. Optimization of resource costs for technological preparation of production based on the parameters of design documentation (DD) and technological system (TS) of the existing production.
3. Pre-production. Development of technological processes for processing a product and selection of methods and control parameters based on the parameters of design documentation (CD) of a new product and existing technological processes (TS) of production.
4. Control and management of production processes. Control of the parameters of the product processing modes and optimization of the control parameters of technological equipment in real time based on the parameters of the product and the technological process (TP) of production.
5. Monitoring of technical and economic indicators (KPI) in real time and optimization of the corrective control of the technological process (TP) of production. Analysis of controlled parameters of finished products to identify technological reserves of production.

The main methods for solving technological problems using system-information models of processes and systems in single information space of digital production can be attributed.

1. Methods based on absolute system-information models of design documentation parameters and receiving statistical control parameters, which are used to control at the level of technological equipment for processing processes.
2. Methods based on relative system-information models that characterize the efficiency of technological equipment and are used to control production at the level of the technological process.

3. Methods based on mixed system-information models that are used to manage production at the enterprise level and ensure production efficiency. These methods make it possible to compare the efficiency of various production links in terms of resource costs per one bit of system information of the manufactured product.
4. Methods based on equivalent system-information models, which are used by the marketing, planning, and accounting services of an enterprise to assess the competitiveness of a product. Also, these models are used by research and development departments to assess the quality of design characteristics and optimize the consumption of production resources in the manufacture of this product.

The use of software in a single information space of digital products based on system-information methods and models allows you to automate the production tasks of forecasting, optimizing, and evaluating production efficiency at the levels of the workplace, site, workshop, and production in real time.

The list of tasks and methods for their solution for a single system-information space is established in the process of creating a specific digital twin of production when setting goals and setting tasks for real production.

Using the proposed approach, methods were developed for calculating system-information models (SIM) of controlled parameters at the level of design documentation and statistical control.

1. Method for calculating the characteristics of system-information models of controlled production parameters based on design documentation (DD)

$$I_{\text{param}} = \log_2 \frac{X_i}{\text{TI}_i}, \quad (18)$$

where  $X_i$ —the controlled parameter of the DD,  $\text{TI}_i$ —the accuracy tolerance of the parameter.

#### 1.1 Absolute system-information model (ASIM) of DD parameters

$$\text{ASIM}_{\text{DesDoc}} = \sum_{i=1}^k \log_2 \frac{X_i}{\text{TI}_i}, \quad i = 1 \dots k, \quad (19)$$

where  $k$  is the number of controllable parameters of the DD.

#### 1.2 Relative system-information model RSIM

$$\text{RSIM}_{\text{DesDoc}} = \frac{(\text{KPI}_{\text{plan}})}{\text{ASIM}_{\text{DesDoc}}}, \quad (20)$$

where  $\text{KPI}_{\text{PLAN}}$ —planned technical and economic indicators of production.

2. Method for calculating the characteristics of system-information models of production based on acceptance statistical control.

#### 2.1 The arithmetic mean value of the controlled parameters of the product

$$X_{\text{aver}} = \frac{\sum_{i=1}^n X_i}{n}, \quad (21)$$

where  $n$  is the number of controlled statistical parameters.

## 2.2 Scattering range of qualitative characteristics

$$R_j = X_{\text{max}} - X_{\text{min}}. \quad (22)$$

## 2.3 Absolute system-information model (actual) ASIM

$$\text{ASIM}_{\text{fact}} = \log_2 \frac{\sum_{i=1}^n X_{\text{aver}}}{\sum_{j=1}^l R_j}. \quad (23)$$

## 2.4 Relative system-information model (actual) RSIM<sub>FACT</sub>

$$\text{RSIM}_{\text{fact}} = \frac{\text{KPI}_{\text{fact}}}{\text{ASIM}_{\text{fact}}}. \quad (24)$$

The results of development and research (algorithms and software products) to improve the technical and economic indicators were tested in production. The developed algorithms for solving technological problems are based on system-information models of technological methods for processing parts by cutting based on the controlled parameters of the product and the technological process. Algorithms functionally link the accuracy and quality of parts processing with the technical and economic indicators of the technological process, which are used for the economic analysis of machine-building production.

# 4 Conclusions

The article considers a systematic approach to modeling a single system-information space of virtual digital production. The proposed methodology for modeling the information space does not contradict the basic principles of modern virtual digital production but complements it with those features that are methodologically incorporated into it.

The introduction of the terms system information, information norm, and information measure in the work correspond to the generally accepted classical definition of the concept of “information”. The concept of conditions and principles for the emergence, transmission, and processing of system information is considered. The analysis of system-information models of processes and systems is carried out, and the definition of system information as a proportional ratio of a general attribute to its particular value is proposed. The methodology of system-information modeling of the information space is proposed.

The analysis of methods for solving production problems using system-information models was carried out, which showed additional possibilities for using intelligent information processing in virtual digital twins, which are based on the developed terms of system information, information norm and measure.

Note that system-information models of processes and systems describe the interaction of objects at the information level. The complexity of determining the state of systems lies in the uncertainty in measuring the quantitative and qualitative characteristics of the true values of physical quantities that reflect the laws of the properties of the environment. This can lead to distortion and loss of information due to a possible change in the characteristics of the process or system. Therefore, as a direction for further research, it is supposed to obtain a description of the information interaction of objects under uncertainty using models of intelligent information processing, in particular, neural network models or fuzzy inference models. Their use will improve the efficiency of system-information modeling of the information space.

## References

1. Prakash N, Prakash D (2020) Novel approaches to information systems design. Hershey, New York
2. Hasan FF (2018) A review study of information systems. 17(18):15–19
3. Ferguson B (2012) Key stages of strategic information system planning (SISP) methods and alignment to strategic management planning concepts. Columbia Forest Products, Greensboro
4. Lutsky S (2021) System-information approach to uncertainty of process and system parameters. *Innov Technol Sci Solut Ind* 3(17):91–106
5. Egorov AA (2001) Application for discovery. In: International association for scientific discovery. Russia. No. A-242 dated December 28
6. Korablyov M, Lutsky S (2022) System-information models for intelligent information processing. *Innov Technol Sci Solut Ind* 3(21):6–13



# Creating a Happy Life Through Body Sensations



Shuichi Fukuda

**Abstract** The World 1.0, the world of the industrial society, is getting close to its ceiling and many issues are emerging, such as decreasing labor force due to decreasing childbirth and aging population. And although AI is expected to contribute to the next generation, it consumes 10,000 times more energy than that of a human brain. So, although AI is indeed wonderful, we need to pinpoint the applications, which cannot be solved otherwise. Thus, it is time now to develop another world for the next generation. We make our life happy by “self-sustaining” and “self-satisfying”. Then, we can reduce the burden to the “society” to the minimum. Still, we can enjoy the maximum happiness and the feeling of achievement. To achieve this, we need to pay attention to our “instinct” and utilize it to the maximum. In short, we should pay attention to the body sensations and make the most of our proprioception.

**Keywords** Next world · Self-sustaining and self-satisfying · Body sensations · Instinct · Decision making · Ordinal · Mahalanobis distance · Pattern

## 1 Digital and Analog

Digital transformation (DX) is getting wide attention these days. But we should remember that the purpose of DX is to process things faster, especially with the current number-based computers. The progress of computers are impressive and remarkable. But it is mainly due to the progress of hardware.

The industrial society is product-based, so we have been paying efforts to establish the way to evaluate value a reasonable, objective and quantitative basis. And the industrial revolution introduced “division of labor” and it changed our working style from working for ourselves to working for others to enable mass production. This shift satisfied our material needs. But what characterize humans is we can see the

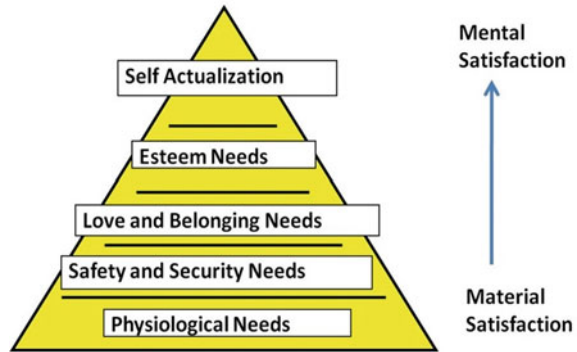
---

S. Fukuda (✉)

Keio University, SDM Research Institute 4-1-1, Hiyoshi, Yokohama 223-8526, Japan

e-mail: [shufukuda@gmail.com](mailto:shufukuda@gmail.com)

**Fig.1** Maslow's human needs



future. In old time, we worked for ourselves. But lining things other than humans worked for now. But as we, humans, can see the future, we wanted to make our dreams come true. We live for tomorrow. Yes, we worked for ourselves as other living thing did. But what we did was to challenge. Challenge is the core and mainspring of all human activities. But in the industrial society, there were no many challenging jobs. We repeat the same jobs again and again from day to day.

Abraham Maslow published a paper on human motivation [1] and clarified that “self-actualization” is our highest need. We want to demonstrate how capable “self” is by challenging many possibilities (Fig. 1).

About forty years later, Edward Deci and Richard Ryan proposed “self-determination theory” [2]. They made it clear that we get the maximum happiness and the feeling of achievement, when we do the job internally motivated and self-determined. And no external reward can provide this level of happiness and the feeling of achievement. And they also pointed out that pursuing this maximum happiness and the feeling of achievement plays an important role in our growth.

## 1.1 *Finite Element Analysis and 3D Printing*

The discussion about analog and digital reminds us of two very important tools.

### **Finite Element Analysis**

One is finite element analysis (FEA). Until FEA was developed, we were at a loss how to deal with the problem of analyzing the objects. Most of them are analog and continuous. So, we were forced to solve the problem case by case. But FEA provided us with the comprehensive approach. This was a big surprise and brought us the feeling of relief and happiness.

This is nothing other than DX. In this case, D implies discrete, but in fact it is digitalization. Not in the sense of current digitalization, which aims to reduce the number of degrees of freedom to one, or cardinal number based. But FEA is to change analog data to be able to be processed by discrete mathematics.

### 3D Printing

The other is 3D printing. This goes the other way, from digital to analog. Now we can process digital data very effectively. But most objects are analog. We need a tool to assemble data to form an object which is analog. 3D printing emerged to satisfy this need of ours.

But 3D printing brought us another big surprise. The industrial society was product-centric. We focused our attention to products, its performance, cost-effectiveness, etc. But it changed our perspective. It reminds us of the joy of process. 3D printing shifts our society from product-focused to process-focused. We realize how process provides us with the joy of creation.

Lego is the pioneer. Danish value individuality and personal development. Lego is just pieces of plastics. But we enjoy to create what we wish to realize. Cost? Nothing. Pleasure and Joy? Maximum as Maslow and Deci and Ryan pointed out. In fact, the name of Lego is excellent. Lego means “I put together” in Old Latin and it is the current Danish word with the same meaning. We really put many pieces together to realize what we wish.

Industries are trying to utilize 3D printing to enhance Industries. But frankly, they forgot to how we would like to actualize ourselves. Happiness and Joy vary from person to person. What we want is to make our own dreams come true. We want to “work” for ourselves. Such activity as “makers” demonstrates how we want to actualize ourselves. It should be emphasized that business people need to change their perspective and realize that the society is shifting from product to process. It will bring them much better market and they can produce much higher market share. It will bring Win–Win for the producer and the customer. As the name “customer” tells us we want to customize our way.

## 2 Creatures

Living things are called “creatures”. Why? It is because we create movement to survive. Animals create movements to satisfy their needs for now. But we, humans, can see the future. We live for tomorrow. We want to make our dreams come true. We want to customize our way of life.

We must remember there are two movements in another sense. Movement by “reflex” and movement by “awareness”. Reflex movement is nothing other than just an action without any consciousness. Our body reacts to the outside stimulus. The signal is processed when it reaches spinal cord in the midway to the brain.

But other movements reaches the brain and we become aware of the environment and situations and we make decisions what actions we should take.

Thus, we literally create movement to adapt to the changing outside world. We must remember that although it is widely discussed that our world is coming to change very frequently, extensively and unpredictably, we have experienced such contexts since we are born.

Jean Piaget developed theory of cognitive development [3] and made it clear that babies learn by themselves how they should adapt to the changing outside world. And Betty Edward made it clear [4] that children draw sketches as they see up to seven years old. But after seven, they start to draw on concept. We change from concrete world to abstract world. Why does such change occurs? It is because as we grow, data around us increases tremendously in amount and in variety, so we need to reduce our burden. But this is to make a decision to take better actions. But most of the tools we have are effective for processing data, but the tools for decision making in real sense is very few.

Then, how can we cope with the continuously changing environments and situations which we encounter everyday in our daily life? Yes, we utilize our instinct. But we have forgotten its importance until now.

### 3 Cardinal and Ordinal

We have been focusing our attention to how we can solve the problem, i.e., to How. But for decision making, what matters is to prioritization. We should focus on goal finding, or What and Why. To describe it another way, our primary world has been cardinal. But we need to develop ordinal approaches for making decisions.

But “instinct” itself is not elucidated yet. So, what we can do right now for decision making is not to study instinct and make it clear how it works, but to support instinct to work better.

#### 3.1 Mahalanobis Distance (MD)

Most of our current approaches are cardinal and are processed in Euclidean space. It requires orthonormality among datasets and interval-scale with units. We must remember that such units as weights and length. Come from our physical experience. Weight, for example is defined based on our experience of lifting things up things and we understand what is weight and provide unit for it. This works well for data processing. But to make decisions, we need ordinal approach, or Non-Euclidean approach.

Mahalanobis developed non-Euclidean ordinal approach [5]. He is a researcher of design of experiments and he wanted to remove outliers and improve his dataset for his work. To remove outliers, we need to prioritize which one to remove first, second, ... . So, he developed ordinal approach.

It is very simple. It indicates how far away the point P is from the mean of the current single dataset we are dealing now. So, we can remove outlier first which is the most far away. Then, we can remove the second far away outlier. We keep on removing until we remove all outliers in an ordinal way.

## 4 Intelligence A B C D E

When we hear the word “intelligence”, what comes up in most people’s mind is brain intelligence, which is “knowledge”-based. But “intelligence” means “understand”. This “understand” implies “perception” or “awareness”. As materials are getting softer and softer with the progress of material engineering, we became aware of the environment and situation and make decisions to take appropriate actions.

In fact, there are many “intelligences”. For example, AI, artificial intelligence is a buzzword today. BI, business intelligence, is another buzzword in business sector. In the case of CI, there is no such well-known one. Creative intelligence is one which comes up. Creative intelligence is unique, because other intelligences are focused on data processing or How. Creative intelligence focuses on goal finding or what and why. Robert Sternberg proposed “Triarchic Theory of Intelligence” [6] relating to creative intelligence. D is known as decision support system. But we may call it decision support intelligence. As described, intelligence means to perceive and understand the context. But decision making is increasing its importance today. So, it would be better to call it decision support intelligence. Emotional intelligence is well known.

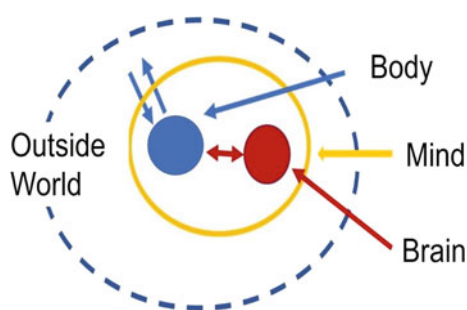
As emotional intelligence is deeply related with the subject of this paper, it will be discussed in detail here. When we say emotional intelligence, we often think that it is a matter of an individual. But we must remember that when we form a group, another emotion emerges. In fact, why we get excited when we watch team sports, we, all watchers, get excited and this excitement is transmitted to the players and they get excited, too. Yes, emotion plays a very important role in communication. It transfers from person to person and from group to group.

It is true non-verbal communication and it is not transmitted by movement, but by air. Communication by air is rarely discussed in technology or engineering. But it is communication in real time. And it is deeply associated with heart. Brain processes knowledge. But it may be called “reflex”, i.e., “mental reflex”.

## 5 Brain and Heart

Brain is getting wide attention these days. But we must remember that our true death comes when our heart stops working and there is no blood flow. Even after brain death, heart is working and blood is flowing. So, even after brain death, we can transplant organs.

Brain corresponds to discrete mathematics. Network and graph play a basic role. Heart, on the other hand, corresponds to continuous mathematics. It deals with analog. And when it comes to emotion, heart plays a leading role. We can easily understand this if we recall William Wordsworth poem “My heart leaps up when I behold a rainbow in the sky” [7].

**Fig. 2** Mind-body-brain

Heart responds to the outside world in real time. Brain, on the other hand, receives information from body and structure them into knowledge. Thus, brain intelligence is knowledge intelligence. But heart intelligence is “body intelligence”, which is nothing other than “wisdom” (Fig. 2).

## 6 Outside and Inside

As described earlier, living things are called “creatures”, because we create “movement” to survive. But we must remember that there are two movements. In the case of human, external movement is called motion. And internal movement is called motor, which includes muscles, etc.

Haptics is getting wide attention these days, because materials are getting softer and softer, so that we cannot understand what the object is and how we should handle it. Thus, haptics becomes increasingly important. But it is only skin sensation. We need to “feel” with our whole body to recognize what it is and to understand how we should handle it. Thus, the “body sensations” are rapidly increasing its importance. Westerns divide outside and inside. Walls, therefore, play an important role.

The phrase “Cross the wall” describes “western culture”. Japanese culture, on the other hand, does not divide outside and inside. What matters to Japanese is the roof, not the wall. “Roof” protects us from the outside world. And there is no wall between outside and inside (Fig. 3).

“Roof” in Japanese house has the same meaning as “umbrella” in western culture. It protects us, but it also serves to share the feelings.

## 7 Growing Worlds

Our worlds are growing and shifting from one world to another (Fig. 4). World 1.0 indicates the industrial society. It is now getting close to its ceiling. So, many issues are emerging, such as decreasing labor force due to decreasing childbirth and aging

**Fig. 3** Japanese house  
(seiken.jp)



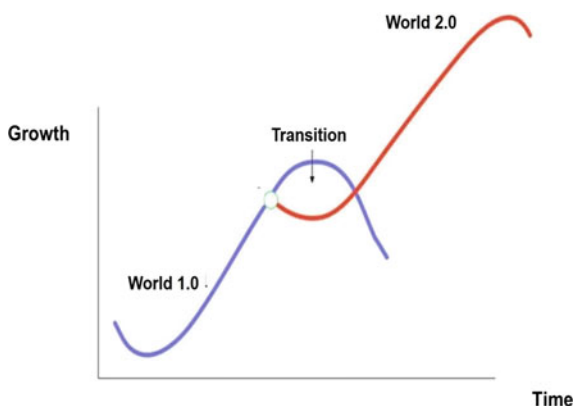
population. AI, which is expected to solve many issues, consumes 10,000 times energy than a human brain. So, it is indeed a wonderful tool, but we need to narrow down and pinpoint the issues which really need AI.

As we have been working for others the industrial society, we could not satisfy our human needs. Maslow pointed out that “self-actualization” is our highest need and Deci and Ryan proposed SDT and made it clear that when we do the job internally motivated and self-determined, we get the maximum happiness and the feeling of achievement. And it not only provides the maximum, but it is necessary for us to grow.

Therefore, it is time for us now to develop the human-focused World 2.0 and satisfy our basic needs as the human. In developing such a world, we need to broaden our perspective and make everybody happy. To achieve this goal, we need to go the other way. Instead of advancing technology and making it higher and higher, we need to introduce “low technology”. We need to shift from hi-tech to low-tech.

To describe this another way, it is “exploration”. As Theodore Roosevelt told us, we need to “do what we can, with what we have, where we are”. It is frequently pointed out that today changes occur frequently, extensively and in an unpredictable manner, but we are experience this in our daily life. Everyday, our environments and

**Fig. 4** Growth curve



situations change continuously. But we lead a daily life. Babies teach us why. They start to crawl and walk by themselves. They use their inborn “instinct”.

But “instinct” is analog and continuous. We cannot process it the way we are doing now with other things. What we are doing now is to process based on the experienced world. But babies are challenging the unexperienced world. They are striving to gain new experiences.

Then, how can we do that? What we can do is to leave the job to our “Instinct”. But we help them make better decisions. And we should not forget that it is not rational, objective and quantitative. It is emotional, subjective and qualitative. We need to develop ordinal approach.

## 8 Mahalanobis Distance-Pattern (MDP): An Instinct Support Tool

Thus, an instinct support tool “Mahalanobis distance-pattern (MDP)” is developed.

Let me explain by taking swimming as an example. Water is changing continuously, so we cannot identify parameters and cannot apply mathematical approaches. This is the situation of the real world today.

Put wearable sensors on the swimmer or take motion images. Then, we can produce such a data sheet shown on the right. Each row corresponds to muscle at each location.

We calculate MD between time T1 and time T2. If MD is decreasing, we know we are moving that muscle in the right way. But if MD is increasing, we need to change its movement. As muscles are analog and we cannot process analog data, we leave to our instinct how we coordinate our muscles and balance our body to swim. Thus, the decision is left to our “instinct”, but MDP provides a guideline how we should improve swimming (Fig. 5).

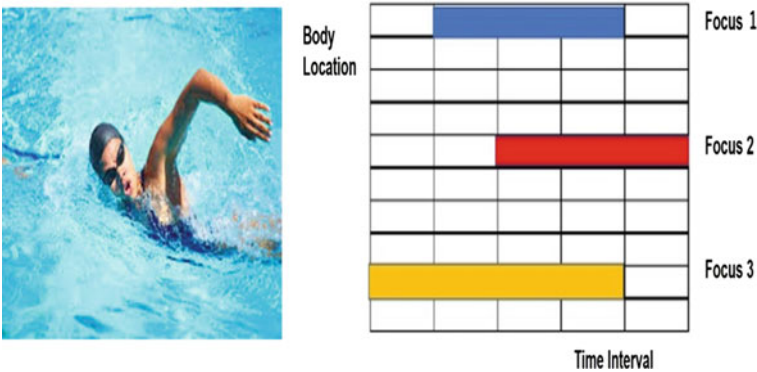


Fig. 5 Mahalanobis distance-pattern (MDP)



## 9 Make Your Life Happy Body Sensation Way

The World 2.0, which comes after the industrial society, will be the world of DIY. We make our life happy by “self-sustaining” and “self-satisfying”. Then, we can reduce the burden to the “society” to the minimum, but can enjoy the maximum happiness and the feeling of achievement. To achieve this, we need to pay attention to our “instinct” and utilize it to the maximum. In short, we should pay attention to the body sensations and make the most of our movement inside of us.

## References

1. Maslow AH (1943) A theory of human motivation. *Psychol Rev* 50(4):370–396
2. Deci EL, Ryan RM (1985) *Intrinsic motivation and self-determination in human behavior*. Plen, New York
3. Piaget JWF. [https://en.wikipedia.org/wiki/wiki/wikipedia.org/wiki/Jean\\_Piaget](https://en.wikipedia.org/wiki/wiki/wikipedia.org/wiki/Jean_Piaget)
4. Edwards B (1979) *Drawing on the right side of the brain*. Tarche, New York
5. Mahalanobis PC (1936) On the generalized distance in statistic. *Proc Natl Inst Sci India* 2(1):45–49
6. Sternberg RJ (1985) *Beyond IQ: a triarchic theory of human intelligence*. Cambridge Univerity Press, New York City
7. [https://en.wikipedia.org/wiki/My\\_Heart\\_Leaps\\_Up](https://en.wikipedia.org/wiki/My_Heart_Leaps_Up)

# Virtual Training System for the Autonomous Navigation of an Omnidirectional Traction Robot



De La Cruz Aida, Tapia Edison, and Víctor H. Andaluz

**Abstract** This paper presents the development of a virtual environment for the training of autonomous omnidirectional drive vehicle control. The virtual system considers the virtualization of structured and unstructured environments. Therefore, the virtual environment considers mathematical models of the omnidirectional robot in order to simulate more realistically the behavior and motion constraints of the robot. The integration of the control schemes is considered in the MATLAB software, for which a communication between the Unity3D graphic engine and MATLAB is considered through the use of DLL libraries. For the validation of the control algorithms on the virtual training environment, the construction of an omnidirectional traction vehicle with mecanum configuration. In addition, the constructed prototype will be used for the identification and validation of the mathematical models that represent its behavior. Finally, a usability analysis of the developed training system is presented.

**Keywords** Omnidirectional robot · Virtual training · Autonomous control · Dynamic modeling

## 1 Introduction

Currently, technological development has allowed advances in the area of robotics to be focused not only on the industrial area [1, 2]. The latest developments are oriented to non-industrial applications, for example, mining, agriculture, security,

---

D. La Cruz Aida (✉) · T. Edison · V. H. Andaluz  
Universidad de Las Fuerzas Armadas ESPE, Sangolquí, Ecuador  
e-mail: [alde2@espe.edu.ec](mailto:alde2@espe.edu.ec)

T. Edison  
e-mail: [eftapia2@espe.edu.ec](mailto:eftapia2@espe.edu.ec)

V. H. Andaluz  
e-mail: [vhandaluz1@espe.edu.ec](mailto:vhandaluz1@espe.edu.ec)

construction, health, among others. Among the most developed robots are land, aerial and aquatic mobile robots [3, 4]. Different applications that are usually performed by humans are nowadays developed by robotic platforms, e.g., domestic cleaning, crop spraying, traffic surveillance, among others [5, 6].

Considering the different mechanisms of terrestrial mobile robots, it can be described: (i) *unicycle robots*, have a mechanical structure of two wheels independently controlled by DC motors [7, 8]; (ii) *car-like robots*, are based on the Ackerman system with its linear velocity and angle of rotation [9, 10]; and (iii) *omnidirectional robots*, consisting of wheels with rollers that allow frontal, lateral and angular displacement [11, 12]. Due to their mobility, omnidirectional robots have different applications, e.g., inspection of hazardous environments, cargo transportation, among others.

For the above reasons, this work presents a virtual training system for the control and autonomous navigation of an omnidirectional traction robot [13]. It considers the digitization of laboratory and industrial environments, with the purpose of executing load transfer tasks through mobile robots [11, 14]. Mathematical modeling of an omnidirectional traction robot is determined, in order to be implemented in the virtual environment and in the development of control algorithms [11, 15]. The proposed dynamic model considers as input signals two linear velocities and an angular velocity; in addition, it considers the displacement of the center of mass, which is caused by placing a displaced load on the robotic platform. In order to implement different control strategies, MATLAB software is considered, which through the use of DLL libraries communicates in real time with Unity3D [16]. A cascade control scheme is proposed consisting of: kinematic controller and an adaptive dynamic compensation controller. To validate the mathematical models and evaluate the proposed control scheme, a mecanum-type four-wheeled omnidirectional robotic prototype was built. Finally, a usability test was implemented to different engineering users in order to evaluate the developed virtual training system.

## 2 Training System Methodology

Figure 1 shows the methodology developed for the digitalization of a virtual training system for the autonomous control of a mobile robot. The implemented methodology considers four main stages: construction of the mobile robot; mathematical modeling; digitization in the graphics engine; and design of advanced control algorithms.

- (i) *Construction*, this stage considers the design and mechanical and electrical construction of an omnidirectional traction robotic prototype. For the construction of the robotic system, a mecanum configuration of four wheels controlled by four independent motors coupled with encoder each motor is considered, in order to know the velocity and position of the robotic system. In addition, each motor has a driver that will allow controlling the velocity of the wheels, as

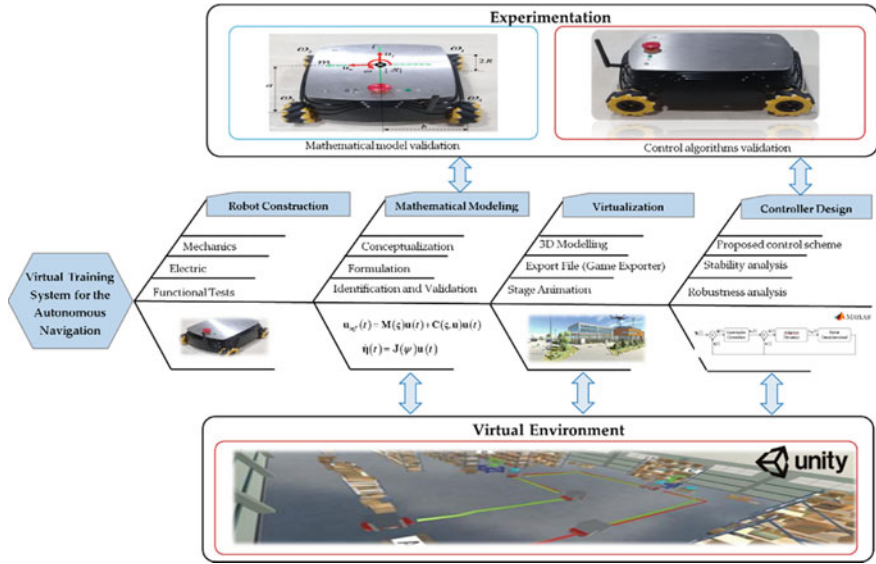


Fig. 1 Methodology for the virtual training system

- well as the direction of rotation, thus separating the control and power stages, respectively. A PID controller is implemented for each motor.
- (ii) *Mathematical Modeling*, through the heuristic method, the mathematical models of the omnidirectional robot is determined, in order to represent the behavior and motion restriction of the mobile robot. Mathematical models will be incorporated in the virtual environment and in the control scheme. The dynamic parameters of the mobile robot are determined experimentally.
  - (iii) *3D Digitalization*, the mobile robot and the virtual environment are modeled with CAD tools. Furthermore, elements are considered to simulate disturbances and different types of surfaces that affect the displacement of the omnidirectional robot. Then, using 3DS Max software, it is exported to Unity3D software.
  - (iv) *Control Scheme Design*, the main objective of the virtual environment is to be a theoretical-practical tool in the teaching–learning process, specifically in robotics for the design and evaluation of advanced control algorithms. In this work, a cascade adaptive control scheme. Finally, for the evaluation of the training system, simulation and experimental tests are considered: (a) *Simulation*, bilateral communication between MATLAB and Unity3D software through the use of DLL libraries; and (b) *Experimentation*, experimental tests are considered for the identification and validation of the mathematical models and to evaluate the considering control algorithms in this paper. For both simulation and experimental tests, a sampling period  $T_o = 100$  [ms] is considered.

### 3 Mobile Robot with Omnidirectional Traction

This item, describes the mathematical models of the platform with omnidirectional traction, considering a mecanum configuration. The mathematical models representing the behavior of the mobile platform are considered in the control scheme and in the 3D virtual environment.

#### 3.1 Kinematic Modeling

The mobile platform with omnidirectional traction considered in this work has a mecanum configuration, as shown in Fig. 2. Where  $\eta(t)$  defines the position and orientation of the control point with respect to  $\{R\}$ .

The kinematic model of the mobile platform with omnidirectional traction considered in this work is considering four velocities represented with respect to the moving reference system  $R(l, m, n)$ . The movement of the platform is defined by three linear velocities  $u_l$ ,  $u_m$  and  $u_n$  and an angular velocity  $\omega$  rotating about the vertical axis of the moving reference system  $R(l, m, n)$ .

Therefore, the motion of the mobile platform is defined as

$$\begin{cases} \dot{\eta}_x = u_f \cos(\psi) - u_l \sin(\psi) - \omega \eta_l \sin(\psi) - \omega \eta_m \cos(\psi) \\ \dot{\eta}_y = u_f \sin(\psi) + u_l \cos(\psi) + \omega \eta_l \cos(\psi) - \omega \eta_m \sin(\psi) \\ \dot{\eta}_\psi = \omega \end{cases} \quad (1)$$

Equation (1) can be described as

$$\dot{\eta}(t) = \mathbf{J}(\psi)\mathbf{u}(t) \quad (2)$$

where  $\dot{\eta}(t) \in R^n$  represents the velocity vector of the control point on  $\{R\}$ ;  $\mathbf{J}(\psi) \in R^{m \times n}$  with  $m = n = 4$  represents the motion characteristics of the mobile platform;

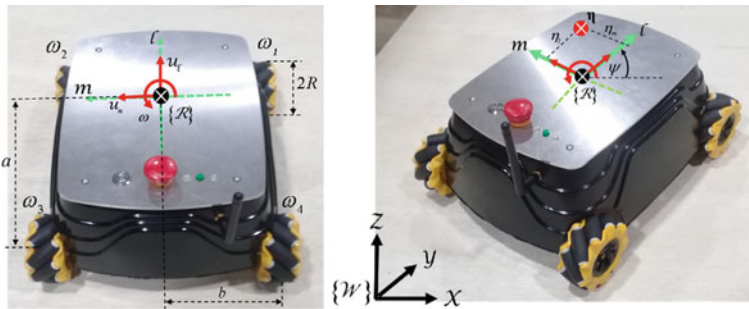


Fig. 2 Mobile robot with omnidirectional traction

and  $\mathbf{u}(t) \in \mathbb{R}^n$  is the maneuverability vector of the platform. The matrix  $\mathbf{J}(\psi)$  is of full rank with  $|\mathbf{J}(\psi)| = 1$ ; therefore, there exists the inverse matrix of  $\mathbf{J}(\psi)$ . The maneuverability vector  $\mathbf{u}(t)$  can be defined as

$$\mathbf{u}(t) = \mathbf{J}^{-1}(\psi) \dot{\eta}(t) \quad (3)$$

being  $\mathbf{J}^{-1}(\psi) = \frac{1}{|\mathbf{J}(\psi)|} \mathbf{J}^T(\psi)$  where  $\mathbf{J}^T(\psi)$  is the matrix transpose of  $\mathbf{J}(\psi)$ .

In addition,  $\mathbf{W} = [\omega_1 \ \omega_2 \ \omega_3 \ \omega_4]^T$  represents the angular velocity of each wheel, defined as

$$\mathbf{W}(t) = \mathbf{\Gamma}(a, b) \mathbf{u}(t). \quad (4)$$

Considering the mechanical characteristics of the  $\mathbf{\Gamma}$  platform it is defined as

$$\mathbf{\Gamma}(a, b) = \frac{1}{R} \begin{bmatrix} 1 & 1 & (a+b) \\ 1 & -1 & -(a+b) \\ 1 & 1 & -(a+b) \\ 1 & -1 & (a+b) \end{bmatrix}. \quad (5)$$

### 3.2 Dynamic Modeling

The dynamic of the mobile platform was developed based on the Euler–Lagrange formulation, for which it is considered that the power energy is equal to zero, since it has no displacement in the  $Z$  axis with respect to  $\{R\}$ . Therefore, the kinetic energy of the robot is defined as

$$L = E_C = \dot{\eta}^T \mathbf{M}_{R1} \dot{\eta} + \mathbf{W}^T \mathbf{I}_1 \mathbf{W}, \quad (6)$$

where  $M_{R1} = \frac{1}{2} \text{diag}\{m_R, m_R, I_R\}$  with  $m_R$  and  $I_R$  defined the mass and inertia of the robot, respectively. Also,  $I_1 = \frac{1}{2} \text{diag}\{I_W, I_W, I_W, I_W\}$  where  $I_W$  is the inertia of the wheels. Thus, the dynamic of the platform is defined as

$$\overline{\mathbf{M}} \ddot{\boldsymbol{\eta}} + \overline{\mathbf{C}} \dot{\boldsymbol{\eta}} = \mathbf{E}^T \boldsymbol{\tau}_i. \quad (7)$$

Now, considering that the moving platform is driven by DC motors, it is possible to define

$$\boldsymbol{\tau}_i = \frac{k_{pa}}{R_{pa}} (v_i - k_{pb} \mathbf{W}_i), \text{ with } i = 1, 2, 3, 4, \quad (8)$$

where  $v_i$  represents the input voltage to each motor;  $k_{pa}$ ,  $R_{pa}$ ,  $k_{pb}$  electrical constants of the motor. In addition, one PD controller per motor is considered

$$\mathbf{v}_v = \mathbf{K}_P(\mathbf{u}_{\text{ref}} - \mathbf{u}) - \mathbf{u}\mathbf{K}_D, \quad (9)$$

where  $\mathbf{K}_P > 0$  and  $\mathbf{K}_D > 0$  weigh control errors. Through (8)–(10) the mathematical model with velocity reference signals is obtained.

$$\begin{aligned} \begin{bmatrix} u_{\text{fref}} \\ u_{\text{lref}} \\ \omega_{\text{ref}} \end{bmatrix} &= \begin{bmatrix} \varsigma_1 & \varsigma_2 & \varsigma_3 \\ \varsigma_2 & \varsigma_4 & -\varsigma_5 \\ -\varsigma_6 & \varsigma_7 & \varsigma_8 \end{bmatrix} \begin{bmatrix} \dot{u}_f \\ \dot{u}_l \\ \dot{\omega} \end{bmatrix} \\ &+ \begin{bmatrix} \omega\varsigma_9 + \varsigma_{10} & -\omega\varsigma_{11} & -2\omega\varsigma_{12} \\ -\omega\varsigma_{13} & -\omega\varsigma_9 + \varsigma_{14} & -2\omega\varsigma_{15} \\ \omega\varsigma_{12} & \omega\varsigma_{15} & \varsigma_{16} \end{bmatrix} \begin{bmatrix} u_f \\ u_l \\ \omega \end{bmatrix}, \\ \mathbf{u}_{\text{ref}}(t) &= \mathbf{M}(\varsigma)\dot{\mathbf{u}}(t) + \mathbf{C}(\varsigma, \mathbf{u})\mathbf{u}(t), \end{aligned} \quad (10)$$

where  $\varsigma = [\varsigma_1 \ \varsigma_2 \ \dots \ \varsigma_j]^T \in R^j$  with  $j = 16$  represent the dynamic parameters of the mobile platform.

## 4 Control Scheme

The scheme for adaptive autonomous control for a robot with omnidirectional drive is presented in this item. The proposed control scheme for solving the motion control problem of a mobile platform is shown in Fig. 3. The control scheme is based on the mathematical models of the mobile platform:

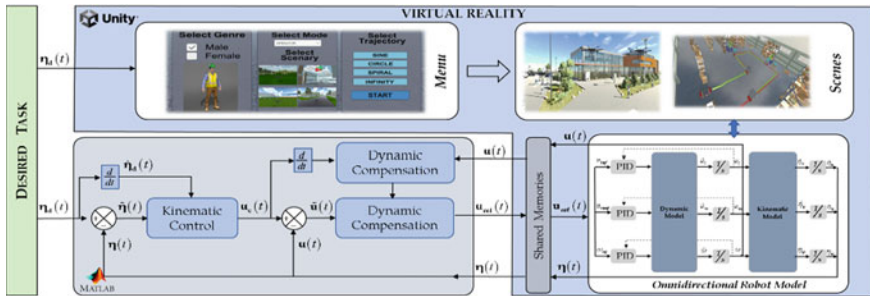


Fig. 3 Control scheme in the virtual training system

## 4.1 Kinematic Controller

The controller is based on the kinematic of the omnidirectional platform. Therefore, considering (3) we have

$$\mathbf{u}_c = \mathbf{J}^{-1}(\dot{\eta}_d + \Gamma \tanh(\alpha(\eta_d - \eta))), \quad (11)$$

where  $\mathbf{u}_c = [u_{lc} \ u_{mc} \ \omega_c] \in R^3$  is the calculated velocities;  $\mathbf{J}^{-1}$  is the inverse matrix;  $\Gamma > 0 \in R^{3 \times 3}$  weight control errors; and  $\alpha \in R^+$  that defines the saturation slope of the control errors.

## 4.2 Adaptive Dynamic Controller

The output of the dynamic controller in each sampling period considers the adaptation of dynamic parameters of the mobile platform. The dynamic parameters of the robot may vary as a function of the load carried by the robot or the surface on which it performs the task. Hence, the dynamic model (12) can be expressed as

$$\begin{aligned} \begin{bmatrix} u_{fref} \\ u_{lref} \\ \omega_{ref} \end{bmatrix} &= \underbrace{\begin{bmatrix} \varsigma_1 & \varsigma_2 & \varsigma_3 \\ \varsigma_2 & \varsigma_4 & -\varsigma_5 \\ -\varsigma_6 & \varsigma_7 & \varsigma_8 \end{bmatrix}}_{\mathbf{M}} \begin{bmatrix} \dot{u}_f \\ \dot{u}_l \\ \dot{\omega} \end{bmatrix} \\ &+ \underbrace{\begin{bmatrix} \omega u_f & u_f & -\omega u_l & -2\omega^2 & 0 & 0 & 0 & 0 \\ -\omega u_l & 0 & 0 & 0 & -\omega u_f & u_l & -2\omega^2 & 0 \\ 0 & 0 & 0 & \omega u_f & 0 & 0 & \omega u_l & \omega \end{bmatrix}}_{\chi} \begin{bmatrix} \varsigma_9 \\ \varsigma_{10} \\ \vdots \\ \varsigma_{16} \end{bmatrix} \\ \mathbf{u}_{ref}(t) &= \mathbf{M}(\varsigma) \dot{\mathbf{u}}(t) + \chi(\varsigma, \mathbf{u}), \end{aligned} \quad (12)$$

Based on dynamic model (13) we propose the following control law

$$\mathbf{u}_{ref} = \mathbf{M}\sigma + \chi, \quad (13)$$

where  $\sigma \in R^3$  defined as:  $\sigma = \dot{\mathbf{u}}_c + \kappa(\mathbf{u}_c - \mathbf{u})$  with  $\kappa = \text{diag}(\kappa_{ul}, \kappa_{um}, \kappa_{\omega}) > 0 \in R^{3 \times 3}$  weight velocities control errors; and the velocities control errors are defined as  $\tilde{\mathbf{u}} = \mathbf{u}_c - \mathbf{u}$ . Now, rewrite (14)

$$\mathbf{u}_{ref}(t) = \mathbf{\Omega}(\sigma, \mathbf{u}) \varsigma(t), \quad (14)$$

where



$$\mathbf{\Omega} = \begin{bmatrix} \dot{u}_f & \dot{u}_l & \dot{\omega} & 0 & 0 & 0 & 0 & 0 & \omega u_f & u_f & -\omega u_l & -2\omega^2 & 0 & 0 & 0 & 0 \\ 0 & \dot{u}_f & 0 & \dot{u}_l & -\dot{\omega} & 0 & 0 & 0 & -\omega u_l & 0 & 0 & 0 & -\omega u_f & u_l & -2\omega^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\dot{u}_f & \dot{u}_l & \dot{\omega} & 0 & 0 & 0 & \omega u_f & 0 & 0 & \omega u_l & \omega \end{bmatrix};$$

$\varsigma = [\varsigma_1 \varsigma_2 \cdots \varsigma_j]^T \in R^j$  with  $j = 16$ . In the case of any uncertainty in the unicycle robot parameters, the control law

$$\mathbf{u}_{\text{ref}} = \mathbf{\Omega} \hat{\varsigma} = \mathbf{\Omega} \varsigma + \mathbf{\Omega} \tilde{\varsigma} = \mathbf{M} \sigma + \chi + \mathbf{\Omega} \tilde{\varsigma}, \quad (15)$$

where  $\varsigma$  y  $\hat{\varsigma}$  are the actual dynamic parameters and estimated parameters of the robot; hence, the errors is defined as  $\tilde{\varsigma} = \hat{\varsigma} - \varsigma$ .

Similar to the work of Martins [17], stability analysis can be performed. Therefore, if  $\Gamma > 0$ , to ensure that  $\tilde{\eta}(t) \in R^3 \rightarrow 0$  when  $t \rightarrow \infty$ .

## 5 Analysis and Results

This item presents the most relevant results of the implemented system. This item considers the construction of a prototype; the digitalization environment; the implemented control scheme; and the results of the usability test.

### 5.1 Robot Omnidirectional

This work presents the construction of a robot with omnidirectional traction, developed in the Master's program in Electronics, mention Industrial Networks. Figure 4 shows the robotic prototype built.

**Fig. 4** Omnidirectional robotic prototype



From the electrical point of view, the omnidirectional robot is mainly composed of four main stages, see Fig. 5: (i) *Power system*, it consists of a LIPO battery, a bank of current protections, and a DC/DC converter module that allows powering all the devices that are part of the robotic platform, i.e., actuators, sensors, control system and communication modules; (ii) *Motor driver*, it is responsible for generating the voltage corresponding to the DC motors through H-Bridges, in addition, a PID controller is considered for each motor; (iii) *Actuators*, consists of four DC motors with encoder, each motor supporting a current of 2A; (iv) *Control system*, is composed of a control unit where the closed-loop control algorithms are implemented; and finally (v) *Communication*, manages the wireless communication between the robot and an external computer.

On the other hand, the identification of the dynamic parameters that represents the behavior of the omnidirectional robot was performed with the robotic prototype built. The method used for the identification is through general identification, or black box, where the objective is to establish the input–output relationship of the system, without making physical interpretations on the composition of the mathematical model, see Fig. 6a.

The identification of the dynamic parameters of the mobile platform, was developed based on the minimization of an objective function:  $\min f(\zeta)$  where  $\zeta \in \mathbb{R}^{16}$  are the parameters of the system to be identified. The identification of the parameters is carried out through the optimization algorithm, when the value of  $f(\zeta)$  reaches a minimum the search delivers the vector of parameters  $\zeta$  that satisfy the minimization of the function. Figure 6b presents the validation results of the dynamic model of the mobile robot.

The kinematic and dynamic models obtained from the constructed prototype are implemented in the digitalization environment. Furthermore, perturbations in the robot output are considered according to the type of surface (Fig. 7).

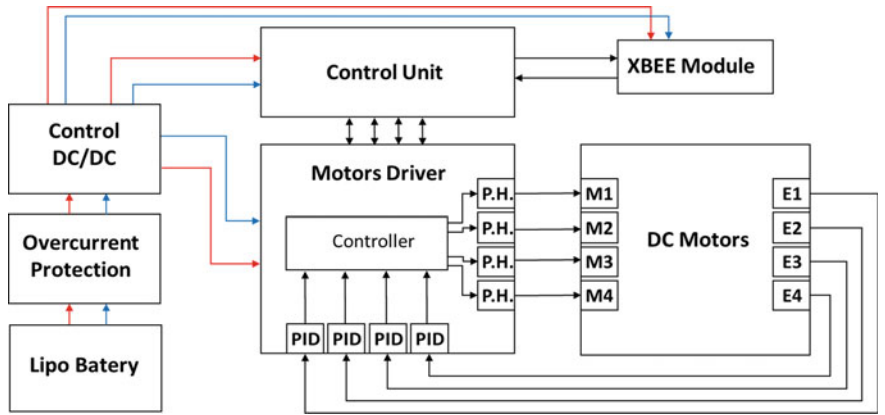


Fig. 5 Omnidirectional robot electrical diagram

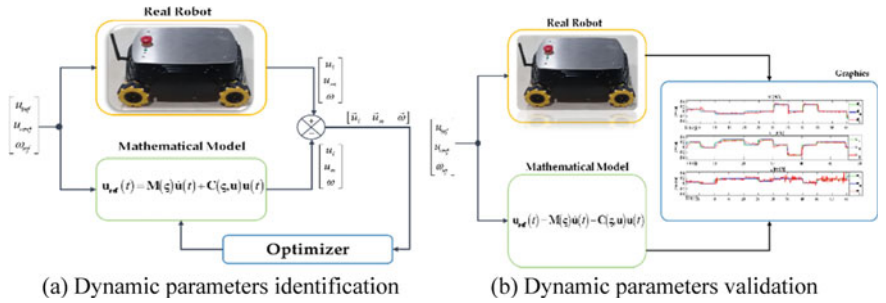
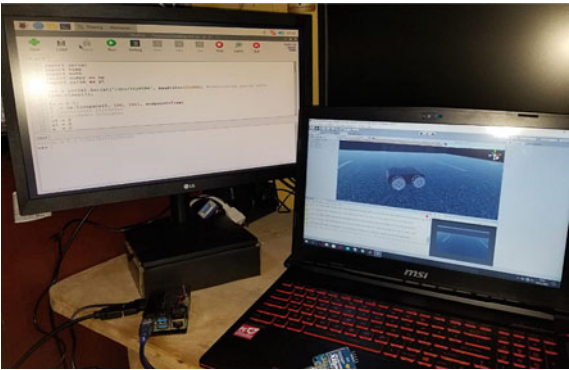


Fig. 6 Dynamic parameters

Fig. 7 Virtual training system—Unity3D graphics engine



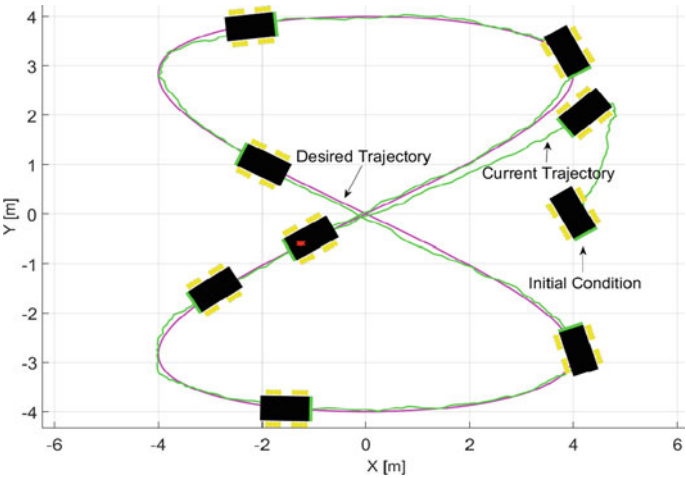
5.2 Control Scheme Implementation

Relevant results are presented in this subsection. Each test was run both in the virtual environment and experimentally. Figure 8 presents the stroboscopic movement of the omnidirectional traction robot based on experimental data. Figure 9 presents the control errors evolution  $\tilde{\eta} \in R^3$ , which are  $|\tilde{\eta}(t)| < 0, 2$  [m]. Finally, Fig. 10 presents the control actions for the omnidirectional robot.

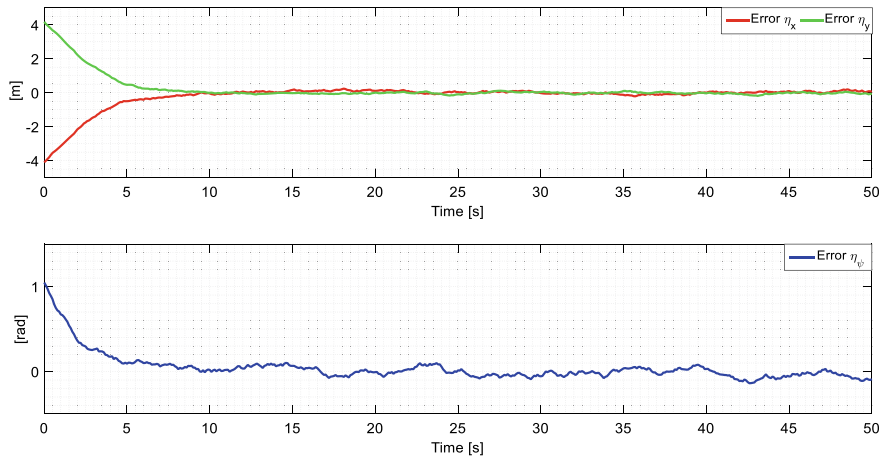
Finally, through the usability test (SUS [17, 18]) a weighting of 78.5% was obtained from a group of fifteen students, so it can be concluded that the training system developed is good.

6 Analysis and Conclusions

This work presented the development of a digitalization system for the autonomous control of omnidirectional traction robots. The system considers a digitized environment in Unity3D, in which the mathematical models of the omnidirectional robot are

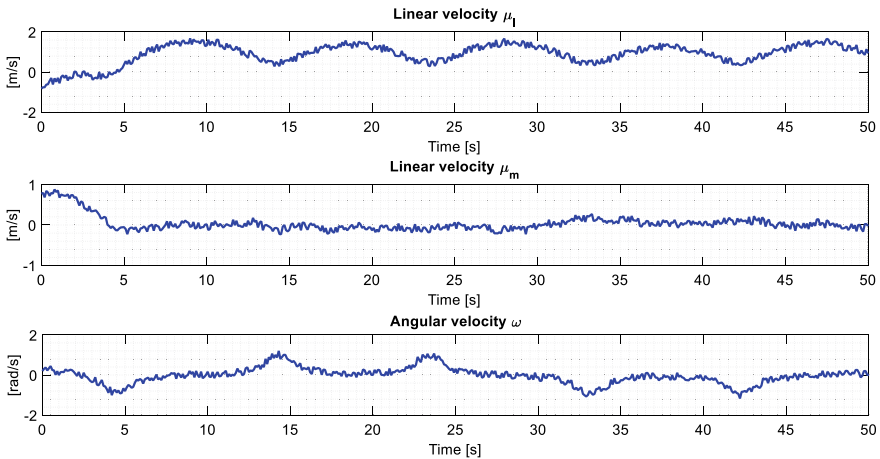


**Fig. 8** Stroboscopic motion of the mobile platform based on experimental data



**Fig. 9** Control errors evolution

implemented. The mathematical models obtained were validated with a built robotic prototype, which consisted of four mecanum-type wheels. In order to evaluate the training system developed, a cascade control scheme was implemented. Finally, from the usability tests performed on students, a SUS percentage of 78.5% was obtained, which is within the acceptable margin.



**Fig. 10** Velocity commands of the mobile robot

**Acknowledgements** This article shows the results of the Master’s Degree Program in “Master’s Degree in Electronics and Automation with mention in Industrial Networks” of the Postgraduate Centre of the University of the Armed Forces-ESPE. Finally, the authors would like to thank the ARSI research group for their advice in the development of the degree work.

## References

1. Sánchez H, Martínez LS, González JD (2019) Educational robotics as a teaching tool in higher education institutions: a bibliographical analysis. *J Phys Conf Ser* 1391:012128. <https://doi.org/10.1088/1742-6596/1391/1/012128>
2. Ruspini E (2019) The robot and us: an “antidisciplinary” perspective on the scientific and social impacts of robotics. *J Tourism Futures* 5:297–298. <https://doi.org/10.1108/JTF-09-2019-089>
3. Sharma A, Sinha A (2021) Approach of automation manufacturing and optimization in a robotic industry. *Int J Robot Autom* 7:18–35
4. Andaluz V, Varela Aldás J, Chicaiza F, Quevedo W, Ruales Martínez M (2019) Teleoperation of a mobile manipulator with feedback forces for evasion of obstacles. *RISTI—Revista Iberica de Sistemas e Tecnologias de Informacao* 2019:291–304
5. Pal S, Gupta S, Das N, Ghosh K (2022) Evolution of simultaneous localization and mapping framework for autonomous robotics—a comprehensive review. *J Autonom Vehicles Syst* 2. <https://doi.org/10.1115/1.4055161>
6. Carvajal CP, Proaño L, Pérez JA, Pérez S, Ortiz JS, Andaluz VH (2017) Robotic applications in virtual environments for children with autism. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. 10325 LNCS, 175–187
7. Zeini M, Pirmoradian M (2021) Design and construction of a unicycle robot controlled by its center of gravity. *J Simul Anal Novel Technol Mech Eng* 13:59–73
8. Sun Z, Xia Y, Dai L, Campoy P (2020) Tracking of unicycle robots using event-based mpc with adaptive prediction horizon. *IEEE/ASME Trans Mechatron* 25:739–749. <https://doi.org/10.1109/TMECH.2019.2962099>

9. Chen X, Huang Z, Sun Y, Zhong Y, Gu R, Bai L (2022) Online on-road motion planning based on hybrid potential field model for car-like robot. *J Intell Robot Syst* 105:7. <https://doi.org/10.1007/s10846-022-01620-5>
10. Chen H, Yang H, Wang X, Zhang T (2018) Formation control for car-like mobile robots using front-wheel driving and steering. *Int J Adv Rob Syst* 15:1729881418778228. <https://doi.org/10.1177/1729881418778228>
11. Ortiz JS, Molina MF, Andaluz VH, Varela J, Morales V (2018) Coordinated control of a omnidirectional double mobile manipulator. In: Kim KJ, Kim H, Baek N (eds) *Proceedings of the IT convergence and security 2017*. Springer, Singapore, 278–286
12. Yunardi RT, Arifianto D, Bachtiar F, Prananingrum JI (2021) Holonomic implementation of three wheels omnidirectional mobile robot using dc motors. *J Robot Control (JRC)* 2:65–71. <https://doi.org/10.18196/jrc.2254>
13. Saenz A, Santibañez V, Bugarin E, Dzul A, Ríos H, Villalobos-Chin J (2021) Velocity control of an omnidirectional wheeled mobile robot using computed voltage control with visual feedback: experimental results. *Int J Control Autom Syst* 19:1089–1102. <https://doi.org/10.1007/s12555-019-1057-6>
14. Andaluz VH, Pérez JA, Carvajal CP, Ortiz JS (2019) Virtual environment for teaching and learning robotics applied to industrial processes. In: De Paolis LT, Bourdot P (eds) *Proceedings of the augmented reality, virtual reality, and computer graphics*. Springer International Publishing, Cham, 442–455
15. Wu M, Dai S-L, Yang C (2020) Mixed reality enhanced user interactive path planning for omnidirectional mobile robot. *Appl Sci* 10:1135. <https://doi.org/10.3390/app10031135>
16. Andaluz VH, Chicaiza FA, Gallardo C, Quevedo WX, Varela J, Sánchez JS, Arteaga O (2016) Unity3D-matlab simulator in real time for robotics applications. In: *Augmented reality, virtual reality, and computer graphics*; De Paolis LT, Mongelli A (eds) *Lecture notes in computer science*, vol 9768. Springer International Publishing, Cham, 246–263 ISBN 978–3–319–40620–6
17. Martins F, Celeste W, Carelli R, Sarcinelli-Filho M, Bastos-Filho T (2008) An adaptive dynamic controller for autonomous mobile robot trajectory tracking. *Control Eng Pract* 16(11):1354–1363
18. Salvendy G (2012) *Handbook of human factors and ergonomics*. Wiley. ISBN 978–1–118–12908–1

# NFTs: Inside the Twitter Discussion



Victor Hernández-Manrique, Rodrigo Carmona-Herrera,  
Francisco J. Cantú-Ortiz, and Héctor G. Ceballos-Cancino

**Abstract** NFTs, which stand for *non-fungible tokens*, have become a technology with ups and downs. Through the years, the idea of what it could mean for the society, its advantages, and the future of the tokens, has changed, and nowadays, the users are expressing their opinions and experiences about this topic. Twitter is the platform where people around the world can discuss any subject with 280 characters per tweet. The following study provides a sentiment analysis of the tweets related to this matter in order to understand the public view toward them. The purpose behind this study is to provide an insight on how the consumers feel with this emerging technology, allowing companies to decide if an approach to them through this is viable to enhance the interaction.

**Keywords** NFTs · Twitter · Sentiment analysis · Topic analysis · Public perception

## 1 Introduction

NFTs are unique digital assets created by using blockchain technology. *Blockchain technology* was conceived as the ultimate collective tool for data protection. The reason is that it uses an encryption technique that allows every user on the blockchain network to verify and approve a change, making it virtually impossible to be falsified. This is the technology that was used to create NFTs. They have a wide variety of

---

V. Hernández-Manrique (✉) · R. Carmona-Herrera · F. J. Cantú-Ortiz · H. G. Ceballos-Cancino  
Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, NL, Mexico  
e-mail: [A01731594@tec.mx](mailto:A01731594@tec.mx)

R. Carmona-Herrera  
e-mail: [A01731508@tec.mx](mailto:A01731508@tec.mx)

F. J. Cantú-Ortiz  
e-mail: [fcantu@tec.mx](mailto:fcantu@tec.mx)

H. G. Ceballos-Cancino  
e-mail: [ceballos@tec.mx](mailto:ceballos@tec.mx)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_32](https://doi.org/10.1007/978-981-99-3091-3_32)

applications, even though the most common one is about using it for generating digital art so it has potential at long-term. An example for this is that it can be used as a authentication method since it has an unique signature [1]. Or its use for marketers by persuading costumers into buying them thanks to its characteristics [2]. Times are undeniably changing, and topics such as the metaverse and NFTs are gaining terrain among the society. This has a great implication in all levels for institutions, artists, buyers, and investors because they are the ones that are more affected by this technology [3]. However, even with all the noise that has been generated through the years about NFTs, there is still so much disinformation about it. Many potential users are not costumers simply because they do not know how this trade technology works or how to begin using it. The purpose of this research is to analyze user thoughts, ideas, and feeling toward NFTs in order to determine their future impact. From a professional point of view, it was decided to use Twitter as a source of information due to the extreme honesty from its users.

## 2 Background

### 2.1 Understanding the Revolution

The main idea behind the revolution of NFTs is the ownership of a digital matter, such as a picture and video. The buyer has the legal rights to sell it, but this does not exclude other users to download or obtain for personal or business use the same art. On the other hand, the artist/creator receives a royalty every time his/her piece is sold. From its beginning, several ideas of what it could become attract attention for those who were not related to this subject. For example, this was a new way for creators to obtain profits from their work after it was sold. For the buyers and users, having a one-of-a-kind artwork could result in an increase of the original price. In early 2021, an art piece was sold for \$69M USD [5]. The cryptocurrency is another topic related to the NFTs. The volatility of this market positioned several pieces like the *bored ape yacht collection* as prime artwork. Celebrities, such as Eminem or Jimmy Fallon, bought them and their prices went through the roof [6]. Months later, that same volatility drastically reduces the cost of expensive NFT to just a couple of dollars.

## 3 Methodology

Even though a dataset related to this topic exists, we are looking to analyze the latest reactions since the NFTs technology has been evolving, and the experience is different from one user to another. The creation of the new dataset is possible through *Twitter API* [4]. As the platform mentions in its Website:



The Twitter API is a set of programmatic endpoints that can be used to understand or build the conversation on Twitter. This API allows you to find and retrieve, engage with, or create a variety of different resources.

All the content related to the tweet can be analyzed, such as the text, the region where it comes from, the date of creation, and furthermore the information of the user itself. The details of the tweet creators will not be taken into account for privacy concerns. The examination of the subjectivity and polarity of each tweet will concede a sentiment approach through diverse functions.

## 4 Data

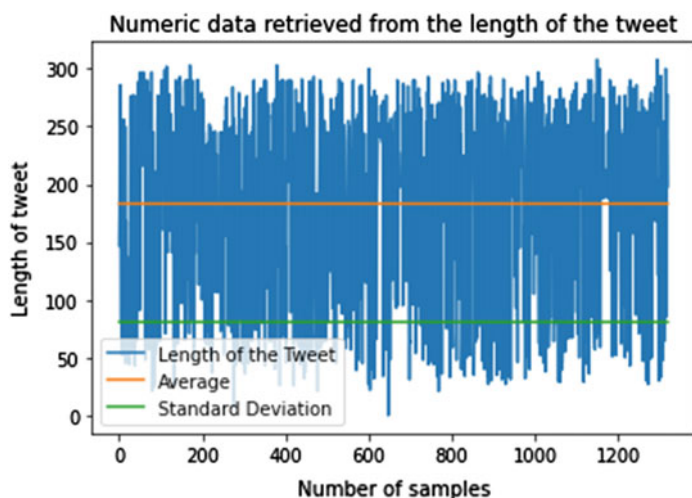
In order to prepare the analysis of data, some libraries are required. Since this is a sentiment analysis using Twitter application programming interface (Twitter API), the library *Tweepy* was used. Other libraries were incorporated for necessary features. We used *Pandas* for data manipulation, *Textblob* for the processing of textual data and *Wordcloud* to generate an image with the most repeated words on the tweets. Additionally, some more classical and must be libraries such as *Numpy*, *re*, and *Mathplot* were also added.

The next step was to actually extract the tweets related to the NFT opinion/thoughts. To obtain a larger sample, we decided to include tweets made in English and Spanish. The reason for this decision was made based on the fact that these communities are the largest consumers of NFTs, creating the biggest marketplace to sell and buy these items [9]. That is the basis behind the keywords language for this research. Then, a filter for retweets was added and finally, and we limited the search to one hundred thousand (100,000) tweets. With all this information on hand, the data frame was also built. All the tweets were cleansed from special characters, links, hashtags, and mentions that could have been written in any tweet. This occurs because those characters could potentially affect the sentiment analysis.

## 5 Results

### 5.1 Information Obtained from Tweets

Even though we set a limit of one hundred thousand tweets, our program was only able to obtain thirteen hundred records. We attributed this to the Google Collab and Twitter API limitations, since the latest only cover tweets up to seven days ago. From all the records, we acquired the attributes time, language, source, tweet, length of the tweet, subjectivity, and polarity. The *sentiment analysis* was based on the last variables mentioned before. These attributes were chosen because we considered them as sufficient for relevant information for our analysis. The way the program



**Fig. 1** Tweet's length

was written allowed us to extract always the latest tweets; ideally, it should always be the last 100 thousand, but because of the limitations of Google Collab, it was reduced at the end.

It is worth mentioning the analysis of the minimum and maximum length of the tweets. The minor one registered had a length of zero characters which its suprising because it should not even return any value in the search of the tweets. On the other hand, the greater value of characters was three hundred and eight (308), which at the beginning was incoherent because the maximum of characters allowed in Twitter per Tweet is 280. But then, it all fit when we investigated that, even though emojis consume only one character space, at the count, they take a value of two characters. So, at the end, we attributed the largest tweet length to a spread use of emojis.

Figure 1 shows the length of every tweet along the **average** size and the **standard deviation**. That information tells us how much people got to tell about the subject; in other words, it allow us to measure how much interest people have about NFTs.

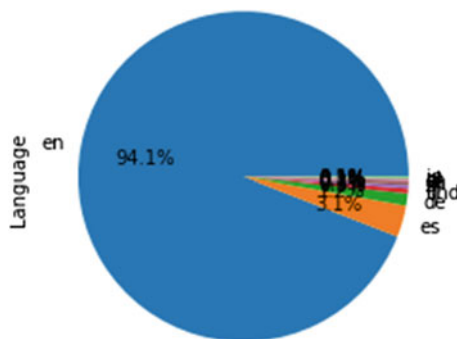
The average character length of NFT sentiment tweets is 183, compared with the average tweet length of 28 characters [8], it shows that the user that expresses an opinion or thought related to this topic has to provide more information than usual in order to communicate his/her message.

Next, we analyzed the most repeated words in all those tweets. The work was done using the Word Cloud library, which helped us with the visualization of the words repeated with higher frequency by remarking them with bright colors and inserting them in a collage. As imagined, it works by counting how many times certain words were repeating and then counting how many times were repeated in all samples. The ones with the highest mentions get a higher font size and brighter color. Figure 2 shows that, besides 'opinion' and 'NFT', the users relationed the



Fig. 2 Word cloud with the most repeated words

Fig. 3 Language of tweets



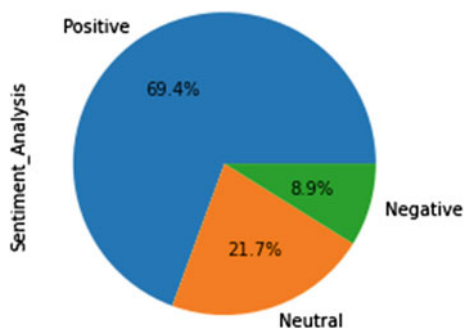
subject with the ‘cryptocurrency’ as well as words such as ‘right now’ or ‘hottest’, which lead to understand that this technology is worth discussing on these days. As it was mentioned in Sect. 3, we made a general search of tweets written in English or Spanish. With that information on hand, it was plotted a graph containing all the languages found during that search. The analysis of this particular aspect was done as a control method to have an idea on which language community the NFTs topic has the most impact. The diverse dialects found on this research are shown in the Fig. 3.

It was also obtained the source of the tweets. This refers to what publishing method was used to post the tweets. The results are presented in the Fig. 4.

## 5.2 Sentiment Analysis

In order to measure the sentiment of the tweets, two functions from the *TextBlob* library were used. The first one, called *subjectivity*, analyzes the text and returns a



**Fig. 6** Sentiment analysis

tion. All positive values obtained a *positive* classification. And only values of 0 can get a *neutral* classification. This method allowed us to graph the sentiment with a great ease. The result of this classification is shown in the Fig. 6.

## 6 Discussion and Related Work

The main objective of this research was to present a sentiment analysis related to NFTs in order to determine whether the industry should invest in this subject to enhance the interaction between sellers and consumers. The results obtained offered an important insight on which media do the users spend their time commenting this topic or which language do they speak, allowing to analyze which community is greater or where they must focused in the long run. Even though the recent news showed that the value of NFTs has dropped over the months [10], the majority of the public has a positive opinion on this technology, statement supported by the Fig. 6. This might be due to the fact that more security features and government stipulations supported them. Unquestionably, the english community has a strong presence on this theme since they are the main consumers of this issue, as it was seen in the Fig. 3. It is interesting to see that the majority of tweets with the word 'NFT' and its alternatives are published from an iPhone (Fig. 4). A common conception nowadays is that Apple's users have more valuable purchase and the acquisition of NFTs support this case, since those are not cheap. Related work presented a similar approach to our investigation, since the examination of subjectivity and polarity is the base of a *sentiment analysis* [11]. Another technique, such as sentiment scores, could be applied to this type of research. Its wide range of results (from zero to ten) may present more accurate results [12].

## 7 Conclusion and Future Work

The presented article provides a sentiment analysis to the Twitter community regarding the subject ‘NFTs’. The data inspection is valuable for companies since their specialize ads invite users to spend on them. At this moment, the general sentiment is *positive*, concluding that, even though the current state is negative and its value is falling, clients are still buying and selling NFTs. Future work is related to the analysis of the processes behind these products, which will let us understand how the users feel about the purchase procedure to buy them, for instance. Other sentiment analysis might be performed to understand specific aspects of the tweets, such as the correlation between more variables, but the main core of this labor is to deconstruct the text and examine its subjectivity and polarity, as it was made on this research.

**Acknowledgements** We would like to thank CONACyT and Instituto Tecnológico y de Estudios Superiores de Monterrey campus Monterrey for their support toward this investigation.

## References

1. Khokhariya U, Shah K, Pancholi N, Kumar S (2023) DAMBNFT: document authentication model through blockchain and non-fungible tokens. In: Part of the lecture notes in networks and systems book series LNNS, vol 396. [www.scopus.com](http://www.scopus.com)
2. Chohan R, Paschen J (2022) How marketers can use non-fungible tokens (NFTs) in their campaigns. Business Horizons
3. Belk R, Humayun M, Brouard M (2022) Money, possessions, and ownership in the metaverse: NFTs, cryptocurrencies, Web3 and wild markets. J Bus Res 153:198–205
4. TwitterDev (2022) Get started with the Twitter developer platform. Twitter. <https://developer.twitter.com/en/docs/platform-overview>. Last accessed on 27 Sept 2022
5. Park A, Kietzmann J, Pitt L, Dabirian A (2022) The evolution of nonfungible tokens: complexity and novelty of NFT use-cases. IT Prof 24(1):9–14
6. van Slooten J (2022) Predictive value of tweet sentiment and volume on the Bored Ape Yacht Club’s trading volume and price
7. Kharif O (2022) NFTs have ‘Fallen off the cliff’ as sales Sink to lowest levels in a year. Bloomberg. <https://www.bloomberg.com/news/articles/2022-06-29/nfts-have-fallen-off-the-cliff-as-sales-sink-to-lowest-in-year>. Last accessed on 27 Sept 2022
8. SMK Editor (2022) The average Tweet length is 28 characters long, and other interesting facts. [SMK] Social Media Knowledge. <https://smk.co/the-average-tweet-length-is-28-characters-long-and-other-interesting-facts/>. Last accessed on 23 Oct 2022
9. White B, Mahanti A, Passi K (2022) Characterizing the OpenSea NFT marketplace. In: Companion proceedings of the web conference 2022, pp 488–496
10. Howcroft E (2022) NFT sales plunge in Q3, down by 60% from Q2. Reuters. <https://www.reuters.com/technology/nft-sales-plunge-q3-down-by-60-q2-2022-10-03/>. Last accessed on 29 Oct 2022
11. Alsaeedi A, Khan MZ (2019) A study on sentiment analysis techniques of Twitter data. Int J Adv Comput Sci Appl 10(2)
12. Ahuja S, Dubey G (2017) Clustering and sentiment analysis on Twitter data. In: 2017 2nd international conference on telecommunication and networks (TEL-NET)

# Integrating Analog PIR Sensor Telemetry with TinyML Inference for On-The-Edge Classification of Moving Objects



Ritha M. Umutoni , Marvin Ogore , Damien Hanyurwimfura ,  
and Jimmy Nsenga 

**Abstract** Identification of moving objects plays an important role in different real-time applications such as security monitoring, social distancing for infection disease spreading surveillance, and so on. Image-based sensor technology has been the go-to approach at the expense of huge processing time, resources and complexity. As an alternative, digital passive infrared (PIR) is used to overcome the above challenges of image-based sensors but fails to address the challenge of moving object identification due to business logic being hard-coded in hardware logic. This research takes a different approach by using analog PIR signals as the data source and leveraging the emerging tiny machine learning (TinyML) technology to identify moving objects with limited real-time resources. The lack of an open-source public dataset has driven this research to start by collecting primary analog PIR data of moving humans, dogs, goats, and windblown vegetables. Then, a TinyML classification model has been trained and deployed on a resource-constrained embedded microprocessor for real-time classification inference. The pilot experimentation shows a performance accuracy of 80.8% which may be improved over time using reinforcement learning. The proposed integration of analog signals and TinyML can be extended and applied in monitoring infectious diseases between humans, and between humans and animals.

**Keywords** Analog PIR · Deep learning · TinyML

## 1 Introduction

The integration of the Internet of Things (IoT) and machine learning (ML) with inference at the edge enables the development of real-time applications that require 24 h operations (always on); namely monitoring of moving objects in the surrounding environment, management of social distance with wearable devices, home intrusion

---

R. M. Umutoni (✉) · M. Ogore · D. Hanyurwimfura · J. Nsenga  
African Center of Excellence in Internet of Things (ACEIoT), University of Rwanda, Kigali,  
Rwanda  
e-mail: [rithamarie9@gmail.com](mailto:rithamarie9@gmail.com)

surveillance, proximity detection in dangerous or unauthorized locations, monitoring construction sites, and so on [1, 2]. The use of image processing through cameras has been the standard go-to sensing technology for such types of applications. However, image processing implicates a heavy cost of resources, cost, and time [3] limiting its use in many real-time applications. Indeed, image processing requires high bandwidth up to hundreds of megabytes for high-quality images [4], energy consumption for real-time monitoring, and lack of privacy (sensing information that is not necessary to the use case such as image quality). These challenges motivate the need for a low-cost and lightweight solution to enable different existing and emerging privacy-sensitive and real-time resource-constrained use cases.

There have been various innovations that address these challenges. For instance, researchers propose to detect moving objects by combining a digital passive infrared (PIR) sensor and camera, with the PIR first detecting the movement of the object and then the camera recording the movement of objects. Shaikh et al. [5] created an optical camera combined with a digital PIR sensing device for external intruder detection and classification. In [6], authors developed a system that uses graph spectral clustering to detect moving objects targeting applications like video surveillance, security, enforcement, and self-driving cars. In [7], a sensor tower platform which is made of 8 PIR sensors is combined with an optical camera to classify and detect any intrusion in the outside environment. All the above solutions rely on a digital PIR for which the object detection logic is implemented in hardware to provide a binary response about whether a movement has been detected or not. However, the raw analog signal waves collected by PIR sensors feature a lot of patterns that may be used to classify moving objects.

Analog PIR has been used in various applications. For instance, in [3] an activity recognition method was developed based on analog output and PIR sensors that can track the user's precise movements and detect which activities are being performed using ML. In [4], a system with PIR sensors that can calculate resting heart rate (RHR), as an effective and affordable solution for heart rate monitoring is the analog PIR-based system. Additionally, a system based on a single PIR sensor for monitoring purposes was developed in [8] using also a PIR sensor integrated with ML to monitor human-related activities. In these systems, they showed how a single PIR sensor can be used in person monitoring, person counting, activity recognition, motion monitoring, and even for security purposes. In [9], a combination of low-cost PIR and Deep Learning (DL) has been experimented with to optimize the performance of human counting and localization for both one and multiple persons. In this research, a two-stage network has been designed, one using signal separation for counting persons and another being used for determining their locations. To the best of our knowledge, the application of analog PIR with ML to distinguish the types of moving objects has not been explored yet.

Therefore, this paper aims at designing and prototyping a cost-effective and evolutive smart embedded device that senses the analog PIR signal patterns of moving objects and then uses Deep Learning to make the moving object classification. To enable real-time operation, the training of collected data is done using the TinyML technology to enable deployment of the resulting model on real-time



resource-constrained analog PIR device, enabling therefore to infer the classification directly at the edge; thus as a way to minimize wireless communication with the cloud, decreases energy consumption (i.e., increases battery life) and also reduce the decision latency feedback since the processing is done directly on-the-edge device.

The rest of the paper is organized as follows; Sect. 2 defines the requirements for edge classification of moving objects. Section 3 presents the design and prototyping of a TinyML-based analog PIR device for classification inference of moving objects at the edge, the performance analysis, and inference accuracy. Finally, future works and conclusions are presented in Sect. 4.

## 2 Requirements for Edge Classification of Moving Objects

Classification of moving objects is a requirement for several real-life applications, especially the ones related to security. Before using PIR-based sensors, camera-based solutions were the dominant solutions but encountered several limitations such as privacy invasion, high bandwidth, and high energy consumption [8, 9]. Later on, a PIR-based solution emerged to overcome some of the problems of a camera-based solution as summarized in the Table 1.

PIR sensors rely on infrared radiation change to detect humans across their field of view. Thanks to their non-invasive and non-contact features, various indoor and outdoor applications have been using PIR sensors to perform different tasks such as monitoring different activities, human identification, activity recognition with energy harvesting, and movement directions [10, 11].

PIR sensors are of two categories depending on the output produced. (1) analog PIR sensors which produce analog output, (2) digital PIR sensors which produce the digital output [12]. On the one hand, digital PIR sensors have binary output; “1” if an object is detected or “0” if no object is detected. This binary outcome limits further analysis of the output from the sensor. On the other hand, analog PIR sensors

**Table 1** Comparison of object detection between the camera and analog PIR

Object detection using the camera	Object detection using analog PIR
Intrusion of privacy	No intrusion of privacy
High power consumption	Low power consumption
Low resolution caused by low light condition	Doesn't affected by light condition
Expensive hardware	Cheap hardware
Long setup time in case of deploying	Short setup time for deploying
Camera capture many things in the environment	Analog PIR detects the targeted objects

produce analog voltage output which can be processed with various machine learning algorithms to extract patterns and classify them [13].

Analog PIR data cannot differentiate different moving objects because it's almost the same by checking the numbers. In short, simple if else statements in code to interpret object detection and classification using PIR data. That is why the ML learning process of feature extraction and pattern recognition was needed.

Integrating analog PIR with ML allows code signing a hardware/software system optimized in terms of bandwidth, latency, real-time performance, cost, and power consumption [14]. The enhanced development of hardware architectures and digital signal processing (DSP) capabilities has made it possible to build these systems while maintaining low costs and power consumption [15]. TinyML processes raw analog PIR time-series data to identify hidden features for human analysts, which allows distinguishing between moving objects, as opposed to digital PIR, which uses a kind of rule-based configuration in hardware (HW).

Analog PIR time-series inference is very important to consider latency since it is applicable in sensitive applications like security which concerns human life. On-device ML mostly called edge computing overcomes these challenges by using resource-constraint embedded devices by optimizing and compressing the Deep Learning model (DL), low memory, cheap hardware devices, low computation time, and low power. Various DL models for edge computing are utilized in a variety of industries and sectors, including smart transportation, smart buildings, smart health, and smart cities. [16]. Outdoor sound event detection is one of the on-time-series DL based at the edge [17]. In-door human monitoring, counting, and localization using PIR and ML [7–9].

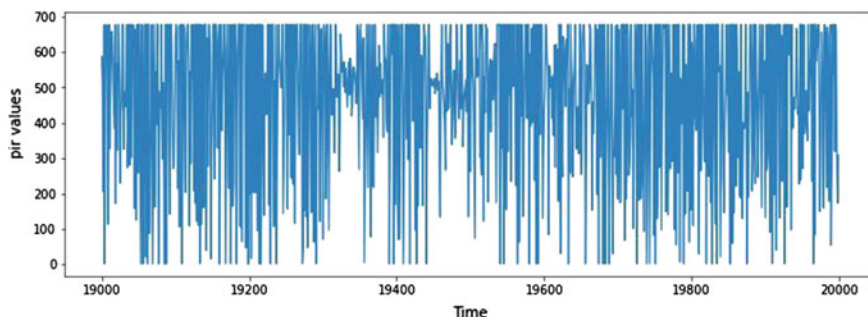
### **3 Design and Prototyping of a TinyML-Based Analog PIR Device for Classifying Moving Objects at the Edge**

#### ***3.1 Data Collection Campaign***

As described in Sect. 2, Raw analog PIR time-series data requires further processing and manipulation through data visualization and data manipulation an integrated with ML to find features that are hidden and allowed to detect the different moving objects. The biggest challenge presented finding open-source datasets in public repositories. Therefore, primary data was collected using an IoT device (see Fig. 1) made up of an AMN24112 analog PIR sensor working at 166 Hz, RTC, SD card, Arduino Uno boards, switch buttons, and LED. The device was used to collect human-related data, animals (dogs, goats), windblown vegetation, and analog readings where there was no object detected. The data collection process took three months. Since the quality of the data was a big concern, the data was collected in a different scenario but the ones more accurate to the use case were chosen.



**Fig. 1** Device developed for collecting analog PIR time-series dataset



**Fig. 2** Output raw data from a class person made of 1000 samples

During the data collection process, it has been found that the strength of the signal from analog PIR sensors can change with environmental contexts. For example, when the temperature changes the range of signals also changes.

In addition, the data collected from highly populated areas like churches, have a different signal range from quiet areas like homes or schools.

In addition to the lesson learned, there were various challenges including (1) the availability of targeted objects for example presence of wind, (2) forcing animals to move, and (3) the presence of noise (movement of the not targeted object). After overcoming some of the challenges, a dataset was formed with 213,760 rows collected in each of the four classes. In each class, 1287 observations were produced, totaling 5148 observations for all 4 classes. See in Fig. 2 how 1000 samples from a class person. In Fig. 3, Raw data from an animal class was shown in one observation of 166 rows, as number 110 in the class.

### 3.2 Machine Learning Training

The moving object classification model was trained using the following steps:

1. After data collection, data were pre-processed by cleaning through the removal of null values

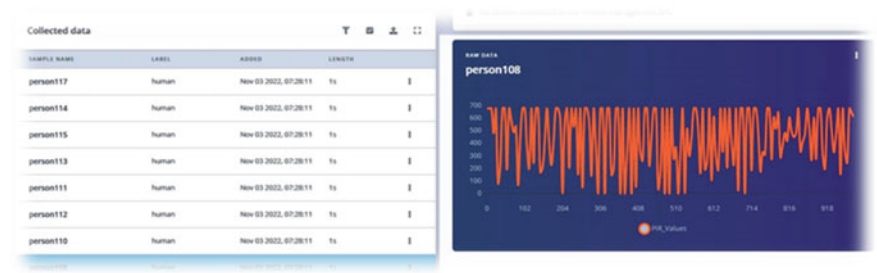


Fig. 3 Raw data from animal class, observation number 110

- 2. Data transformation—Data was converted from raw values to a time-series dataset. Fast Fourier transform as a further step is used to decompose the data into frequency components.
- 3. Dataset is then split into 80% for training/validation and 20% for testing.

Digital Signal Processing (DSP) is used to implement mathematical approaches and statistical analysis to visualize and manipulate time-series data acquired from our analog signal for analysis. through extensive experimentation with different settings, We can see a clear distinction between features extracted from respective classes this, in turn, helped The features that produced the beast features were based on: The proposed model was trained using deep neural network architecture as shown in Fig. 4, at a learning rate Of 0.0005, input layer (11 features), dense layer (48 neurons), dense layer (20 neurons), and 4 classes of the output layer. Anomaly detection was included to locate outliers, which is proven good for identifying unidentified states and completing classifiers [18].

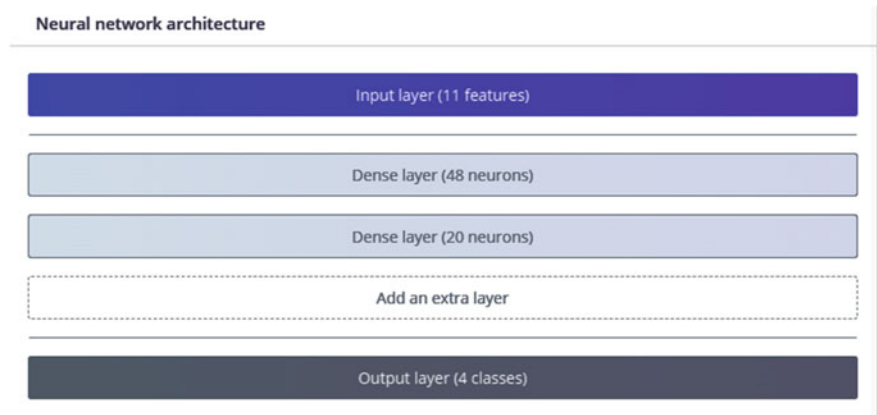
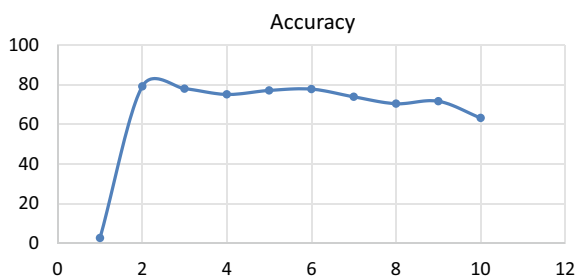


Fig. 4 Neural network architecture

**Fig. 5** Accuracy of all-time intervals from 1 to 10 seconds



### 3.3 Results and Performance Analysis

#### 3.3.1 Feature Analysis

The time-series dataset was done using different time intervals of observation from 10 to 1 s. The dataset for 10-s time intervals has 128 observations while 1 s has 1287 observations. The dataset of 1-s observations was chosen since it has proven that it has more accuracy than others, see in Fig. 5, and also an analog PIR sensor was used that was able to detect an object in 6 ms.

#### 3.3.2 DSP and Classification Analysis

Classification performance was analyzed based on the training accuracies for selected input features. TinyML model is designed to classify moving objects. Upon testing, animals are correctly classified with 80.1%, animals and humans are 17.5%, animals and no object are 0.9% related, and animals and wind are 1.4% related. Humans are correctly classified with 75.4%, humans and no object are 0% related, human and wind 4.4% related, no object and animal are 0% related, no object and Human are 0% related, no objects are 100% correctly classified, no object and wind are 0% related, wind and animal are 4.9% related. Wind and human are 27.2% related, wind and no object are 0% related, and the wind is 68% correctly classified. So, the  $F1$  score as one of the significant evaluation metrics in ML was used, where the animal has 78%, the human has 68%, no object has 100% and wind has 78%, as shown in Fig. 6.

Thus, the features based on DSP: PIR values RMS, PIR values skewness, PIR values kurtosis, PIR values spectral power 5.21–15.62 Hz, PIR values spectral power 15.62–26.04 Hz, PIR values spectral power 26.04–36.46 Hz, PIR values spectral power 36.46–46.88 Hz were able to give us the best performing model with accuracies reaching 80.8% with the loss of 0.47 for both training and testing classification for targeted classes: animals, humans, no objects, and wind (windblown vegetation) and anomaly detection shows that all inputs for model inference are valid for the targeted moving (Table 2).

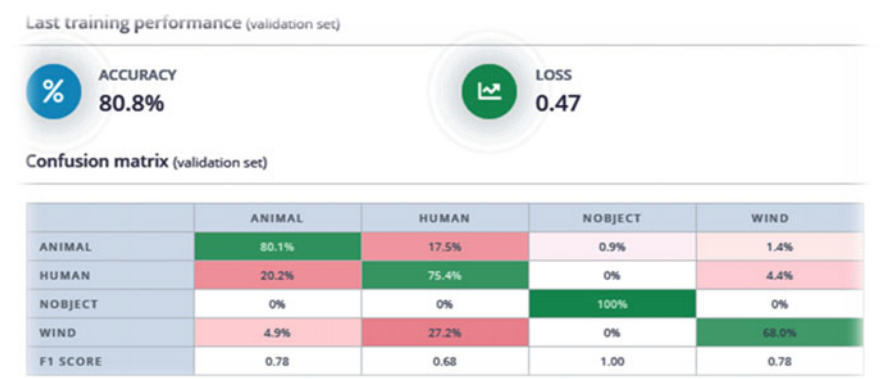


Fig. 6 Confusion matrix for training performance

Table 2 Comparison of DSP block parameter’s performance

DSP block parameters	Input layers	Accuracy (%)	Loss
Raw values	166	24.8	1.38
Flatten	7	79.8	0.70
Spectral analysis	11	80.8	0.47

3.3.3 Real-time Inferencing on Embedded Platform

The performance of the TinyML model during the training phase was measured at first using a cloud-based run-time environment as shown in Fig. 7. This part is the first step toward deployment focusing on an edge run-time environment. We had to ensure that the model works properly before converting it to TinyML.

The TinyML model for the NN classifier developed in the machine learning training section was optimized and compressed into the Arduino library designed

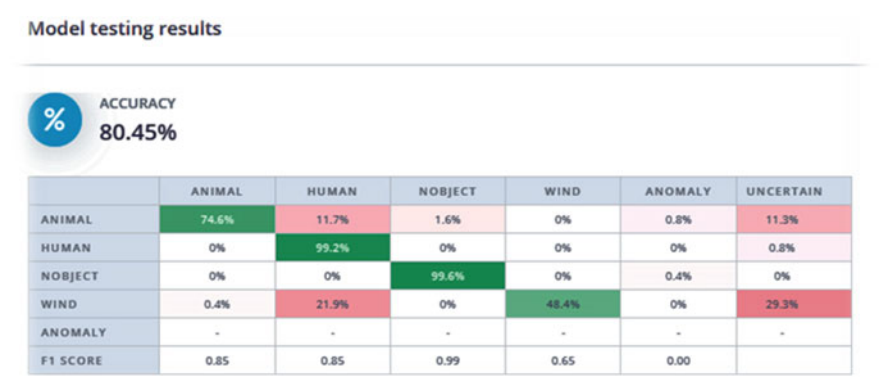


Fig. 7 Confusion matrix for test data

```
19:09:19.609 -> run_classifier returned: 0
19:09:19.609 -> Predictions (DSP: 27 ms., Classification: 0 ms., Anomaly: 4 ms.):
19:09:19.609 -> animal: 0.87891
19:09:19.609 -> human: 0.12109
19:09:19.609 -> nobject: 0.00000
19:09:19.609 -> wind: 0.00000
19:09:19.609 -> anomaly: -0.427
```

**Fig. 8** On-device inferencing targeting animal class

to run on Arm-based development boards on device inference which was done on different moving objects. We defined our DSP block size to schedule our inferencing intervals. Based on the frequency of the sensor, we feed 166 values which are equivalent to one observation then proceed to perform inference. Thus, the inference happens every minute, as shown in Fig. 8.

## 4 Conclusion and Future Works

This paper developed a cost-effective solution that performs the detection of moving objects at the edge by using TinyML with analog PIR time-series data as a viable real-time alternative to the intensive image processing solutions with respect to real-time resources such as cost, energy, processing time, and so on. The trained model classifies moving objects based on patterns of PIR signals bounced back from different objects. This solution has the potential to be used on different types of applications requiring non-invasive sensing such as monitoring house security, surveillance of eligible objects at a given place, proximity sensing for infection diseases spreading limitation, and so on. With respect to the state of the art, the main contributions of this paper are (1) constructing a dataset of analog PIR time-series data for humans, dogs, goats, and windy vegetables which can be used by other researchers, and (2) co-design of DSP techniques and TinyML to process analog PIR data in order to achieve high accuracies of classification of moving objects. Through pilot experimentation, an inference performance accuracy of 80.8% has been achieved. This accuracy may be improved over time using reinforcement learning. Future work includes applying the method to develop a wearable device that can monitor infectious diseases.

**Acknowledgements** This work was jointly supported by the African Center of Excellence in Internet of Things (ACEIoT) College of Science and Technology, University of Rwanda, and the National Council for Science and Technology (NCST) for Rwanda.

## References

1. Nguyen CT et al (2020) A comprehensive survey of enabling and emerging technologies for social distancing—part II: emerging technologies and open issues. *IEEE Access* 8:154209–154236. <https://doi.org/10.1109/access.2020.3018124>
2. Wang F, Zhang M, Wang X, Ma X, Liu J (2020) Deep learning for edge computing applications: a state-of-the-art survey. *IEEE Access* 8:58322–58336
3. Choubisa T et al (2017) An optical-camera complement to a PIR sensor array for intrusion detection and classification in an outdoor environment. In: *Proceedings of the 2017 IEEE 42nd conference on local computer networks workshops (LCN workshops)*. <https://doi.org/10.1109/lcn.workshops.2017.63>
4. Mondal A, Shashant R, Giraldo JH, Bouwmans T, Chowdhury AS (2021) MovingObject detection for event-based vision using graph spectral clustering. In: *Proceedings of the 2021 IEEE/CVF international conference on computer vision workshops (ICCVW)*. <https://doi.org/10.1109/iccvw54120.2021.00103>
5. Shaikh SH, Saeed K, Chaki N (2014) Moving object detection approaches, challenges, and object tracking. In: *Moving object detection using background subtraction*, pp 5–14. [https://doi.org/10.1007/978-3-319-07386-6\\_2](https://doi.org/10.1007/978-3-319-07386-6_2)
6. Yun J, Woo J (2020) A comparative analysis of deep learning and machine learning on detecting movement directions using PIR sensors. *IEEE Internet Things J* 7(4):2855–2868. <https://doi.org/10.1109/JIOT.2019.2963326>
7. Andrews J, Vakili J, Li J (2020) Biometric authentication and stationary detection of human subjects by deep learning of passive infrared (PIR) sensor data. In: *Proceedings of the 2020 IEEE signal processing in medicine and biology symposium (SPMB)*. <https://doi.org/10.1109/spmb50085.2020.9353613>
8. Yang T, Guo P, Liu W, Liu X, Hao T (2021) Enhancing PIR-based multi-person localization through combining deep learning with domain knowledge. *IEEE Sens J* 21(4):4874–4886. <https://doi.org/10.1109/jsen.2020.3029810>
9. Nguyen CT et al (2020) A comprehensive survey of enabling and emerging technologies for social distancing—part I: fundamentals and enabling technologies. *IEEE Access* 8:153479–153507. <https://doi.org/10.1109/access.2020.3018140>
10. Using PIR sensors for motion detection (2019) Upverter Blog, Oct 17, 2019. <https://blog.upverter.com/2019/10/17/using-pir-sensors-for-motion-detection/>. Accessed 04 Nov 2022
11. Kashimoto Y, Fujiwara M, Fujimoto M, Suwa H, Arakawa Y, Yasumoto K (2017) ALPAS: analog-PIR-sensor-based activity recognition system in smarthome. In: *Proceedings of the 2017 IEEE 31st international conference on advanced information networking and applications (AINA)*. <https://doi.org/10.1109/aina.2017.33>
12. Wu C-M, Chen X-Y, Wen C-Y, Sethares WA (2021) Cooperative networked PIR detection system for indoor human localization. *Sensors* 21(18):180. <https://doi.org/10.3390/s21186180>
13. Mukhopadhyay B, Srirangarajan S, Kar S (2018) Modeling the analog response of the passive infrared sensor. *Sens Actuat A Phys* 279:65–74. <https://doi.org/10.1016/j.sna.2018.05.002>
14. Oraşan IL, Seiculescu C, Căleanu CD (2022) A brief review of deep neural network implementations for ARM cortex-M processor. *Electronics* 11(16):2545. <https://doi.org/10.3390/electronics11162545>
15. Ogore MM, Nkurikiyeyezu K, Nsenga J (2021) Offline prediction of cholera in rural communal tap waters using edge AI inference. In: *Proceedings of the 2021 IEEE globecom workshops (GC Wkshps)*, pp 1–6. <https://doi.org/10.1109/GCWkshps52748.2021.9682128>
16. Jackovich J, Richards R (2018) *Machine learning with AWS: explore the power of cloud services for your machine learning and artificial intelligence projects*. Packt Publishing, Birmingham



17. Cerutti G, Prasad R, Brutti A, Farella E (2020) Compact recurrent neural networks for acoustic event detection on low-energy/low-complexity platforms. *IEEE J Sel Top Signal Process* 14:654–664
18. Train model anomaly (2022). <https://docs.edgeimpulse.com/reference/edge-impulse-api/jobs/train-model-anomaly>. Accessed 30 Oct 2022

# Advanced Signaling Mechanisms for Assurance of User Service Continuity in 4G/5G Mobile Network



Diep Pham Quang, Hung Nguyen Tai, Hoan Nguyen Dac, and Tu Le Minh

**Abstract** The Evolve Packet Core (EPC) System is a complicated network and have to manage a huge number of connections with E-UTRAN, up to thousands of SCTP connections and each of connection with E-UTRAN supports services for thousands of user subscriptions. Therefore, the reliability of these SCTP connections are critical because any problem with them will affect the large number of users. Theoretically, there are some solutions to assure the high availability for these connections but reality shows that we still lack of the field-proven one. That said, on this paper, we present our proposed solution that combine the mechanism of equal-cost multipath with Re-initiate mechanism of SCTP to provide the high reliability for connection between EPC and E-UTRAN in 4G/5G networks. Our solution has been tested on the real network environment and with huge traffic volume and big number of subscriptions.

**Keywords** EPC · E-UTRAN · SCTP connections · Switching · High availability · ECMP · Resilient hashing

---

D. P. Quang (✉) · H. N. Dac · T. Le Minh  
Viettel High Technology Corporation, Hanoi, Vietnam  
e-mail: [dieppq@viettel.com.vn](mailto:dieppq@viettel.com.vn)

H. N. Dac  
e-mail: [hoannd14@viettel.com.vn](mailto:hoannd14@viettel.com.vn)

T. Le Minh  
e-mail: [tulm4@viettel.com.vn](mailto:tulm4@viettel.com.vn)

H. N. Tai  
School of Electrical and Electronic Engineering, Hanoi University of Science and Technology,  
Hanoi, Vietnam  
e-mail: [hung.nguyentai@hust.edu.vn](mailto:hung.nguyentai@hust.edu.vn)

# 1 Background

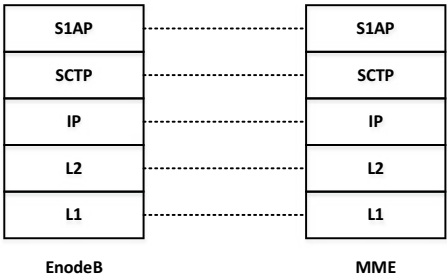
The Long Term Evolution (LTE) Network called Evolved Packed System (EPS) is an end-to-end (E2E) all IP network; EPS is divided into two parts: radio access network (E-UTRAN) and core network (EPC). An E2E all IP network means that all traffic flows—from a UE all the way to a Packet Data Network (PDN) which connects to a service entity—are transferred based on IP protocol within EPS. EPC system includes functional blocks of Mobility Management Entity (MME), Serving Gateway (SGW) and PDN Gateway (PGW). More details of EPC systems are on [1, 2].

In EPC, MME is a logical entity responsible for authentication, session management and mobility management for the subscriber [3]. It also connects E-UTRANS (eNodeB) and EPC using the S1AP interface (Fig. 1), which make use of SCTP at the transport layer. The major requirements of the interface between MME and E-UTRAN are as below:

- Single IP Service: MME has only one IP service for S1AP interface, which used to listen the connection from E-UTRAN.
- High number of connections: up to thousands of connections. For example in Viettel (one of the biggest MNOs in Vietnam) networks, one MME can manage over 17,000 SCTP connections with E-UTRAN. Each of EnodeB in E-UTRAN can support 3000 ~ 5000 user subscription.
- Ensure High Availability: minimize the impact when switching connections between systems upon crashing or during maintenance.

Currently, 3GPP has no specifications of how to meet this requirements of high availability and heavy traffic conditions and that causes a lot of difficulties when implement and operate the 4G (and 5G too) networks in reality. On this paper, we will present our proposition of using advanced mechanisms on signaling and routing of SCTP connections to meet those application requirements in the real network operation environment. The proposed solution has been tested and proved high efficiency and feasibility.

**Fig. 1** Protocol stack at S1-MME interface



## 2 Previous Solutions

### 2.1 Overview SCTP Protocol

SCTP is a reliable transport protocol operating on top of a connectionless packet network such as IP. It offers the following services to its application users [4]:

- Acknowledged, error-free and non-duplicated transfer of user data
- Data fragmentation to conform to discovered path MTU size
- Sequenced delivery of user messages within multiple streams
- With an option for order-of-arrival delivery of individual user messages
- Optional bundling of multiple user messages into a single SCTP packet
- Network-level fault tolerance through supporting of multi-homing at either or both ends of an association.

SCTP provides the means for each SCTP endpoint to provide the other endpoint (during association startup) with a list of transport addresses (multiple IP addresses in combination with an SCTP port) through which that endpoint can be reached and from which it will originate SCTP packets. SCTP supports procedures for establishing and managing connections and for fault management as specified in RFC 4960 [4].

Normally, when SCTP server have operational problem, SCTP clients will detect and terminate SCTP connection, after that try to reconnect to SCTP Server. However, in S1-MME interface of mobile networks, that practice will interrupt the user services because of the connection lost. As such, we need to find the solution to overcome it.

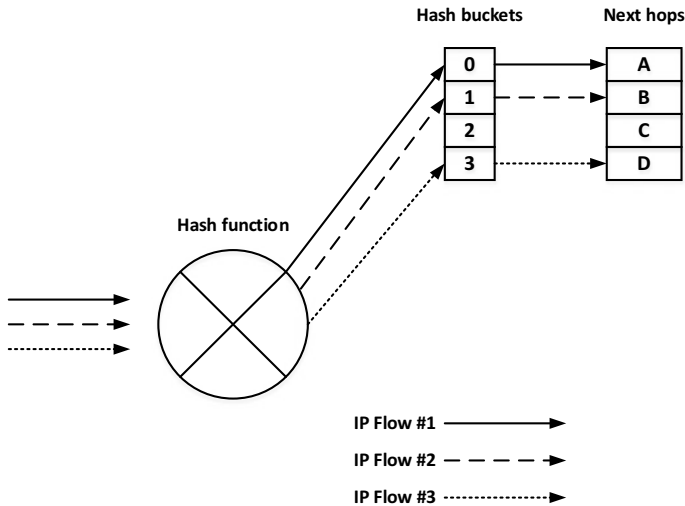
### 2.2 Overview of the Equal Cost Multipath (ECMP) Function

ECMP is a routing technique that allows the router to distribute flows to the single target destination in to different paths rather than in a single one and that brings following benefits:

- Increase transmission performance: packets can be forwarded from the router to destination concurrently on different paths.
- Improve fault tolerance: When one path is deleted from the routing table, packets that previously forwarded on that path are redistributed on remaining path.

ECMP can be implemented in two methodologies [5]:

- **Per-packet load balancing:** Packets are forwarded to paths evenly using Round-robin algorithm. This method ensures IP flows are distributed equally in all paths. However, different transmission latency between paths could cause packets in the same IP flow to arrive at the destination in wrong order, hence badly affected to the throughput of protocols that require sequencing of data like TCP and SCTP.



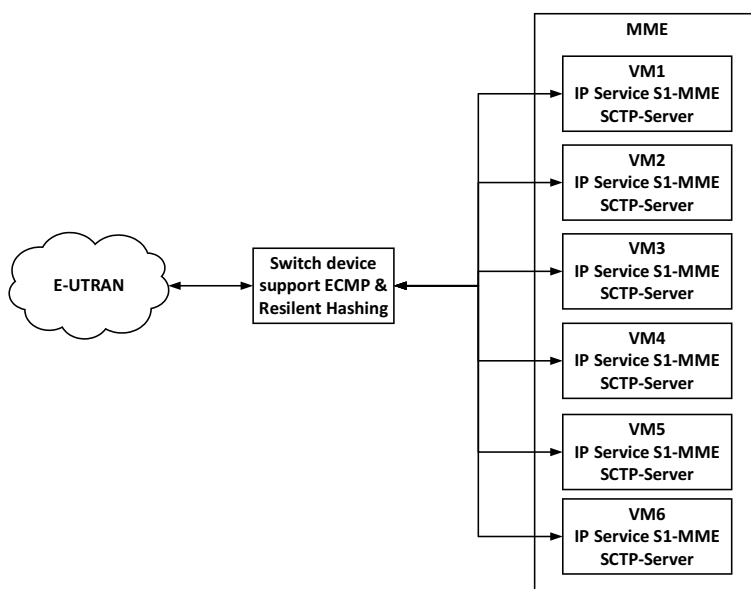
**Fig. 2** ECMP using hash function to distribute IP flows on different routes

- **Per-destination load balancing:** Using a load balancing algorithm to ensure packets in the same IP flow are transmitted at the same path and avoid the wrong ordering transmission problem. The algorithm calculates a hash value of each packet using several fields of packet header, such as source address, destination address, source port, destination port, payload protocol ID, ... Packets in the same IP flow have a similar header, so their calculated hash value are identical. To that end, they will be sent on the same path. The details are depicted in Fig. 2.

In order to minimize IP flows effect when a routing path is added or removed from the routing table, we need to use the so-called Resilient Hashing [6] mechanism. This mechanism will make sure that when a next-hop address corresponding with the path is deleted from the routing table, only flows transmitting on that path are redistributed to others; the rest are not affected.

### 3 Proposed Solution

In order to solve the problem of service interruption upon the (SCTP) connection switching as explained above, our research team has come up with a solution of combining the ECMP routing mechanism with SCTP Re-initiation procedure, the detailed architecture and operation steps are as follow.



**Fig. 3** High availability architecture with advanced mechanisms for the interface between E-UTRAN and EPC

### 3.1 Proposed Architecture

A VM Cluster that contains multiple VMs will act as a logical SCTP server with the same IP Service (single IP service as network's requirements). The application support processing re-init messages from server when any VM have been failure. We implemented all VMs on RHEL 7.9 with the package of `lksctp-tools-devel-1.0.17-2.el7.x86_64`. We have also modified the commercial switch device to support ECMP and Resilient Hashing mechanisms and functions. On the E-UTRAN, we use our in-house developed eNodeB to initiate the big volume of connections to MMEs running on the VM cluster (Fig. 3).

### 3.2 Assurance of User Service Continuity

Typically, eNodeBs have to reconnect when SCTP association to an MME server is lost. To avoid this, other active servers in cluster can directly trigger SCTP INIT message to eNodeB to re-establish the association with the new server (used flow SCTP\_Association restart procedure as specified in the RFC 4960 [4]). Our work has proposed the solution to reduce timeout periods that eNodeB need to wait before re-establishing the connection, as well as not interrupting the user services. The

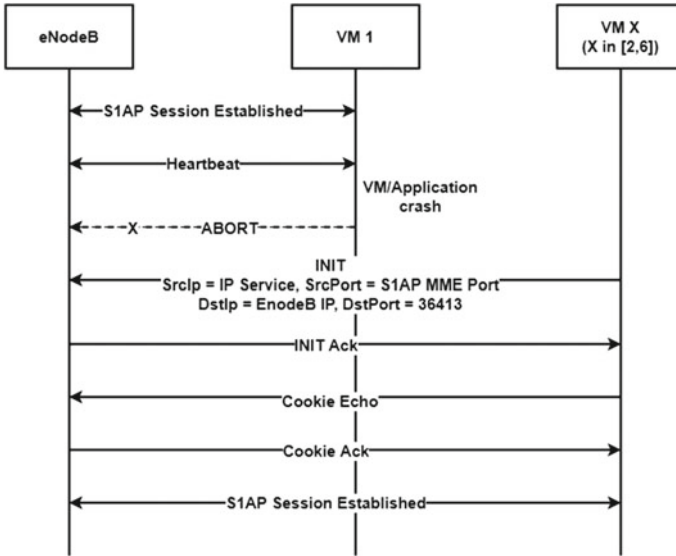


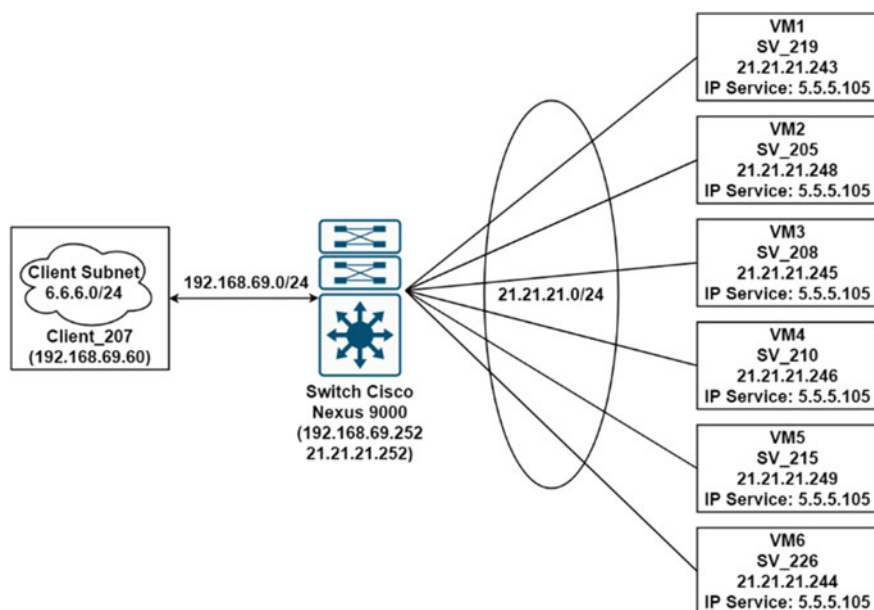
Fig. 4 Working flow of S1AP Connection **re-init** procedure to assure the service continuity

detailed working flow of the proposed solution are depicted in Fig. 4 and explained as working steps below.

- **Step 1:** eNodeB (*eNodeB S1AP IP, eNodeB S1AP Port*) successfully establish association to S1AP GW #1 node (*MME S1AP IP, MME S1AP Port*).
- **Step 2:** VM #1 node crashes, which causes the underlying SCTP stack to abort all SCTP associations. eNodeB does not know about this crash since IPGW blocks the ABORT message.
- **Step 3:** VM #X with X in [2, 6] identifies VM #1 crash and then performs **re-init** procedure with eNodeB by sending **INIT** message with source IP = MME S1AP IP, destination IP = eNodeB S1AP IP, source port = MME S1AP port, destination port = eNodeB S1AP port.
- **Step 4:** eNodeB receives **INIT** message and recognizes that MME has actively re-established the association. It responds with **INIT\_ACK** message to the MME to accept the new association.
- **Step 5:** VM #X node receives **INIT\_ACK** message and continues to establish the association as usual. S1AP connection has not been affected.

## 4 Implementation

We have developed the proposed architecture on our laboratory and setup the test-bed to run and test the working principle as well as the performance and reliability of the system as illustrated on Fig. 5.



**Fig. 5** Test-bed implementation and integration

On this test-bed, six-clustered VMs act as a logical SCTP server, each opening SCTP socket listens at IP service 5.5.5.105:36412. Server broadcasts IP service 5.5.5.105 to switch using OSPF protocol through internal IP subnet 21.21.21.0/24.

Each VM's imitating a SCTP client to establish SCTP connection. Each client opens a socket binding to a unique IP address in the 6.6.6.0/24 subnet. All VMs are running MME applications as SCTP server with following parameters:

- SCTP hb\_interval: 2000 ms
- SCTP path\_max\_retrans: 5
- SCTP rto\_max: 2000 ms
- IPGW for forwarding messages to upper layer (business functions processing) and blocking ABORT message [7].

For the load balancing function, we use the resilient hashing mechanism implemented on the modified switch device. The modified switch then will be configured with following configuration parameters [5, 8, 9]:

- Enabling resilient hashing:  
hardware profile ecmp resilient.
- Configuring IP load sharing:  
ip load-sharing address source-destination port source-destination rotate 32 universal-id 13412341.
- Setting OSPF hello and dead interval:  
ip ospf hello-interval 1.  
ip ospf dead-interval 4.



**Table 1** Performance of the distribution of SCTP connections

Node	Number of clients
SV_205	38
SV_208	36
SV_210	37
SV_215	46
SV_219	42
SV_226	41

## 5 Performance Testing and Reliability Evaluation

### 5.1 Evaluation of the Connection Distribution Function

The goal of this test is to figure out the performance of distribution of SCTP association across logic nodes in the cluster. To achieve this, we will let the SCTP client try to attempt to establish 240 associations to the VMs cluster. After the associations are successfully established, the number of SCTP associations on each logic node is measured as in Table 1.

### 5.2 Evaluation of (VM) Service Recovery

This test's purpose is to evaluate the impacts on the SCTP associations when one logic node in the cluster goes down. The test is executed in the following steps:

- Step 1: Establishing of 240 SCTP associations from SCTP clients as in the previous test.
- Step 2: after the SCTP associations are established, shut down the *ospfd* process on SV\_205. Inspecting the status and distribution of the SCTP associations.
- Step 3: Start the *ospfd* process on SV\_205 and inspecting the impact on the SCTP associations again.

After turning off the OSPFD process on SV\_205 and the period specified by the OSPF dead interval (4 s in this particular case), the switch removes the route to SV\_205 from the routing table. At this time, thus there are only 5 routes with next-hop point to the IP Service (instead of 6 as before), the switch performs update mapping between IP flows previously belong to SV\_205 to the remaining active nodes and forward messages from the client to the those nodes. The SCTP stack on the new serving node (e.g., SV\_208), after receiving a packet that belong to the association on SV\_205, sends back an INIT message to the client process. Finally, the SCTP client update new association and change the connection to the new serving node without interruption.

The testing results show that all the SCTP associations between the client processes and SV\_205 are transferred to a new node in the cluster.

Start the *OSPF*D process on node SV\_205 again. The switch, after determining that node SV\_205 is back on service, redistributes all the flows that previously connect to SV\_205 to SV\_205 again. SCTP stack on SV\_205 receives SCTP packets of the current active association between SCTP client processes and the serving node (e.g., SV\_208), figures out that the association is unknown and proceeds to send an *INIT* message back to the client to update association. After receiving the *INIT* message from SV\_205, the SCTP client process update association to SV\_205. Finally, the old association on the previous serving node times out (because the SCTP packet from the SCTP client is now forwarded to SV\_205) and gets terminated.

The number of SCTP clients on each node after SV\_205 is back on service is shown in the following Table 3.

We can clearly see from Tables 2, 3 that the SCTP client distribution is exactly the same as before SV\_205 goes down. On top of that, only those clients served by SV\_205 are affected; while the others work seamlessly during the entire (interruption) test.

**Table 2** Performance of the distribution of SCTP connections after SV\_205 goes down

Node	Number of clients
SV_205	0
SV_208	43
SV_210	50
SV_215	52
SV_219	49
SV_226	46

**Table 3** Performance of the distribution of SCTP connections after SV\_205 come back

Node	Number of clients
SV_205	38
SV_208	36
SV_210	37
SV_215	46
SV_219	42
SV_226	41

**Table 4** Distribution of Sctp connections after SV\_205,208 down

Node	Number of clients
SV_205	0
SV_208	0
SV_210	61
SV_215	63
SV_219	62
SV_226	54

**Table 5** Distribution of Sctp connections after SV\_205 come back

Node	Number of clients
SV_205	53
SV_208	0
SV_210	34
SV_215	56
SV_219	50
SV_226	47

5.3 Connection Impacts When 2 Nodes Go Down

This test’s purpose is to evaluate the impacts on the Sctp associations when two logic nodes in the cluster goes down. The test is executed in the following steps:

- Step 1: Establishing of 240 Sctp associations from Sctp clients as in the previous test.
- Step 2: After the Sctp associations are established, shut down the *ospfd* process on SV\_205 and SV\_208. Inspecting the status and distribution of the Sctp associations.
- Step 3: Start the *ospfd* process on SV\_205 and SV\_208, respectively, and inspecting the impact on the Sctp associations again.

After shutting down the ospfd service on SV\_205 and SV\_208, the Sctp associations currently served by these nodes are redistributed to the remaining active nodes, just like the case one server goes down above (Tables 4, 5, 6).

After turning on the service on SV\_205, several clients on active nodes are moved to SV\_205 to ensure a relatively equal load between nodes:

When SV\_208 is back on service, the client distribution goes back to the state when all nodes are active as well. That make sure the user services are un-interrupted during the MME failure.

**Table 6** Distribution of Sctp connections after SV\_208 come back

Node	Number of clients
SV_205	38
SV_208	36
SV_210	37
SV_215	46
SV_219	42
SV_226	41

6 Conclusion

On this paper, we have studied the problem of the MME node-crashing on the real 4G mobile networks and proposed the solution of combining the ECMP routing function with the Sctp re-init mechanism to overcome the said problem and ultimately assuring the user service continuity. We have implemented and evaluated the proposed solution on the test-bed setup that resuming the real network operation environment. We emulated the evenly distribution of the Sctp connections to different VMs (meaning MMEs), the testing results show that when a specific VM goes down, the existing Sctp connections on that VM will be distributed to the remaining VMs (aka. MME) and immediately initiating the process of re-init from Sctp server to EnodeB to ensure high availability. All the tests are done with high number of Sctp connections (aka. Clients) as specified on the tables, those figures are almost same with the real 4G commercial network environment.

References

1. <https://www.netmanias.com/en/post/techdocs/5904/lte-networkarchitecture/lte-network-architecture-basic>
2. 3GPP TS 23.401 V15.5.0 (2018–09) 3rd Generation partnership project; technical specification group services and system aspects; general packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access
3. 3GPP TS 36.413 V15.10.0 (2020–04) Evolved universal terrestrial radio access network (E-UTRAN); S1 application protocol (S1AP)
4. Network working group request for comments: 4960 R.Stewart, Ed. September 2007 stream control transmission protocol
5. <https://www.cisco.com/c/en/us/td/docs/security/securefirewall/management-center/device-config/710/management-centerdevice-config-71/routing-ecmp.pdf>
6. [https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/interfaces/configuration/guide/b\\_Cisco\\_Nexus\\_9000\\_Series\\_NXOS\\_Interfaces\\_Configuration\\_Guide\\_7x/b\\_Cisco\\_Nexus\\_9000\\_Series\\_NX-OS\\_Interfaces\\_Configuration\\_Guide\\_7x\\_chapter\\_0111.html](https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/interfaces/configuration/guide/b_Cisco_Nexus_9000_Series_NXOS_Interfaces_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_Interfaces_Configuration_Guide_7x_chapter_0111.html)
7. <https://www.redhat.com/sysadmin/iptables>
8. <https://www.cisco.com/c/en/us/support/docs/ip/open-shortest-path-firstospf/7039-1.html>
9. <https://hub.packtpub.com/introduction-to-open-shortest-path-first-ospftutorial>

# The BB84 Quantum Key Distribution Protocol and Potential Risks



Maria E. Sabani, Ilias K. Savvas, Dimitrios Poulakis, George C. Makris,  
and Maria A. Butakova

**Abstract** Quantum computing is not only a revolution in information science, but it is going to cause colossal changes in almost all sciences and especially in telecommunications and cryptography. In this work, we present a quantum key distribution protocol, BB84, and also two possible attacks against this cryptographic scheme. Even though quantum cryptosystems are considered to be secure as they are based on the laws of quantum mechanics, always there are loopholes, and it is really urgent to prepare for the fast-approaching era of quantum computing.

**Keywords** Quantum computing · Quantum cryptography · Quantum key distribution (QKD) · Intercept resend attack · Trojan horse attack

## 1 Introduction

Quantum computing is a hodgepodge of quantum physics, mathematics, and computing and is a fascinating method that uses quantum mechanics to achieve extremely

---

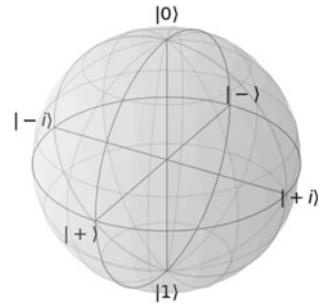
M. E. Sabani (✉) · I. K. Savvas · G. C. Makris  
Department of Digital Systems, University of Thessaly, Larissa, Greece  
e-mail: [masampani@uth.gr](mailto:masampani@uth.gr)

I. K. Savvas  
e-mail: [isavvas@uth.gr](mailto:isavvas@uth.gr)

G. C. Makris  
e-mail: [makris@uth.gr](mailto:makris@uth.gr)

D. Poulakis  
Department of Mathematics, Aristotle University of Thessaloniki, Thessaloniki, Greece  
e-mail: [poulakis@math.auth.gr](mailto:poulakis@math.auth.gr)

M. A. Butakova  
Smart Materials Research Institute, Southern Federal University, Rostov, Russia  
e-mail: [mbutakova@sfedu.ru](mailto:mbutakova@sfedu.ru)

**Fig. 1** Bloch sphere

fast computations as well as in a flash of eye secure telecommunications. In a classical computer, the basic unit of information is the bit, however, the quantum bit or qubit (quantum BIT) is the cornerstone of quantum computational systems. It is a mathematical object that can be used in a real physical system.

A bit can take value 0 or 1 while a qubit can be in basic states  $|0\rangle$  or  $|1\rangle$  in every linear combination of the two states, such that  $a|0\rangle + b|1\rangle$ ,  $a, b \in \mathbb{C}$ ,  $\wedge a^2 + b^2 = 1$ . Geometrically, qubits can be visualized by using a geometric shape known as Bloch Sphere (Fig. 1), named by the Swiss physicist Felix Bloch. Bloch sphere is a geometric representation in a three-dimensional Hilbert space of the state of a quantum system of two levels or qubits. The north and south poles of the sphere represent the typical basis vectors  $|0\rangle$  and  $|1\rangle$ , respectively. A fundamental principle of quantum is quantum superposition where two quantum states can be added, and their sum is another valid quantum state [1]. If we want to measure the state of a qubit, we take the value  $|0\rangle$  with possibility  $|a|^2$ , the value  $|1\rangle$  with possibility  $|b|^2$ , such that  $|a|^2 + |b|^2 = 1$ . A qubit in a quantum computer can be or stored in two states simultaneously, and therefore, the quantity of information that can be stored in a quantum computer is huge. Quantum supremacy is a property of quantum computers and is a proof that a quantum computer can solve any problem that a classical computer cannot in a reasonable time.

Another important and rare property that qubits have is the quantum entanglement [1]. Quantum entanglement is the phenomenon that two particles interact each other irrespective of the space there is between them. If one particle is at one point of the universe and the other in the other point of the universe, if something happens to one of them, the other will react instantly. While the qubits are in contact the information travels at lightning speed. This phenomenon had been characterized by Albert Einstein “*Spooky action at distance*”. This property is very useful as it allows us to measure the state of one qubit and know the state of the other qubit without any other measurement.

As a fact quantum computing is a new scientific field that is developing rapidly, and nowadays, real quantum computers exist with limited number of qubits and many technical problems, which in turn reduce their reliability [2, 3]. Quantum computers will shake up our world and will bring revolution to computer science, chemistry, drug

development, economy, cryptography, telecommunications, and almost all sciences will benefit from this revolution of information technology.

Cryptography is the science of secure communication, and fast and secure cryptographic schemes are a basic requirement both in computing and in our daily life [4]. Cryptography uses algorithms and cryptographic schemes to ensure the safe transmission of information and therefore the communication between users. Cryptosystems are based on mathematical methods, computer programs, and the right way of managing them, and they are composed by messages, keys, and encryption and decryption functions [4]. Certainly, quantum computers will have powerful capabilities that solve a wide range of problems and are capable of generating new threats at unprecedented speed and scale in modern cryptographic protocols. The rapid development of the Internet and the cyberattacks require strong cryptosystems, and the security of modern cryptographic schemes, like RSA, is disputed. The security of RSA protocol is based on fundamental mathematics and the problem of factoring the product of two large prime numbers [5]. In 1994, professor Peter Shor presented an algorithm, Shor's Algorithm, that with the aid of a quantum computer could solve the factoring and the discrete logarithm problem in polynomial time, and change the things up for the cryptosystems' security [6].

In this work, we present the applications of cryptography in quantum computers and how we can use quantum computers to exchange cryptographic keys with the method of quantum key distribution (QKD). We discuss about the most famous quantum key distribution protocol, the BB84, and we present some attacks on the cryptographic scheme.

The rest of the paper is organized as follows. In Sect. 2, it is presented some basic issues about quantum cryptography and the quantum key distribution protocols. In Sect. 3, it is given some types of attacks in quantum cryptosystems. Finally, Sect. 4 concludes this work and provides some future directions.

## 2 Quantum Cryptography

Quantum cryptography is the science that applies the principles of quantum mechanics to the encryption and transmission of information [7]. The methods of quantum cryptography practically exploit the properties of a physical quantum system to transmit or store data in a secure way. Typically, quantum cryptography has been applied classical cryptographic methods but these are extended and improved with the help of quantum mechanics [8]. One main issue in cryptographic schemes is the key exchange problem that is the method by which cryptographic keys for symmetric cryptosystems, as part of the algorithm, are exchanged between the two parties that communicate using a public not secure channel. The key in a cryptosystem is extremely important, as it is the one that can encrypt and decrypt a message exchanged by two users without being intercepted by a third user. In quantum cryptography, the quantum key distribution (QKD) is of great importance, and it is the prime achievement of quantum cryptography.

## 2.1 Quantum Key Distribution

Quantum key distribution is a procedure that exploits quantum physics to establish a secure telecommunication between two parties. So, quantum key distribution is the process of creating a private key between the sender and the receiver through a quantum channel, and moreover, this key is been used for encrypt messages through these two users. The main challenge is how this key can be exchanged secure in advance, and quantum mechanics provides a solution. Quantum key distribution is a robust procedure against an eavesdropping and retransmission attempt by a fraudulent user, as in the quantum world, the result of a measurement is not considered to reveal a specific value of a quantum state.

Cryptographic protocols that are deal with the processing and the execution of quantum key distribution use three basic quantum mechanical characteristics: quantum superposition, quantum entanglement, and the uncertainty principle. The uncertainty principle, or Heisenberg's uncertainty principle, mentions that we cannot measure at the same time two quantum states with perfect accuracy [9]. So, if someone wants to eavesdrop on communication will cause changes to the quantum system and will be detected by the two parties that communicate.

Assuming that we have two parties, A and B (users or telecommunication devices), that want to create a cryptographic key, and it is possible the existence of an eavesdropper. The QKD system consists of two channels, one quantum channel where is created and reproduced the private key and the classical channel where the two parties communicate each other [7]. With the laws of quantum physics and the polarization of light particles (photons), QKD encrypts data and becomes the safest communication method. Over the quantum channel, they are transmitted photons, and each photon has a random quantum state. The two devices A and B have an identical receiver that collects the photons and measures the polarization of each one (Fig. 2).

Due to the uncertainty principle, this measurement can detect any intrusion in the quantum key distribution transmission and reveal an eavesdropper. An eavesdropper in the effort to read the photons will change the state of the photons, and this change will reveal the information that have been elicited. Since there are many key

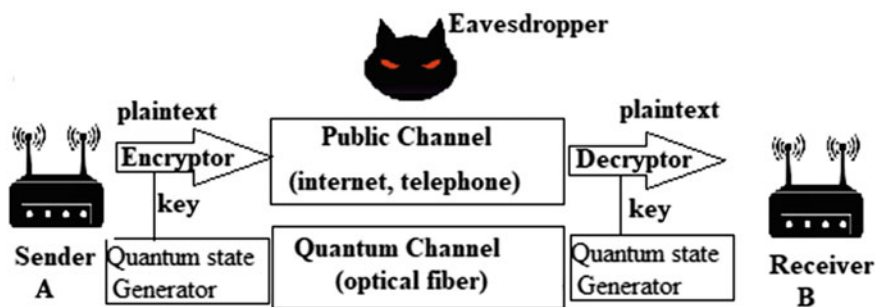


Fig. 2 Quantum key distribution



agreement protocols in public key cryptography, Diffie-Hellman key exchange protocol, KEA, etc., why QKD is useful? The answer is that the most of these protocols are based on discrete logarithm problem, a problem that Schor algorithm can solved in polynomial time. So, in the post quantum area, all these protocols are not secure. The first quantum key distribution protocol has been proposed by Bennett and Brassard and is known as BB84.

2.2    *The BB84 Protocol*

In 1984, Charles Bennett and Gilles Brassard presented a QKD protocol that is named by its creators and the year that was published [10]. Since then, many quantum key distribution protocols have been proposed but BB84 is been extensively analyzed and implemented. At BB84, protocol is used a quantum channel so that is created and reproduced a private key and a classical channel. The first user, Alice, selects a random bit (0 or 1) and therefore a random bit string. The polarization of the photons can be controlled with different perpendicular situations (bases), in order to encode 0 or 1. Consequently, Alice selects a sequence of polarization bases at random (+ or x) and sends to the second user, Bob, photons in one of the two bases chosen for that bit position. The rectilinear basis (+) encodes 1 in horizontal 0° bit value and 0 in vertical 90° bit value and the diagonal basis (x) encodes 0 in diagonal 45° bit value and 1 in diagonal 135° bit value, as it is presented in Table 1.

Bob receives the photons and decides, at random, to measure the photon’s polarization and explains the measurement’s result as binary 0 or 1. Then, the two users communicate in the public channel to compare the basis they used to polarize every bit. With this comparison, they agree on some situations and note the bits that they measured. When the users choose the same basis, they share the same bit. An example of the BB84 protocol is presented in Table 2.

Due to the fact that the photons are sent with random basis, there is the possibility an eavesdropper to try to change the transmission so as to create contention between Alice and Bob for some bits they used to agree. BB84 is proved to be secure but from theory to practice, there are several imperfections. As a consequence of the imperfections in the creation of photons, but also in their transmission and measurement, there are many ways to perform attacks on a quantum key distribution protocol.

**Table 1**   Polarization basis

Basis	0	1
+	↑	→
x	↗	↘

**Table 2** BB84 protocol

Alice's bits	0	1	1	0	1	0	0	1
Alice's basis	+	+	x	+	x	x	x	+
Photon alice sends	↑	→	↘	↑	↘	↗	↗	→
Bob's basis	+	x	x	x	+	x	+	+
Random receiving basis	↑	↗	↘	↗	→	↗	→	→
Bits received by Bob	0	0	1	1	0	0	1	1
Sifted key	0		1			0		1

### 3 The Security of Quantum Key Distribution Protocol

In theory, quantum key distribution protocols have been founded by the laws of quantum physics, and there are properly designed and implemented. The original papers that present QKD protocol, proved that is secure against certain attacks [10], and therefore there have been proposed several proofs of the security of the QKD protocols (for example [11, 12]). However, it cannot be achieved absolute security, meaning that there are loopholes at the application and the implementation of QKD protocols. The quantum system's flaws, the natural limitations in the detectors of the single photons, and the optoelectronic interfaces' imperfections are some of the reasons that question the security of the QKD protocols [13].

The main purpose of an eavesdropper is to manage to obtain information about the secret key through the telecommunication, without being detected by the users that communicate. So, there is a series of attentions and strategies to attack a QKD protocol and obtain information about the communication of the two users, and some types of attacks exploit these weaknesses and imperfections of the practical QKD system, which we mentioned before. Some of the most known attacks that can be employed are the intercept and resend attack (IRA), the faked state attack, the trojan horse attack, the photon number splitting attack, etc. [14, 15]. We will present two specific examples of attacks on QKD protocol, the intercept and resend attack and the trojan horse attack.

#### 3.1 Intercept and Resend Attack

We assume that there is an eavesdropper, Eve, that wants to obtain information of the communication of Alice and Bob. This type of attack must be realized in an ideal environment. Eve intercepts the signals, the light photons, that sends Alice and measures each of them. Eve according to the measurement result will prepare and send new signals to Bob though a noisy channel. Eve, since the environment is ideal and the detectors are efficient, holds the photons that Alice sends, she replaces them and sends them to Bob as per his polarization basis. Bob will receive the photons

**Table 3** Intercept resend attack

Alice’s bits	0	1	1	0	1	0	0	1
Alice’s basis	+	+	x	+	x	x	x	+
Photon alice sends	↑	→	↘	↑	↘	↗	↗	→
Eve’s basis measurement	+	x	+	+	+	+	x	+
Polarization Eve and sends	↑	↘	↑	↑	↑	↘	↑	↑
Bob’s basis	+	x	x	x	+	x	+	+
Bob’s measurement	↑	↗	↘	↗	→	↗	→	→
Shared secret key	0							1

with the same rate. Eve works as a person in the middle that performs the detection of the photons both from Alice’s and Bob’s side. Eve makes a successful attempt if she gains  $1/\sqrt{2}$  information of Alice [16]. An example of intercent and resend attack is presented in Table 3.

3.2 The Trojan Horse Attack

In most type of attacks against a QKD system, the eavesdropper attempts to measure the quantum states transmitted by the two parties that communicate. However, there is plenty of attacks that exploit the loopholes in the devices of the two parties or the optical setup which they have [17]. The trojan horse attack is also known as light injection attack, and Eve concentrates to the devices that are being used in the QKD protocol and the quantum channel’s weakness [18]. So, the eavesdropper tries through the quantum channel to have access at Alice’s or Bob’s apparatus. Eve absorbs the spatial, temporal, or frequency modes of the quantum channel to probe the apparatus of Alice. With the aid of an auxiliary source, which modulates, and a detector, analyzes the backscattered signal. Eve sends light pulses toward Alice’s or Bob’s setup, and therefore, the light pulse is reflected and returns to the detection apparatus that Eve possess [19].

This signal that is being reflected gives information regarding the basis, which Alice uses and the communication of Alice and Bob is insecure as Eve can obtain information on the secret key. When the reflected photons traverse a basis selector, they can obtain relevant information in the basis of the polarizer or the moderator. It is very important for the eavesdropper to maintain unnoticed onto the quantum channel, and in this case, he could break the security of any prepare and measure protocol. During the trojan horse attack, the light pulse that Eve sends, goes backward and forward through the attacked subsystem [20]. With the installation of a passive monitoring device, which is implemented by a detector, Alice can detect a trojan horse attack. This suitable detector can measure the incoming signal and when a pre-characterized signal inserts, the device detects it and raises an alarm.

These attacks can be avoided if the quantum key distribution system is properly prepared. In other words, the QKD apparatuses can be designed with filters and monitoring detectors that counter such type of attacks. An example is the “plug and play” system is presented in 2008 that Alice has already an auxiliary detector [21].

## 4 Conclusions

Since the beginning of the 1990s, the development in the field of quantum computing has been rapid, and we are going through a quantum era in the information science. A quantum computer uses the properties of quantum mechanics, and the astonishing result is its enormous computing power. With the transition to the quantum era, our world and all the sciences will be effected. Information science, cryptography, and telecommunications are the first that will change instantaneously.

Quantum cryptography is a powerful tool of the future to maintain the secure transmission of information. In this work, we discussed about methods that use quantum mechanics for the key exchange among users in a cryptographic scheme and specific the BB84 protocol. BB84 is considered to be a secure cryptographic system to create and reproduce a secret key among two users that communicate. We described the function of the protocol and also an example of it. Moreover, we discussed about the security of the protocol, and we presented two different type of attacks against QKD protocol, the intercept and resend attack and the trojan horse attack. Although quantum key distribution protocols were treated as the oasis of secure communication, due to the possible attacks, their reliability is questionable. So, there are still challenges in the field of quantum cryptography, and it becomes an active field of research and study.





## References

1. Nielsen MA, Chuang IL (2010) Quantum computation and quantum information. Cambridge University Press, New York
2. Galanis IP, Savvas IK, Chernov AV, Butakova MA (2021) Reliability testing, noise and error correction of real quantum computing devices. *Telfor J* 13(1):41–46
3. Galanis IP, Savvas IK, Garani G (2021) Experimental approach of the quantum volume on different quantum computing devices. In: The 14th international symposium on intelligent distributed computing—IDC 2021. Italy
4. Poulakis D (2004) Cryptography, the science of secure communication. Ziti Publications, Thessaloniki
5. Rivest RL, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. *ACM* 21:120–126
6. Sabani M, Galanis IP, Savvas IK, Garani G (2021) Implementation of Shor’s algorithm and some reliability issues of quantum computing devices. In: 25th pan-hellenic conference on informatics (PCI 2021). ACM, New York, USA, pp 392–296
7. Van Assche G (2006) Quantum cryptography and secret-key distillation. Cambridge University Press, New York

8. Bennett CH, Brassard G, Ekert A (1992) Quantum cryptography. *Sci Am* 50–57, Oct 1992
9. Panarella E (1987) Heisenberg uncertainty principle. *Ann Fondation Louis Broglie* 12(2):165–193
10. Bennett CH, Brassard G (1984) Quantum cryptography: public key distribution and coin tossing. In: *International conference in computer systems and signal processing*
11. Shor PW, Preskill J (2000) Simple proof of security of the BB84 quantum key distribution protocol. *Phys Rev Lett* 85:441–444
12. Renner R, Gisin N, Kraus B (2005) Information-theoretic security proof for quantum-key-distribution protocols. *Phys Rev A* 72:012332
13. Jain N, Khan I, Elser D, Leuchs G (2015) Attacks on practical quantum key distribution systems and how to prevent them. *Contemp Phys* 00(00):1–29
14. Gaur V, Mehra D, Aggarwal A, Kumari R, Rawat S (2020) Quantum key distribution: attacks and solutions. In: *3rd international conference on innovative computing and communication, ICICC 2020*
15. Curty M, Lutkenhaus N (2005) Intercept-resend attacks in the Bennett-Brassard 1984 quantum key distribution protocol with weak coherent pulses 3. [ArXiv: Quant-ph](#)
16. Aggarwal R, Sharma H, Gupta D (2011) Analysis of various attacks over BB84 quantum key distribution protocol. *Int J Comput Appl* 20(8):0975–8887
17. Vakhitov A, Makarov V, Hjelme DR (2001) Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography. *J Mod Opt* 48:2023
18. Gisin N, Fasel S, Kraus B, Zbinden H, Ribordy G (2006) Trojan horse attacks on quantum key distribution systems. *Phys Rev A* 73:022320
19. Khan A, Jain N, Stiller B, Marquardt C, Leuchs G (2015) Trojan-horse attacks on continuous-variable quantum cryptographic systems
20. Jain N, Anisimova E, Khan I, Makarov V, Marquardt Ch, Leuchs G (2014) Trojan-horse attacks threaten the security of practical quantum cryptography. *New J Phys* 16:123030
21. Muller A, Gisin N, Herzog T, Huttner B, Tittel W, Zbinden H (1997) “Plug and play” systems for quantum cryptography. *Appl Phys Lett* 70:793–795

# All Vaccinated: Open-Source Web System for the Control of Vaccination Processes in Health Centers



Lucrecia Llerena , Nancy Rodríguez , Ana Osorio , Rino Arias ,  
and John W. Castro 

**Abstract** In the aftermath of the COVID-19 pandemic, more and more people want to be vaccinated to prevent general health problems. In every country, there are health systems, which range from manual to automated systems. In general, private companies develop automated systems that sell software for the health sector. For this reason, it has been decided to develop an open-source web system that facilitates health centers to implement a system to control vaccination processes free of charge. The OSCRUM methodology was used to develop open-source software (OSS). With this methodology, OSS development can be carried out orderly and with a high success rate. OSCRUM facilitates the development of the web-based OSS system called All Vaccinated within the established timeframe. All Vaccinated for the control of vaccination processes is the result of the application of the OSCRUM development methodology. The OSS All Vaccinated web system will benefit health centers or institutions dedicated to vaccination in the optimization of their processes. It is concluded that choosing a development methodology has dramatically facilitated the construction processes for an OSS project and opened the door for new users to collaborate in developing the OSS project.

**Keywords** Open-source software · Software development methodologies · Open-source software development · Vaccines · Health

---

L. Llerena (✉) · N. Rodríguez · A. Osorio · R. Arias  
Universidad Técnica Estatal de Quevedo, Quevedo, Ecuador  
e-mail: [lllerena@uteq.edu.ec](mailto:lllerena@uteq.edu.ec)

A. Osorio  
e-mail: [aosorio@uteq.edu.ec](mailto:aosorio@uteq.edu.ec)

R. Arias  
e-mail: [rino.arias2018@uteq.edu.ec](mailto:rino.arias2018@uteq.edu.ec)

J. W. Castro  
Universidad de Atacama, Copiapó, Chile  
e-mail: [john.castro@uda.cl](mailto:john.castro@uda.cl)

# 1 Introduction

In the current global health context, there is a growing concern about whether people are prepared for a health crisis, such as the emergence of the COVID-19 virus. Based on this concern, many people and health professionals consider it essential for the population to be immunized to prevent diseases that cause significant epidemics, deaths, and sequelae [5]. For this reason, it is crucial and necessary to create a web system that allows the control and registration of the vaccines that the population has received, which are often offered by each country's government, quickly and efficiently by optimizing the vaccination processes. It is considered that being a critical web system. With a possible great demand in the health sector, it is necessary to establish a methodology to ensure the development of this open-source web system. Compared to traditional software engineering processes, the open-source software (OSS) development process emerges informally since rapid releases are made in a cooperative environment, which allows participants to develop software with a high functional and operational level for the community. Projects of this type are organized through virtual development communities, with communicated but globally dispersed participants cooperating to resolve conflicts, with access to source code being a shared resource [2]. OSS development projects do not have a defined development methodology [6]. For this reason, it is not easy to establish the necessary steps to develop the software project.

This research aims to develop an OSS web system using an open-source software development methodology, OSCRUM. The OSCRUM methodology has been chosen to develop this web system in a communitarian way, and that can be used in health centers that control the vaccination processes. This OSS web system will help the population to obtain information on the vaccines available in health institutions. In addition, it will allow system users to register to schedule an appointment and have a vaccine administered. Likewise, it will allow users to learn about the schedule of vaccines (for example, mothers with small children or infants to learn about the vaccines their children need and schedule an appointment), speeding up the process and benefiting a more significant part of the population. In addition, it will help healthcare personnel to optimize their vaccination processes, resulting in many more patients being seen.

This paper presents the following contributions. First, for the OSS development community, we have identified the automation problems in the health area, specifically in the field of vaccination, and propose solutions by implementing an OSS web system. Secondly, for the scientific community, we propose some tables and a web artifact to execute the activities of the OSCRUM development methodology to serve as a guide for OSS development enthusiasts.

The paper is organized as follows. Section 2 describes the related work. Section 3 describes the OSCRUM methodology for developing the OSS web system. Section 4 describes the case study. Section 5 describes the results of the case study. Section 6 discusses the results and mentions the limitations encountered, and finally, Sect. 7 describes the conclusions and future work.

## 2 Related Works

In the work of Aizaga-Villon et al. [1], the design of the computer architecture called “FIREWARE-bases” is presented, which integrates various network tools to collect patient information through a web application. The overall objective of this web application is that physicians and health personnel can manage patient data remotely to provide adequate and timely care. The research work does not record the use of a software development methodology. Tylutki et al. [8] presents a tool designed to predict and visualize the temporal concentration profiles of a parent compound and a metabolite in venous plasma and cardiac tissue following oral or intravenous drug administration. This tool is built on the R-environment framework and supports the features of shiny applications, such as interactive visualization of results, and the default web application interface. However, it does not present a software development methodology.

The work carried out by Siles et al. [9] shows an embedded and integrated accident and crisis management system capable of integrating with the systems of several Brazilian governmental institutions using open-source tools. However, it does not record the use of a software development methodology. In work carried out by Saini et al. [7] mentions that Case-Based Learning (CBL) is a teaching methodology based on discussion and analysis of real-world problems. There are several applications of CBL in medicine, law, and business. However, there is a limited amount of evidence related to applying CBL in the field of Software Engineering (SE). In this paper, an open-source web application called Software Engineering Case-Based Learning Database (SEABED) is presented, but it does not report the use of any software development methodology. The work by Rahman et al. [6] presents an open-source project-oriented software development methodology called OSCRUM, which is an adaptation of the SCRUM methodology and indicates the activities necessary for implementing this development methodology. However, this work does not report software development applying this methodology. In conclusion, only Rahman et al. [6] work reports a development methodology (i.e., OSCRUM). However, none of the works found applies the OSCRUM development methodology for developing an OSS web system. Therefore, more research efforts are required in this line, such as the one reported in the present work.

## 3 Description of the Proposed Methodology: OSCRUM

OSCRUM has been chosen as a development methodology for a web system because it allows the definition of a framework to structure and control the development of the open-source web system [6]. This section describes the selected OSCRUM methodology, its activities, and the technology used for the development of the open-source



web system. OSCRUM is a modification of the Scrum agile development methodology for open-source software development. This methodology has three fundamental pillars: transparency, inspection, and adaptation. This development methodology usually motivates developers to rely on five values for the success of their projects. These values are usually taken as a guideline and are as follows: Individual commitment, open-mindedness, being highly competent, focus on the goal, and collaboration [6].

In the study of Rahman et al. [6] it is described that OSCRUM development methodology differs from SCRUM development methodology in certain concepts such as (i) time box, (ii) stakeholders, and (iii) some SCRUM rules. At the same time, similarities between the two have also been found, such as (i) the characteristics of the team of developers (i.e., highly self-managed, cross-functional), (ii) acceptance of feedback in a short loop, (iii) the frequent release of the working version, (iv) integration and collaboration, and (v) customer engagement [6]. The open-source web system development is carried out under the OSCRUM framework [6] through the following activities, which are summarized in Table 1. The web system to be developed is open source. Therefore, it is proposed the use of open-source development tools, such as (i) MySQL Community [3], (ii) Angular [4], (iii) Node.js, and (iv) NestJS.

## 4 Case Study

The development of the open-source web system is based on the OSCRUM methodology. This methodology is very similar to the SCRUM methodology, which has been used in other projects for the health area [7]. The activities to apply the OSCRUM methodology in developing the All Vaccinated open-source project are detailed below.

- **Problem Discovery and Search for Volunteers.** The OSCRUM development methodology allowed the researchers to propose a starting point for the development of the open-source system, so the first step was to discover the problem and look for volunteers to develop a solution [6]. For this reason, a web artifact was created for this activity: a blog. The link to the blog is the following: <https://allvaccinated.blogspot.com>, and the main idea for developing the software project has been published in it. This blog was created to collect information from the open-source community (regular users of open-source tools registered in the community), such as their names and e-mails. In addition, this web artifact (blog) was used as a communication channel between the members of the development team and the community using this system. The All Vaccinated blog is presented in Fig. 1.
- **Communication.** Undoubtedly an essential aspect in the development of any project is communication, and using the OSCRUM methodology [6], a massive invitation was made through mailing lists and VOIP tools (Meet by Google) to an

**Table 1** OSCRUM's activities for the development of the web system [6]

No.	Criteria	Definition
1	Identification of the problem and search for volunteers	An individual or a small group of individuals, who are not necessarily the main-maintainer or core-contributors, generate an idea and description of the idea, which will be published on the web to gather contributors
2	Communication among development team members	Communicate with core-contributors to generate the feature list for the software product
3	Organization of an initial release planning meeting	Create multiple sprints: <ul style="list-style-type: none"> <li>• Product feature backlog</li> <li>• Interface design</li> <li>• Complete system</li> </ul>
4	Release plan and product status	Once the feature backlog sprint is completed, the activities are listed in the release plan, with their respective status
5	Feature update	During this sprint, if some new features are approved by the main-maintainer and core-contributors, the feature backlog will be updated
6	Source code testing	The project is published in repositories and the community tests it
7	Bug reporting	The community reports bugs found
8	Contributions from external members	Individual contributions are integrated, previously approved by the main-maintainer of the software
9	Bug fixing	Problems or bugs created are corrected by contributors
10	Approval of new features	After features are submitted and resolved, they are re-reviewed and the revision is accepted
11	Iteration of the process	The process is repeated and continued according to the product features

initial communication meeting to discuss the needs of the health area regarding the systematization of the vaccination processes, and the backlog of product features (open-source web system All Vaccinated) was generated.

- **Initial Launch Planning Meeting.** After the initial communication meeting, another meeting was held to plan the launch of the All Vaccinated web system. In this meeting, the functional and non-functional requirements for the project implementation have been discussed and detailed due to the use of information gathering techniques: JAD, Brainstorming, and Competitive Analysis. The initial requirements specification technical report is available at the following link: <https://bit.ly/3SccoLH>. The information corresponding to the application of JAD, Brainstorming, and Competitive Analysis techniques is available at the following link: <https://bit.ly/DocsOSCRUM>. It is considered that being an open-source development, the project is still under development, and new features and functionalities may emerge, which could be implemented as the main-maintainer approves them.

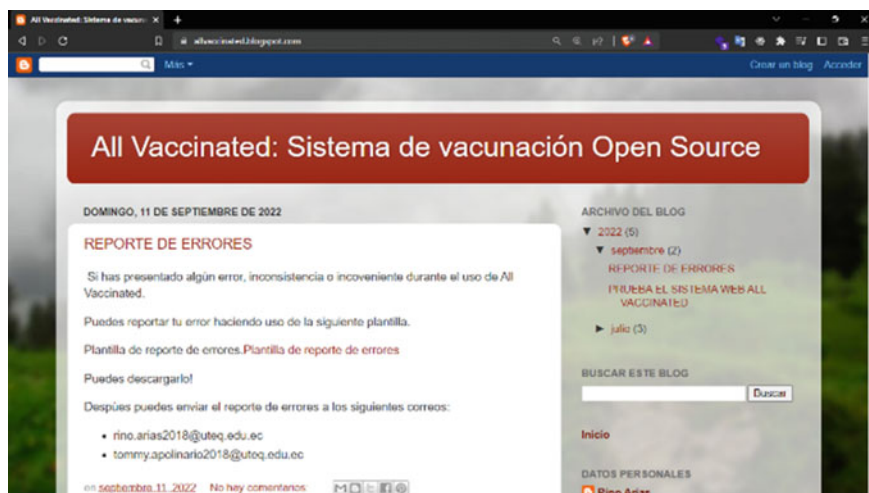


Fig. 1 All vaccinated blog

- **Launch Plan and Status.** Once the feature sprint backlog is completed, the activities are listed in the release plan with their respective status. In the blog <https://allvaccinated.blogspot.com>, a link to the system in production has been inserted through which you can test and leave suggestions and observations in the same blog. All Vaccinated have a release plan that shows all the functionalities/features of the system and the development status.
- **Feature Upgrades.** The next activity performed was to verify if new features, functionalities, or solutions were provided by the community to be subsequently approved by the main-maintainer or core-contributors. Therefore, the backlog of features of the open-source project will be updated. In our case, in the various meetings that have been held, all the project features have been taken into account, so it has not been modified.
- **Testing Source Code.** All Vaccinated is an open-source software project that is publicly and freely available for all users to use as it suits their interests. The source code is hosted in a popular repository in the developer environment called GitHub. GitHub is a cloud hosting service for depositing versions of open-source projects. The following link shows the source code of All Vaccinated: <https://bit.ly/3WF0b5U>. In this GitHub repository, a link to the web system has been created to test the source code of the All Vaccinated web system. Here users can test the web system and its functionalities and leave their comments either to improve the system or fix bugs. The link is as follows: <https://tommyapolinario.github.io/allvaccinated/>
- **Error reporting.** The open-source community (regular users of the open-source tools registered in the community) has evaluated the All Vaccinated system and determined several errors and inconsistencies in the open-source web system, applying the recommendations of the OSCRUM methodology [6]. To carry out

this activity, a template was designed and provided to the evaluators via e-mail. This template reports the problems encountered in using the open-source All Vaccinated project. In the following link, all the bugs reported by the community are listed: <https://bit.ly/3drz38b>.

- **Community Collaboration.** Independent to project member. The OSCRUM development methodology mentions that new contributors can join the project with the approval of the main-maintainer, to add new features, fix bugs, or contribute significantly to the project's development [6]. In the GitHub repository, where the source code is hosted and available, it is expected that new contributors can add solutions to the problems presented or add new features. In this case, no new developers or core-contributors have joined the project's development.
- **Repairs.** Once the errors were reported, we modified the source code to solve the reported errors, following the recommendations of the OSCRUM methodology [6]. These modifications have been uploaded through commits in the repository (GitHub) where the project is hosted. It is important to emphasize that any user can make modifications and repairs, but for them to take effect, they must be approved/validated by the main-maintainer.
- **Approval (validation).** With the repair of errors, several proposals and alternatives have been seen to solve the problems raised by modifying the source code. Once the proposed solutions are listed, the main-maintainer proceeds to review, validate and approve them for the improvement of the system, thus preventing the software project from having errors. This work rhythm was adopted thanks to the OSCRUM methodology [6].
- **Iteration.** The iterative processes of the OSCRUM development methodology [6] have been repeated several times until the functionalities of the All Vaccinated web system have been achieved.

## 5 Case Study Results

This section reports the results of the implementation of the OSCRUM methodology to develop the open-source web system All Vaccinated, as well as its implementation and source code. The results of this case study are mainly based on the functionalities and requirements established in the backlog features of the open-source web system (All Vaccinated). These features are (i) automated registration of patients, (ii) registration and control of vaccines, (iii) scheduling of vaccination appointments, and (iv) patient vaccination history, among others. Therefore, it has been possible to implement an open-source web system for vaccination control, allowing health centers to provide efficient and quality service.

The All Vaccinated open-source web system is intuitive and user-friendly; there are several types of users: (i) Patient User, who can only register, log in, and schedule an appointment; (ii) Nurse/Physician User: Has the same functionalities as the Patient user and can request the patient's vaccination history, vaccinate, and report the vaccination history (it should be noted that the Nurse/Physician user can register new

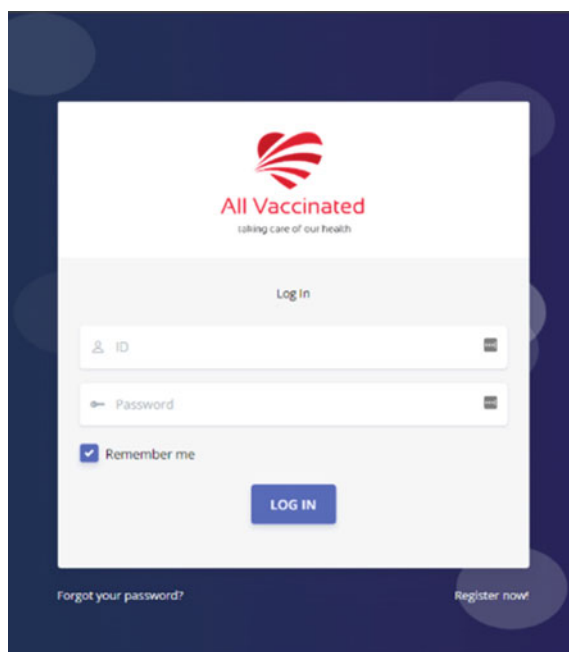
patient-type users); (iii) Administrator user, who has access to all system functionalities, including creating new Nurse/Physician type users, obtaining vaccination history reports, inventory reports, among other functionalities. Regarding the OSCRUM development methodology, the first activity concerns the discovery of the problem and the search for volunteers, together with the second activity, called communication. These two activities were carried out through video calls. In addition, the way to inform the community about the development process was through creating a web artifact (blog), resulting in a backlog of product features (All Vaccinated).

The All Vaccinated feature backlog lists the following features: (i) schedule patient appointments, (ii) register vaccinations, (iii) report vaccination history, among others. These features are detailed in previous paragraphs. Subsequently, a meeting was held to plan the launch of the All Vaccinated open-source web system, where the priorities of each of the features mentioned in the list of functionalities of the software product were specified. Developing the features to be implemented in our open-source web system was monitored. For this purpose, the development status of each of the features was detailed in a table. This table was to see which features were already developed and which were still to be developed, having as an advantage mainly an orderly software development.

Regarding updating the backlog of features, no new functionalities have been added, since no new needs arose. If necessary, the backlog of features can be updated before the main-maintainer approves it and analyzes if the feature, functionality, or set of features is scalable with the system. Some templates were made to register the error reporting, and once the community tested the web system, they could report their errors. Here one of the difficulties that have arisen and that the open-source community has emphasized is the validation of forms but not of the functionality. The next OSCRUM activity has the adoption of a new contributor to the project, where a contributor individually contributes improvements to the open-source project's source code [6]. However, this activity has not been possible to report since no contributors have come forward to contribute to the project with improvements to the source code. The bug reports from the community have allowed us to know the system's shortcomings and identify the problems that could be solved to improve the system. Therefore, once the errors were identified, these software problems were repaired, and corrective maintenance was carried out on the functionalities by writing and modifying the source code. The errors reported by the users were several, such as (i) unsecured login, (ii) error when vaccinating a patient, and (iii) error when modifying vaccines, among others.

Among the advantages of the previous activity, we can highlight the following: (i) the identification of system errors and (ii) the engagement of the open-source community. It should be noted that the corrective maintenance (solution) to these errors was carried out according to the time of the development collaborators, so it was not immediate. These corrective maintenances to the source code were performed by the project's development collaborators, so they are available in the source code. These solutions go through a process where the main-maintainer analyzes, verifies, and approves the software features before allowing changes to the source code. The last step of the OSCRUM methodology refers to iteration [6]. This process is repeated

**Fig. 2** All vaccinated web system login interface

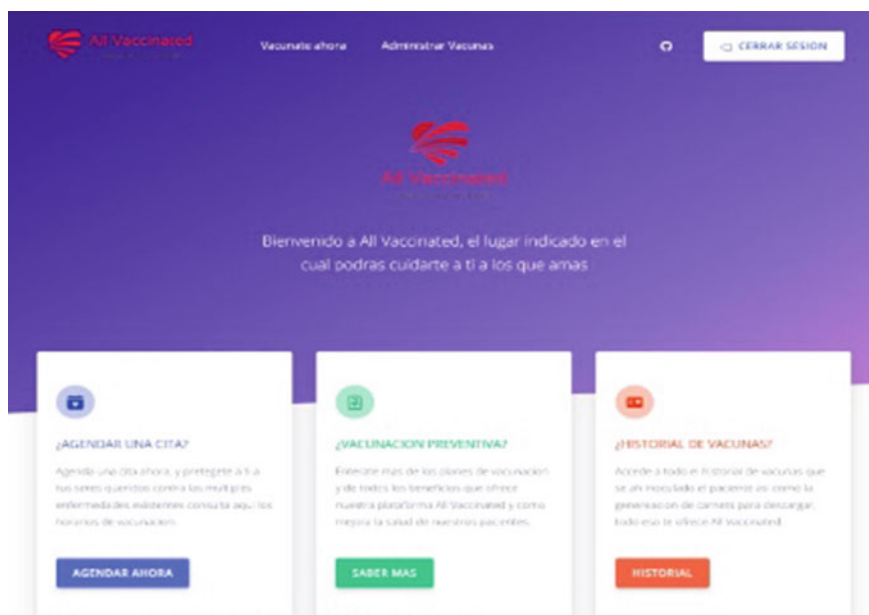


as many times as necessary to obtain new software project features and maintain the functionalities. The open-source project All Vaccinated, results in a web system that allows controlling the vaccination processes. To use the system, you can access the following link: <https://tommyapolinario.github.io/allvaccinated>. The development of the All Vaccinated web system has been completed by a series of activities described in the OSCRUM methodology framework. Figure 2 shows the final interface of the web system, specifically the login interface. Figure 3 shows the All Vaccinated web system dashboard, where the different sections are available.

The source code is hosted in a very popular repository in the developer environment called GitHub. The web system, called All Vaccinated, is available for all users to use as it suits their interests. In the following web link, you can view the source code and download it: <https://github.com/TommyApolinario/all-vaccinated-Frontend>.

## 6 Discussion of Results

The open-source web system All Vaccinated has been designed and developed so that users who want to be vaccinated against various types of viruses and diseases can do so in an efficient and orderly manner. Another objective of implementing the All Vaccinated web system is to optimize the vaccination processes used by



**Fig. 3** All vaccinated system dashboard interface

health centers to attend to the most significant number of patients in the shortest possible time. Also, a correct management of the information of each one of the vaccines, patients, users, etc. The aforementioned is reflected in the workflow of the web system to improve the vaccination processes. It is essential to mention that using a software development methodology has dramatically facilitated this project's planning, allowing for the generation of a quality system. Therefore, OSCRUM has proved to be a practical methodology for developing open-source software projects.

One of the main limitations presented during the elaboration of this research work is the lack of information regarding development methodologies for open-source software projects. Several documents have been found that do contribute significantly to the research. However, they do not go into detail on how to implement a methodology for the development of open-source projects. Despite this limitation, a software project has been successfully developed with several functionalities that contribute significantly to health centers with the processes they use to vaccinate their patients.

Regarding the case study has several limitations concerning the validity of its contributions. Regarding construct validity and internal validity, no problems have been detected. Regarding external validity, the critical limitation of our study is the number of case studies (only one OSS project). Therefore, additional case studies should be conducted to apply the methodology OSCRUM in other OSS projects. Regarding reliability, the case study should involve all types of OSS users, i.e., not only the participation of OSS users. Therefore, before applying the methodology, we

suggest looking for other options to get users who want to participate in this type of research, such as social networking.

## 7 Conclusion and Future Work

This research has aimed to develop an open-source web system using the OSCRUM methodology that facilitates health centers to implement a free vaccination process control system. The activities carried out for its implementation have been very important for the research and development of the web system, allowing us to use a methodology that has adapted very well to the needs of this project, resulting in software that meets the needs of the open-source community. We can mention the information that has been adapted for the realization of this project, such as (i) the creation of web artifacts to keep the community informed and (ii) the use of repositories to host source code.

The OSCRUM development methodology, being an agile methodology, allows easy planning and the incorporation of new functionalities or sections that can be adapted to the workflow. Using the OSCRUM development methodology has allowed the software project to be developed successfully. It is essential to mention that the software project is open source, so the planning for these projects is somewhat more informal. However, thanks to the OSCRUM development methodology, a standard has been established so that the collaborators can continuously and autonomously contribute to the project. As future work, it is proposed to evaluate the system with users and with the results obtained, make improvements that can be implemented to meet specific needs of users.

## References

1. Aizaga-Villon X et al (2022) FIWARE-based telemedicine apps modeling for patients' data management. *IEEE Eng Manag Rev* 50(2):173–188. <https://doi.org/10.1109/EMR.2022.3169991>
2. Cuesta Quintero B, Parra Valencia JA Modelo de desarrollo en proyectos de software libre y de código abierto [FOSS]: una mirada desde la teoría de la cooperación. *Ingenium* 8(20):11. <https://doi.org/10.21774/ing.v8i20.398>
3. Fossati M (2014) Todo Sobre MySQL, pp 1–240
4. Gómez Gutiérrez JA, Boada Oriols M (2018) El gran libro de Angular
5. Loza Chiriboga JS et al (2022) Pandemia en Ecuador: Aceptación de la población ante la aplicación de la vacuna contra la COVID-19. *La Cienc. al Serv. la Salud*; vol 13 Núm. Ed. Esp. Edición Espec. CIEMDO. <https://doi.org/10.47244/cssn.Vol13.IssEd.Esp..693>
6. Rahman S et al (2018) OSCRUM: a modified scrum for open source software development, pp 1–7. <https://doi.org/10.5013/IJSSST.a.19.03.20>
7. Saini V et al (2017) SEABED: an open-source software engineering case-based learning database. In: *Proceedings of the international computer software and applications conference*, vol 1, pp 426–431. <https://doi.org/10.1109/COMPSAC.2017.204>



8. Tylutki Z et al (2019) CardiacPBPK: a tool for the prediction and visualization of time-concentration profiles of drugs in heart tissue. *Comput Biol Med* 115:4–9. <https://doi.org/10.1016/j.combiomed.2019.103484>
9. Vasconcelos, LEG, Leite N, Lopes C (2018) Ballistic impact analysis using image processing techniques. *Inf Technol New Gener* 558:419–427. <https://doi.org/10.1007/978-3-319-54978-1>

# Centralized Tasks Scheduling and Load Balancing on a Cloudlet



Manoj Subhash Kakade, Anupama Karuppiah, Samarth Agarwal,  
Mudigonda Sreevastav, Obulreddigari Gayathri, V. Ranjith,  
Sista Kasi Vishwanath, and Gaurav Basu

**Abstract** Cloudlets are a new technology in IoT, and the architecture and networking of cloudlets is an emerging area of research. The basic building block of cloudlets are SoCs or powerful microcontrollers; whether it is a SoC or a microcontroller, both are severely resource-constrained. In this paper, we are looking at the design of cloudlets using Qualcomm Snapdragon 410c. Though 410c is more powerful when compared to microcontroller-based systems, it is still constrained in terms of processing power and memory. End devices requests task to be executed on the cloudlet. Multiple requests from multiple end devices may be received by the cloudlets system at a given point in time. A cloudlet system is a distributed computing system which cannot run the existing uniprocessing or multiprocessing task scheduling algorithms.

---

M. S. Kakade (✉)

Department of Electrical and Electronics Engineering, BITS Pilani, Pune Center, Pune 411021, India

e-mail: [manoj.kakade@pilani.bits-pilani.ac.in](mailto:manoj.kakade@pilani.bits-pilani.ac.in)

A. Karuppiah · S. Agarwal · M. Sreevastav · O. Gayathri · V. Ranjith · S. K. Vishwanath · G. Basu  
Department of Electrical and Electronics Engineering, BITS Pilani, K.K Birla Goa Campus, Zuarinagar, Goa 403726, India

e-mail: [anupkr@goa.bits-pilani.ac.in](mailto:anupkr@goa.bits-pilani.ac.in)

S. Agarwal

e-mail: [f20190418@goa.bits-pilani.ac.in](mailto:f20190418@goa.bits-pilani.ac.in)

M. Sreevastav

e-mail: [h20210115@goa.bits-pilani.ac.in](mailto:h20210115@goa.bits-pilani.ac.in)

O. Gayathri

e-mail: [h20210119@goa.bits-pilani.ac.in](mailto:h20210119@goa.bits-pilani.ac.in)

V. Ranjith

e-mail: [h20210113@goa.bits-pilani.ac.in](mailto:h20210113@goa.bits-pilani.ac.in)

S. K. Vishwanath

e-mail: [h20210110@goa.bits-pilani.ac.in](mailto:h20210110@goa.bits-pilani.ac.in)

G. Basu

e-mail: [f20200441@goa.bits-pilani.ac.in](mailto:f20200441@goa.bits-pilani.ac.in)

In this paper, we propose a new algorithm for distributed task scheduling with load balancing on cloudlets.

**Keywords** SoC · IoT · Cloudlets · SaaS · PaaS · IaaS · NaaS

## 1 Introduction

The cloud deployment model falls under four categories (a) Public, (b) Private, (c) Community, and (d) Hybrid [1]. Industrial IoT systems generally go in for a remote server or opt for a geographically local private cloud. This is because, in an Industrial IoT system, data privacy and security are paramount. An alternative to geographically distributed cloud systems is a privately located cloudlet that provides similar services.

Cloudlets offer several features, but they are still to be implemented in real-time as of 2022. The cloudlet paradigm may be new, but the notion of the cloudlet is not. Limitations in terms of data bandwidth and connectivity are major issues in clouds. There will always be a communication latency that will be high irrespective of the cloud service used [2].

A cloudlet is defined as “a small scale data center or a cluster of computing devices that are designed to provide cloud services to primarily mobile devices, such as smartphones, tablets and wearable devices that are in close proximity to it” [3].

In this paper, we attempt to design these cloudlets with SoCs. A Cloudlet can be defined as “a trusted cluster of nodes with resources available to use for nearby constrained devices”. Machine to machine (M2M networks) generally connects an industrial system in a hierarchical architecture, as shown in Fig. 1.

Even though end devices at the highest level of the network hierarchy run the control algorithms and do most of the computing, sometimes complex algorithms for non-linear control and overall monitoring of the industrial complex and the data collected from it are usually sent to the cloud, in case of this paper cloudlets. Since most of the complex tasks and large data are sent to the cloudlets; a major aspect of cloudlets is the scheduling of tasks in a distributed environment such that no task misses its deadline; an optimal scheduling algorithm will minimize the computation and transmission latencies.

Some analysts claim that it is better to process the data locally. The processing can take place either on the device that is sensing this data or at the edge devices [4]. Processing the data on the device will mean that the processing complexity of the end device, its memory requirements, power consumption, and form factor will increase as the analytical algorithms that process the big data collected will use Machine Learning or Deep Learning. One option is to go in for lighter ML/DL algorithms, but this is an area still in the nascent stage of research [5]. Again, while lighter models can be used for other IoT domains, it may not be possible to use these compressed models for all industrial applications. Cloudlets provide a unique opportunity and simultaneously satisfy the specification of several IoT applications that fall under the industrial domain.

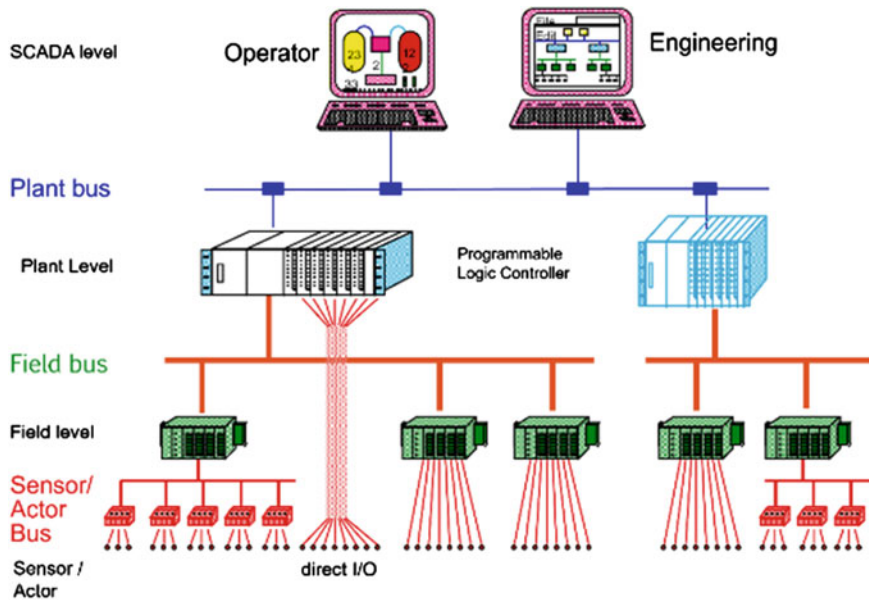


Fig. 1 Hierarchical industrial IoT network

The objective of cloudlets is to bring the processing power closer to the end devices. Tasks that are offloaded from the end devices are processed on the cloudlets.

In this paper, we are attempting to build cloudlets with SoCs. This will reduce the overall cost and increase accessibility while providing the same performance as a full-fledged server. In this paper, we suggest a centralized algorithm for task sharing and storage between various SoCs in the cloudlets.

The organization of the paper is as follows: Sect. 2 gives the background of cloud computing, Sect. 3 compares fog, cloud and edge computing, Sect. 4 describes the architecture of cloudlets, Sect. 5 gives an overview of the SoC used in implementing the cloudlets, Sect. 6 presents the proposed cloudlet architecture, Sect. 7 gives the details of the proposed algorithm for task and load balancing in cloudlet systems. Section 8 presents the results and analysis, and we conclude in Sect. 9.

## 2 Cloud Computing

Cloud computing is sometimes also termed as utility computing [6], because it can be turned off and on with minimum latency; also usage patterns can be varied dynamically according to the application required.

Considering the varying and increasing requirements of IoT applications, it is not possible to maintain the infrastructure for storing the data and computing. IoT uses various protocols for end devices of varying hardware and software architecture to

communicate with each other and with a central infrastructure (In this case, a cloud). Using APIs, we can easily connect diverse end devices to the cloud. A cloud, though defined differently in accordance with the IoT application, can be broadly termed as, “a distributed computing environment that operates over an interconnected network”. A set of computing systems distributed locally or globally functions as a platform for various end users. The services offered by the cloud could be in the form of hardware or software or specific services such as storage. Cloud computing services are available to users in various forms, such as IaaS, PaaS, and SaaS [7]. Since cloud computing has been in use for over a decade now, research in cloud systems is concentrated toward improving communication latency, offering complete services to mobile devices, and meeting real-time requirements while being a geographically distributed system. Due to the sheer geographic scale, cloud computing latencies are high in addition; there are data privacy and security issue to counter this; there is the new paradigm of cloudlets, fog and edge [8].

### **3 Edge Versus Fog Versus Cloud Computing**

Cloud computing evolved from the requirement of “computing as a utility”, offering various internet services. With the evolution of IoT, several limitations have emerged due to the centralized nature of cloud computing [9]. Hence edge computing, fog computing and cloudlet systems have emerged as new paradigms in IoT.

#### **3.1 Edge Computing**

Edge computing uses an enhanced device with more storage processing and data management capability. Edge Computing overcomes some of the shortcomings of an IoT-based system, such as (a) Connectivity, (b) Latency, and (c) Privacy [10].

This is because the edge device is close to the end device. Edge also has a higher availability of service as compared to cloud-based systems.

Figure 2 shows an IoT application using edge architecture; cloud-based IoT systems may be unable to handle real-time deadlines; in such cases the edge systems gain importance and acts as a coordinator between the end device and the cloud, and if the edge device can be made powerful enough then it can run the less latency tolerant application on it.

#### **3.2 Fog Computing**

Fog networks, by nature are heterogeneous and are latency aware. Fog is considered as an alternate implementation of an edge. This is not true. According to Cisco's

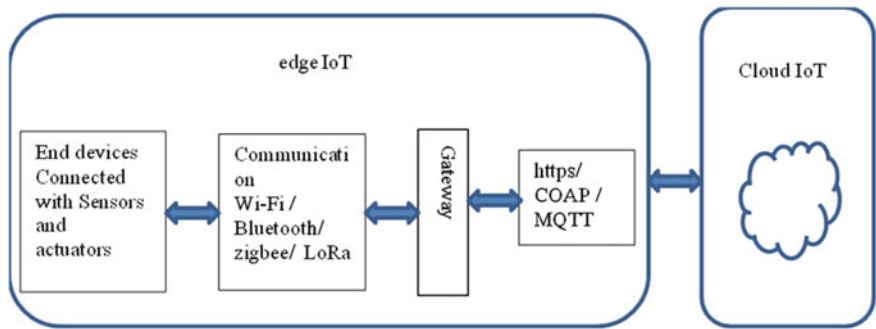


Fig. 2 IoT applications using edge architecture

definition, fog computing is “a highly virtualized platform that provides compute, storage as a network service between end devices and tradition cloud computing data centers, typically, but not exclusively located at the edge of the network” [11]. Four-tier architecture of fog computing is shown in Fig. 3.

The lowest tier usually is the end device with sensors and actuators; the intermediate layer is the edge device which is the coordinator that connects to the end device. The third layer is the fog layer; fog nodes can compute, transmit and temporarily store the data anywhere between the cloud and the end device. The highest layer is the cloud infrastructure; in the case of such an architecture, computational services can be moved from the cloud to the fog to decrement the cloud load and improve efficiency.

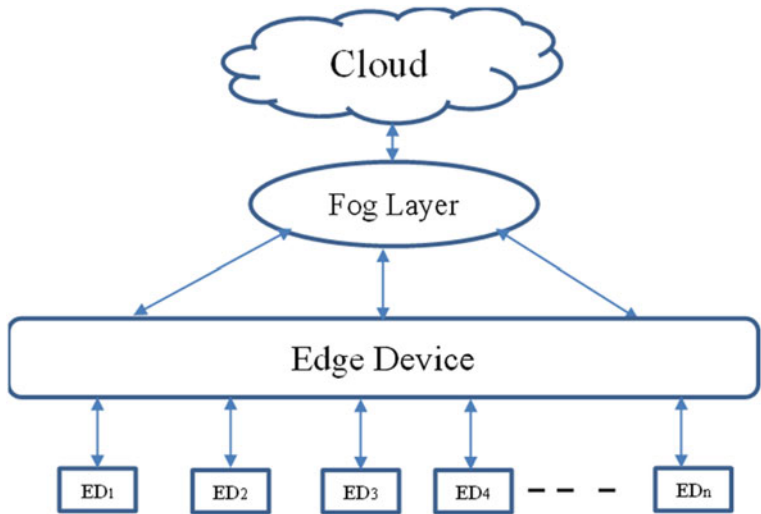


Fig. 3 Four-tier architecture of fog computing

## 4 Cloudlets

Cloudlets are similar to fog; a cloudlet is “a small-scale cloud that provides the services equivalent to that of a cloud” [12]. The primary difference is in the availability of resources—in the case of clouds, it is unlimited; in cloudlets, it is restricted. Also, the communication latency involved in clouds is much higher than in cloudlets. Most of the cloudlets are decentralized in nature; we suggest in this paper a slightly modified Cloudlet architecture, where there is a manager cloudlet node that plays a significant role when tasks need to be offloaded from the cloudlet to the cloud. Cloudlets facilitate collaborative computing and hierarchical architecture.

One way to look at cloudlet is the middle tier of the hierarchy of fog computing [13]. Cloudlet is an architectural element that realizes the convergence between cloud computing and mobile computing [14]. Cloudlets are region-specific and usually used for mobile consumers or devices; hence it is possible for a mobile device that moves from region 1 to region 2 to move from cloudlet 1 to cloudlet 2.

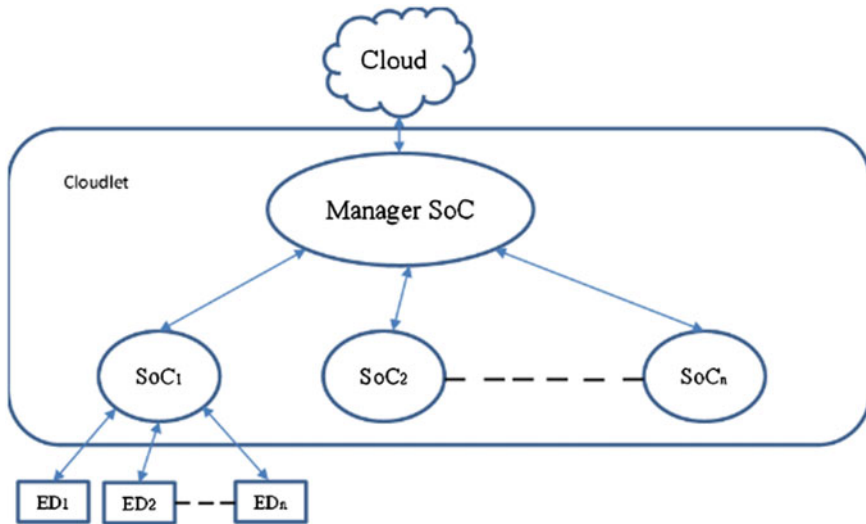
## 5 Snapdragon 410c

Snapdragon 410c is based on the APQ8016E Processor that is built around ARM cortex A-53 quad-core CPU operating at a maximum frequency of 2.4 GHz; it also has a Qualcomm Adreno 306 GPU with an OpenCL library, it has support for both wireless Ethernet as well as Bluetooth 4.0 and has an onboard GPS that can be used for location tracking. Since it has multiple communication interfaces, it is also ideal for distributed systems such as a cloudlet system. Its small form factor and its powerful computing system hence making it an ideal choice as a node in a distributed cloudlet system [15].

## 6 Proposed Architecture

The proposed system is not completely distributed, Fig. 4 depicts the cloudlet architecture as hierarchical. End devices connect to various node SoCs, and the node SoC in turn, connects to a manager SoC. No end device is connected to the manager SoC. The manager SoC interfaces to the cloud. The manager SoC is aware of the state of all the nodes in the cloudlet; the SoCs proactively send their load status, storage availability and network bandwidth at regular intervals of time “ $T$ ”. To further reduce the control overhead due to the update of status, we suggest that there we use a soft threshold ( $\delta$ ) and a hard threshold ( $\gamma$ ).

If the current CPU load on a node is  $\geq$  (previous load +  $\delta$ ), then the system status is transmitted to the manager. At any point in time, if the CPU load goes beyond  $\gamma$ , information is immediately sent to the manager. The same procedure is followed for



**Fig. 4** Proposed cloudlet architecture

storage and network bandwidth usage. The next section gives the complete details of the algorithm for task scheduling on the cloudlet for this centralized architecture.

## 7 Proposed Algorithm

Information is stored at each node in a cloudlet. Each node in the cloudlet will have an information manager with the following statistics: (a) Percentage of CPU utilization (b) Current application software in use (c) Available application software for user (d) Available data space (e) Network bandwidth available for setting up a connection.

**Algorithm:** The algorithm for task allocation in the cloudlet is a centralized one, as we are following a centralized architecture as described in the previous section. The manager SoC has information about the statistics of each SoC connected to it. The information collection process is proactive except when the manager SoC may need to query a node about its instant statistics when the load distribution is non-uniform. The query may be raised by the manager when one or more SoCs are heavily loaded or have less storage space available. Any node can exist in three states heavy, medium, and light.

These states are defined with respect to the following:

- a. CPU load
- b. Data storage availability



c. Network bandwidth.

a. **CPU Load**

**Heavy state:** A node in the cloudlet enters a heavy state when it is running application software or multiple instances of application software. If the CPU utilization percentage is greater than or equal to  $CPU_H$  the node stops accepting requests for running the application. Tasks are migrated to other nodes through the manager SoC.

**Medium state:** If the CPU utilization percentage is greater than that of equal to  $CPU_M$  but lesser than  $CPU_H$  the SOC is in the medium load stage. It can accept application tasks until its load is less than  $CPU_H$ .

**Light state:** If the CPU utilization percentage is less than  $CPU_M$  then it is in a light state. In such a state, the Manger SoC can easily migrate the task from a heavily loaded SoC to this SoC, provided lightly loaded SoC has an affinity toward the task if there is sufficient storage and the network bandwidth and is in the medium state or lesser.

b. **Data storage availability:** Intermediate data storage is done on the cloudlets before a complete backup needs to be done. The amount of storage space can fall into any of the three states heavy, medium, and light.

**Heavy:** if the amount of data storage available on the node is lesser than  $DATA_H$  MB then it indicates high storage utilization; hence any data arriving that is not related to the application currently under execution is to be migrated to the other SoCs in the Cloudlet. Acceptance of new tasks also depends upon storage availability. So while the SoC is in a data storage heavy state, even if the CPU load is medium, it will not accept new application tasks, a condition where the storage utilization is high but CPU utilization is light will rarely occur.

**Medium:** If the amount of data storage space available is greater than or equal to  $DATA_M$  MB but lesser than  $DATA_H$  MB, then it is available for data storage. It can also accept new application tasks if the data space required will not cause the device to go into Data heavy state.

**Light:** If the Data storage utilization is less than  $DATA_M$  MB then it is in a light state. It can accept new applications as well as data. It can accept tasks migrated to it by the manager SoC provided that there is sufficient network bandwidth.

c. **Network bandwidth:** The knowledge of network bandwidth is necessary; for example, if a connection cannot be established due to heavy usage of network bandwidth, the task or data cannot be migrated to a Light state SoC. Network Bandwidth: network utilization might be high, medium, or low based on the number of active connections and available network bandwidth.

### Algorithm

1. At regular intervals of time “ $T$ ” sec every node in the cloudlet system send its states (CPU, DATA and network) to the manager
2. When any new application instance arrives at node
  - a. It checks its state and data storage
  - b. If the state and CPU Utilization and data storage are at medium or light
    - i. If the application instance can be run, it then accepts the task and the related data.
    - ii. Else the task is sent to the manager SoC.
    - iii. Manager SoC finds a node with a medium or light state. Manager Node then decides which node to migrate to by using the following information: Number of Tasks completed in the past window (TP). The number of tasks in the current execution window is (TC), and the number of tasks (TF) in a chosen length of the queue is ready for execution.

$$N = w1 * \frac{TP}{NP} + w2 * \frac{TC}{NC} + w3 * \frac{TF}{NF} \quad (1)$$

Here NP is the Time window of the past tasks, NC is the Time window of the current tasks, and NF is the Time window of the future tasks and  $w1, w2, w3$  are the weights assigned.

- iv. If  $N < Th$  (Threshold set by system), then the task is migrated to the node with the least  $N$  value.
- v. If no node with the value  $N < Th$  is available, task is sent to the cloud.
3. When any new data arrives at a node
  - a. It checks its data storage availability
  - b. If the data storage is at medium or light
    - i. It then accepts the data.
    - ii. Else the data is sent to the manager SoC.
    - iii. The manager SoC looks at the storage utilization rate per node marked as light and medium. If the rate of utilization is  $< R_{TH}$  and if the number of tasks in the NF window of the queue is  $< R_{TH}$  then the data is sent to nodes with light or medium data state and with a minimum rate of storage utilization.
    - iv. If no such node is available, data is sent to the cloud.

## 8 Results and Discussion

To analyze the protocol, we ran it on five Qualcomm Snapdragon 410c. One 410c acts as the manager, and the other four as normal cloudlet nodes. We analyzed the behavior of the Task Scheduling algorithm on the cloudlets by varying parameters such as the

weight, load on the node, percentage of node heavily loaded, and window size. We have assumed a medium data load on all the nodes. Our earlier publication [16] describes the effect of data loading on the available memory on Snapdragon 410c. In this case when we are talking about tasks executed, we mean tasks completed before the deadline.

**Varying Weights:** We ran the task scheduling algorithm for the following (a) varying  $w_1$  between 0.333 and 0.8 and distributing the  $(1 - w_1)$  equally between the weights  $w_2$  and  $w_3$ , (b) varying  $w_2$  between 0.333 and 0.8 and distributing the  $(1 - w_2)$  equally between the weights  $w_1$  and  $w_3$ , (c) varying  $w_3$  between 0.333 and 0.8 and distributing the  $(1 - w_3)$  equally between the weights  $w_1$  and  $w_2$ . The results can be seen in Figs. 5, 6, and 7. This was run keeping the number of past, future tasks windows same at 10-time units. The number of current tasks window was also kept at 10 time units.

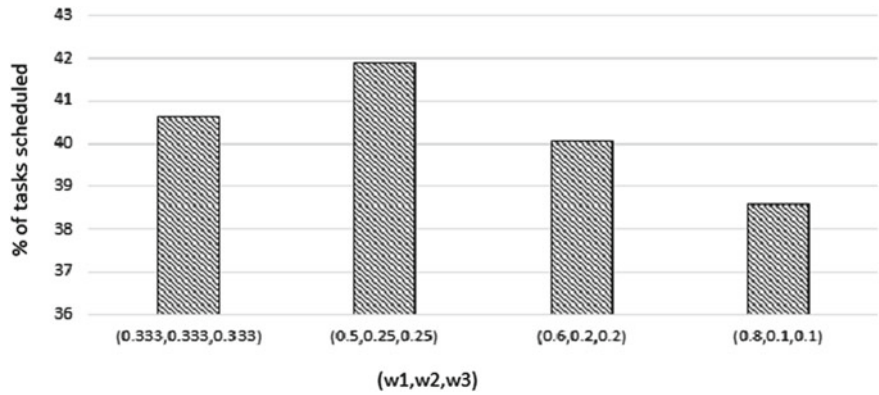


Fig. 5 Task scheduling versus varying  $w_1$  keeping  $w_2, w_3$  equal

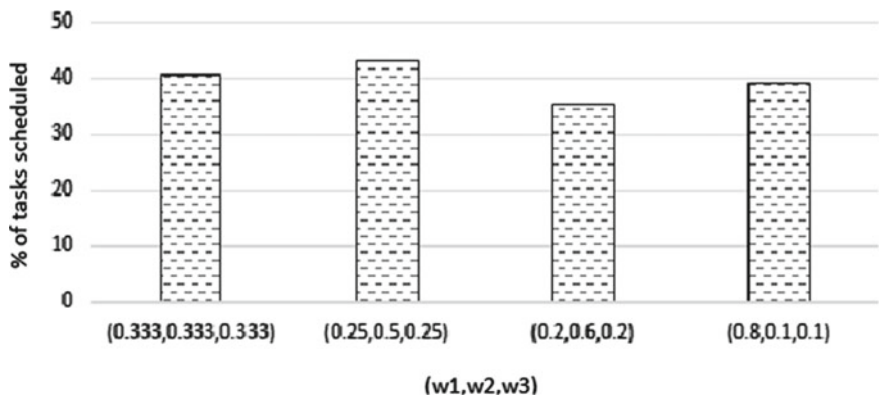


Fig. 6 Task scheduling versus varying  $w_2$  keeping  $w_1, w_3$  same

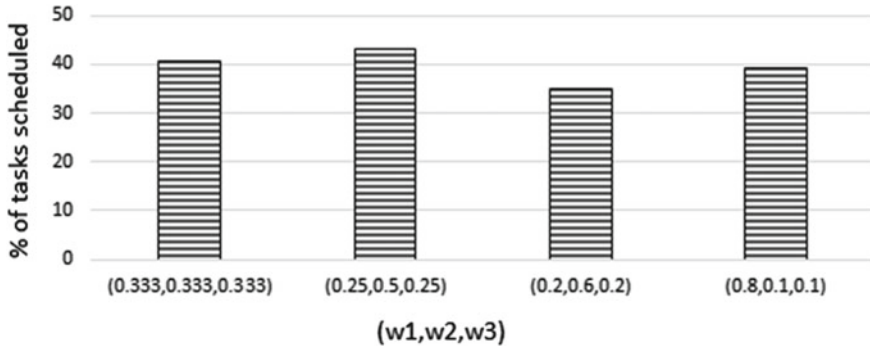


Fig. 7 Task scheduling versus varying  $w_3$  keeping  $w_1$ ,  $w_2$  same

From the figures, it can be seen that the best result is at 0.5 for all  $w_1$ ,  $w_2$  and  $w_3$ . But this is not possible. The best possible solution is keeping the present at 0.4 with the past and future windows at 0.3. This is what we used in the further analysis.

**Varying the High-Level Threshold and Medium Threshold:** We varied the high-level Threshold between CPU being 60% (0.6) loaded and 90% (0.9) loaded in steps of 10% (0.1). We also varied the number of nodes that were heavily loaded from one to four (all the nodes). The results are shown in Fig. 8. From Fig. 8, we can see that we get the best results when we have just one node heavily loaded with a high threshold at 90%. When the high threshold is at 60%, the behavior is better than at 80% because the migrated tasks can be scheduled, but at 80%, neither can the task be scheduled at the node, nor can it be migrated as the other nodes are already working at 80% capacity (Fig. 9).

A similar trend can be observed in the case when the medium load is varied along with the number of nodes that are being mediumly loaded. The best results are better for  $n = 1$ , as expected.

**Varying the Size of Future and Past Windows:** We varied the size of the past window from 10 to 20 time units keeping the future and current window size constant at 10 time units. We also varied the size of future window from 10 to 20, keeping the past and current window size constant at 10 time units. The result can be seen in Figs. 10 and 11.

From the figures, we can see that the more we look at the history, the performance drops. If we have more details of the future tasks, this helps the scheduling, and hence the performance improves. Knowledge of future tasks will help choose a node that may also be lightly loaded in the future; hence, tasks are more schedulable and completed well before the deadline.

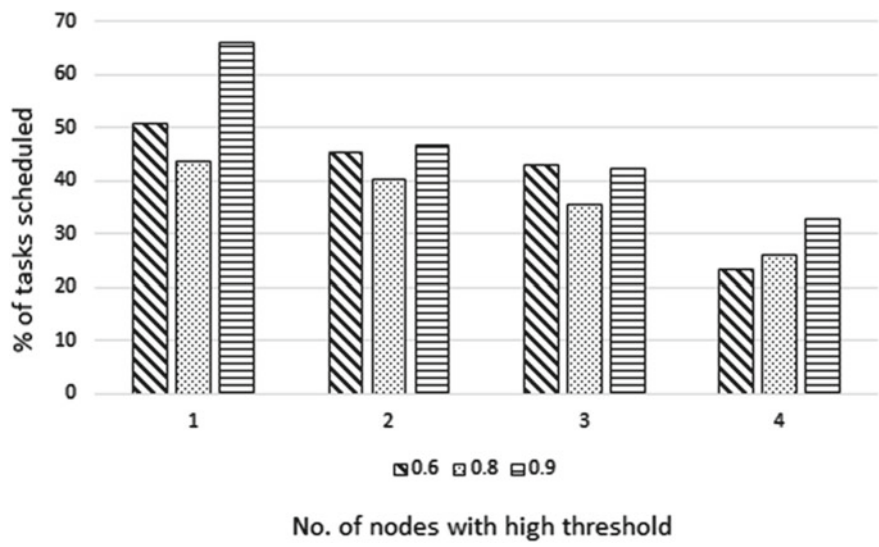


Fig. 8 Varying number of highly loaded nodes versus % of tasks executed

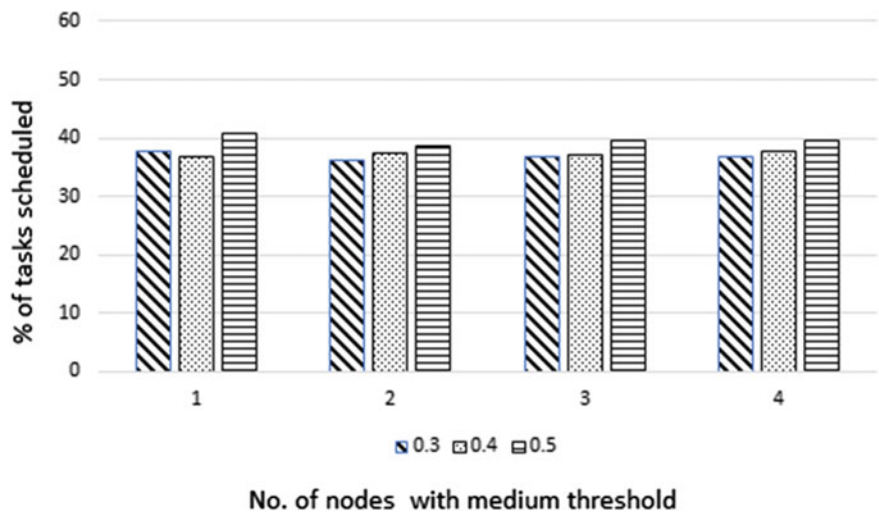


Fig. 9 Varying number of mediumly loaded nodes versus % of tasks executed

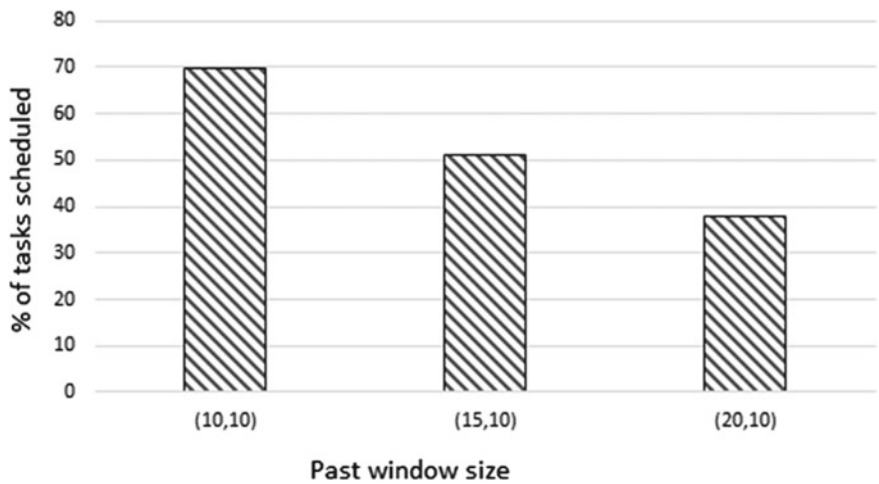


Fig. 10 Varying past window size versus % of tasks executed

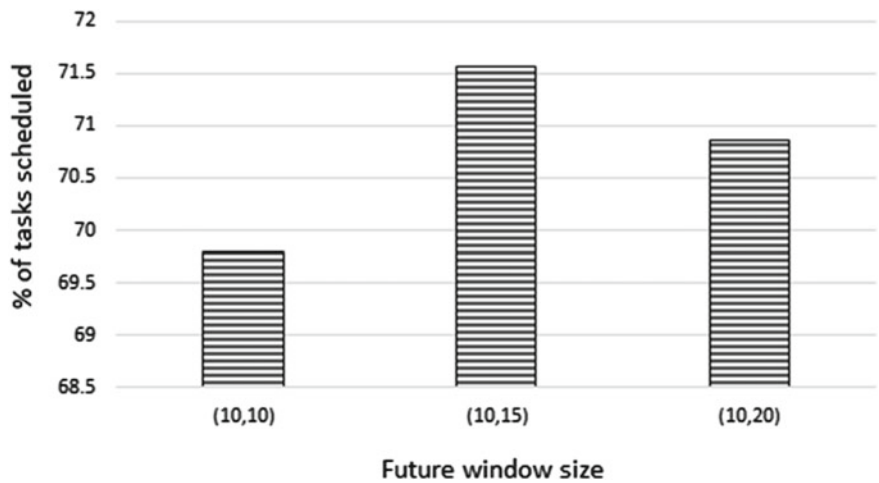


Fig. 11 Varying future window size versus % of tasks executed

9 Conclusion

In this paper, we presented a centralized multi-tasking scheduling algorithm that is extremely simple and can be easily deployed in distributed systems, especially in the case of cloudlets. The percentage of the tasks schedulable is at most 75%, this is fine as we have not executed any tasks on the manager, so if the tasks can be executed on the manager, the percentage of tasks is 91.2%. The rest of the 8.8% can be easily moved to the cloud. Cloudlets being close to the end devices, will suffer

from minimum latency. We can also conclude from our analysis that the weights assigned to history, current and future tasks must be the same. Also, as we increase the high threshold, most tasks get executed at their own node, and the performance improves. When we are able to get a more clairvoyant understanding of the future tasks on the system, the performance improves.

In this paper, we have presented a load-balanced task-sharing algorithm for scheduling tasks on cloudlets. Individual cloudlet nodes were implemented on Snapdragon 410c. We have currently built a cloudlet on four 410cs. The architecture of the cloudlet presented uses a central manager with individual cloudlet nodes. In the future, we plan to make the architecture completely distributed and have a hybrid of multiple SoCs, such as Snapdragon 820c and 833.

## References

1. Goyal S (2014) Public vs private vs hybrid vs community—cloud computing: a critical review. *Int J Comput Netw Inf Secur* 3:20–29
2. Dolui K, Datta SK (2017) Comparison of edge computing implementations: fog computing, cloudlet and mobile edge computing. In: *Proc. Global Internet Things Summit (GIoTS)*, Jun 2017, pp 1–6
3. Satyanarayanan M, Bahl P, Cáceres R, Davies N, University L (2009) The case for VM-based cloudlets in mobile computing. *IEEE Perv Comput* 8 (4)
4. Li C, Xue Y, Wang J, Zhang W, Li T (2018) Edge-oriented computing paradigms: a survey on architecture design and system management. *ACM Comput Surv* 51(2):A34–A39
5. Al-Garadi MA, Mohamed A, Al-Ali AK, Du X, Ali I, Guizani M (2020) A survey of machine and deep learning methods for internet of things (IoT) security. *IEEE Commun Surv Tutor* 22(3):1646–1685
6. Miller M (2008) *Cloud computing: web-based applications that change the way you work and collaborate online*. Que Publishing, Indianapolis
7. Dukaric R, Juric MB (2013) Towards a unified taxonomy and architecture of Cloud frameworks. *Future Gener Comput Syst* 29(5):1196–1210
8. Hamdan S, Ayyash M, Almajali S (2020) Edge-computing architectures for internet of things applications: a survey. *Sensors* 20(22):6441
9. De La Prieta F, Corchado JM (2016) Cloud computing and multiagent systems, a promising relationship. Springer, Cham, pp 143–161
10. Zhang J, Chen B, Zhao Y, Cheng X, Hu F (2018) Data security and privacy preserving in edge computing paradigm: survey and open issues. *IEEE Access* 6:18209–18237
11. Open Fog Consortium (2017) Open fog reference architecture for fog computing [Online]. Available from: <https://www.openfogconsortium.org/ra/>. Feb 2017
12. Satyanarayanan M, Chen Z, Ha K, Hu W, Richter W, Pillai P (2014) Cloudlets: at the leading edge of mobile-cloud convergence. In: *Proceedings of the 6th international conference on mobile computing, application and services*, 2014, pp 1–9
13. Varshney P, Simmhan Y (2017) Demystifying fog computing: characterizing architectures, applications and abstractions. In: *IEEE 1st international conference on fog and edge computing (ICFEC'17)*, pp 115–124
14. Pang Z, Sun L, Wang Z, Tian E, Yang S (2016) A survey of cloudlet based mobile computing. In: *International conference on cloud computing and big data*, pp 268–275
15. Qualcomm snapdragon 410c. Available from: <https://www.qualcomm.com/products/technology/processors/application-processors/dragonboard-410c>

16. Kakade MS, Karuppiah A, Mathur M, Parikh P, Dhir R, Gokhale T (2022) Tasks scheduling and load balancing on a cloudlet system using Qualcomm 410c. In: Proceedings of the 26th world multi-conference on systemics, cybernetics and informatics (WMSCI 2022)



# A Digital Twin Enabled Decision Support Framework for Ship Operational Optimisation Towards Decarbonisation



Antonis Antonopoulos, Bill Karakostas, Takis Katsoulakos,  
Anargyros Mavrakos, Theodosis Tsaousis, and Stathis Zavvos

**Abstract** Management of the combined effects of several factors is needed to achieve the required ship operational performance towards emissions reduction ('decarbonisation') for green shipping. Identifying these factors, defining the effects of them on each other, assessing their importance, and selecting decarbonisation solutions, require a suitable management framework. This paper discusses the potential of the recent IT paradigm of digital twins for the optimisation of ship performance, regarding decarbonisation as the ultimate goal. The management framework described in this paper is underpinned by a data-driven digital twinning platform and assists stakeholders to continuously optimise current ship operations as well as evolve the next generation of energy efficient ships.

**Keywords** Ship decarbonisation · Digital twin · Knowledge Graph · Simulation

---

A. Antonopoulos  
Konnecta, Newbridge, Ireland  
e-mail: [Antonis.Antonopoulos@konnecta.io](mailto:Antonis.Antonopoulos@konnecta.io)

B. Karakostas (✉) · T. Katsoulakos · A. Mavrakos · T. Tsaousis  
Inlecom Systems, Brussels, Belgium  
e-mail: [bill.karakostas@inlecomsystems.com](mailto:bill.karakostas@inlecomsystems.com)

T. Katsoulakos  
e-mail: [takis.katsoulakos@inlecomsystems.com](mailto:takis.katsoulakos@inlecomsystems.com)

A. Mavrakos  
e-mail: [argyris.mavrakos@inlecomsystems.com](mailto:argyris.mavrakos@inlecomsystems.com)

T. Tsaousis  
e-mail: [theodosis.tsaousis@inlecomsystems.com](mailto:theodosis.tsaousis@inlecomsystems.com)

S. Zavvos  
VLTN, Antwerp, Belgium  
e-mail: [stathis.zavvos@vltan.be](mailto:stathis.zavvos@vltan.be)

# 1 Introduction

## 1.1 Background

Maritime, while being one of the greenest modes of transport, still has significant scope for decarbonisation. The European Green Deal [1], the Paris Agreement Objectives [2], the Initial International Maritime Organisation (IMO) Strategy on the reduction of GHG emissions from ships [3] and the CCNR Ministerial Mannheim Declaration, represent key policy developments which provide clear ambitions towards zero-emission waterborne transport by 2050. The vision is to develop low-emission solutions for all main ship types and associated shipping services by 2030, in turn enabling shipping companies to implement strategies for achieving the IMO requirements of reducing carbon compound emissions in shipping by at least 40% by 2030, and by 70% by 2050. This is consistent with achieving the EU goal for zero-emission waterborne transport by 2050. Therefore, vessels, with operational zero emissions, would need to represent a significant portion of newbuilds from 2050 onwards [4].

As part of greenhouse gas emission reduction strategy, IMO has initiated several reporting measures for the shipping industry, including the Energy Efficiency Design Index (EEDI), the Ship Energy Efficiency Management Plan (SEEMP), the Energy Efficiency Operational Indicator (EEOI) and the Energy Efficiency Existing Ship Index (EEXI) [5]. A key shipping industry requirement towards decarbonisation, therefore, is the development of advanced decision support systems to assist ship operators in quantifying the effect of the different decarbonisation pathways on the vessel's emission profile and technical and economic performance improvement. The use of digitalisation to accelerate decarbonisation in the shipping industry, supported by data gathered from sensors and ship systems [6], is a promising approach as it provides many economic, health, social, and environmental benefits [7]. Particularly relevant is Digital Twin (DT) technology, initially used by NASA, which achieved prominence in recent years across different domains (e.g. Industry 4.0, Smart Manufacturing, transport, and Smart Cities) for enhanced decision making, particularly in designing and managing complex systems. In short, DTs enable the creation of a digital representation of a ship, which is fed with data acquired by the physical ship via sensors. These digital representations can then be used to optimise and/or simulate various ship-related functions and processes, and the results can then be used to enact appropriate actions on the physical ship either manually or automatically. The complexity of managing a ship energy system supporting optimally propulsion and cargo operations in a dynamic environment, makes it is easy to see how digital twins can underpin decision support systems for decarbonisation and performance improvement in the shipping sector.

## **1.2 Research Contribution**

To maximise the ship operational overall energy efficiency to comply with required CO<sub>2</sub> targets, it is necessary to select, deploy and manage decarbonisation solutions as part of the overall ship optimisation. This is something not considered by ship design methodologies.

This paper presents research carried out towards optimisation of ship performance that includes emissions reduction ('decarbonisation'), by the EU funded DT4GS project[8]. The paper outlines a management decision support framework underpinned by a data-driven digital twinning platform that helps stakeholders develop, maintain and evolve the ship digital twin to continuously optimise current ship operations as well as evolve the next generation of energy efficient ships. As part of the proposed framework, a novel ship performance control approach, supported by a knowledge graph interlinking variables that affect ship performance and performance metrics is also presented in the paper.

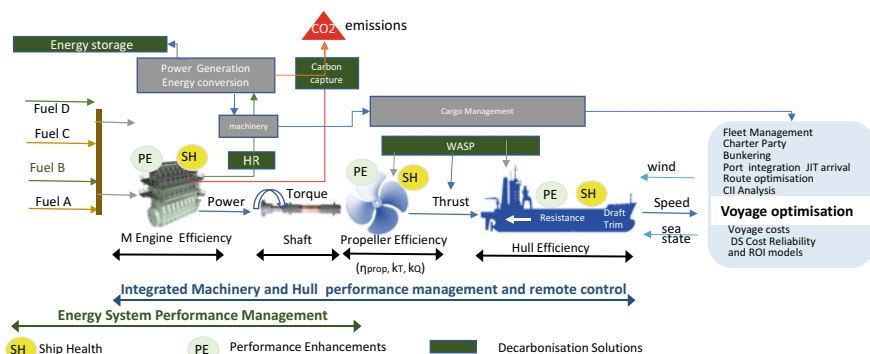
## **1.3 Organisation of the Paper**

The paper is organised as follows: Sect. 2 contains a discussion of digital twins in the context of green shipping. Section 3 presents the management framework for ship operations optimisation, including an overview of the key components of the framework. In the same section, a novel control approach is also introduced, that optimises ship operational parameters in line with given environmental and economic key performance indicators (KPIs). Section 4 contains a discussion of the significance of the presented framework and of its benefits for the shipping industry. Finally, Sect. 5 contains a discussion of recommendations for shipping managers and policy makers regarding the utilisation of the paper's framework and of digital twinning more generally, towards transition to ship decarbonisation. The benefits and pitfalls of digital twinning for shipping are discussed. Obstacles to adoption of green shipping technologies as well as information technologies that can act as enablers of ship decarbonisation are outlined.

# **2 Ship Digital Twin**

## **2.1 Digital Twins**

A ship digital twin is a digital replica of the real (physical) ship in terms of its structure (e.g. hull type, component layout, hull parameters), subsystems (e.g. engines, propeller, rudder) and functions (e.g. propulsion, navigation, loading), as well as its integration in a fleet management system or multimodal supply chains. It provides a



**Fig. 1** Ship modelling domains for digital twin

unique, intelligent ship model, merging technical specifications, component models, and parameters with management information on the components and processes of the ship, ultimately enabling computerised simulation and optimisation of all its functions in terms of performance metrics that address environmental criteria.

Digital Twins use bi-directional communication links with the automation systems and network infrastructure on a physical ship [9]. The communication link from the ship to the Digital Twin is used to monitor the physical ship constantly through several data collection techniques and devices, and communication channels to transfer this information. This allows the virtual Digital Twin to constantly learn from its physical counterpart and evolve, mirroring its lifecycle. As a result, a Digital Twin can be used to get insights into the current state of the physical ship as well as predict future states of the physical ship through simulation or predictive algorithms. The Digital Twin can, therefore, drive automated supervisory control of the ship or inform human decision makers who are required to perform the necessary tasks, such as ship operators and maintenance engineers.

A ship digital twin, thus, consists of ship model parameters (and their values) that we can observe and measure. It can be developed as a series of interlinked/interconnected models for the ship propulsion system, the ship voyage management or fleet management, and can address the entire continuum of ship design, construction and lifecycle management. In our approach this is achieved via a Knowledge Graph that interlinks the different ship optimisation and prediction models that correspond to different ship subsystems. Figure 1 shows the different ship subsystems that can be modelled with a digital twin.

## 2.2 Approaches to Developing and Using Digital Twins

There are three main approaches to developing ship digital twins: ‘White box’ approaches rely on creating analytical models of the ship that are based on the

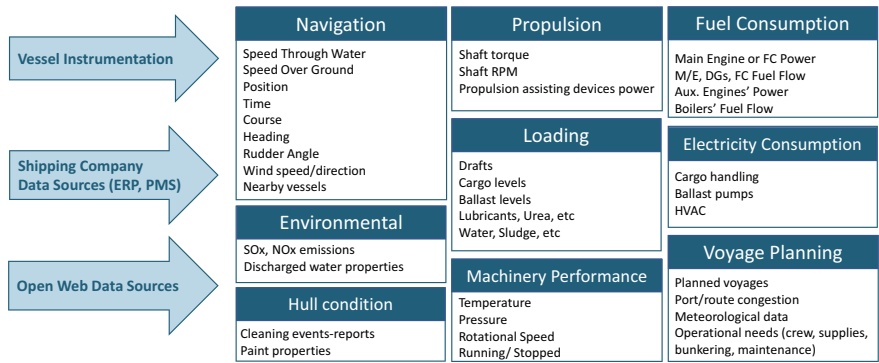


Fig. 2 Sources of ship data for the digital twin

underlying laws of physics. In contrast to white box methods, ‘black box’ approaches utilise statistical techniques to derive the relationships between the digital twin data, without reliance on a prior model of each process that generates them. Hybrid (or ‘grey box’) models, combine the benefits of the white box and black box approaches, meaning that the physical laws governing the relationships between the ship variables (white box) can be used where feasible, together with statistical approaches (black box), in order to decrease the error margin and give better prediction accuracy with reasonable computational times.

As per Fig. 2, there are various sources of data that can be used to populate a ship’s digital twin. These include instrumentation on-board the ship, IT systems of the company managing the ship and external/third party information sources.

2.3 How Digital Twinning Relates to Ship Decarbonisation

Understanding the environmental performance (e.g. fuel consumption, fuel burn by-products) of a ship is a data intensive process. However, most such data typically rely on service providers’ or manufacturers’ data and tests that are carried out under restricted and limited conditions, with the resulting lack of transparency leading to significant uncertainty on the actual energy savings in real life conditions. Additionally, instrumentation installed on board may not offer the required resolution or accuracy to provide reliable results. In addition, some calculations rely on measurement manually conducted by crew and may have unknown accuracy. Moreover, performance data may originate from studies that have no relevance or similarity to the actual operational usage and profile of a particular ship. To develop robust and accurate ship environmental and economic performance models, detailed ship technical/operational data are necessary. Performance models must also be customisable and adaptable to the actual specificities of individual ships and operating patterns, in

order to enable real-time operational optimisation, maintenance triggers and evaluate technological interventions [10].

### 3 Ship Decarbonisation Decision Support Framework

#### 3.1 Framework Overview

The decision support framework presented in this section supports decision makers in planning, implementing and controlling decarbonisation solutions on-board ships while monitoring their performance and planning their evolution (upgrading, replacing, retrofitting). The Ship Integrated Performance Management Metamodel, shown in Fig. 3, integrates the viewpoints and key decision parameters of the operational performance framework.

The Ship Performance Management metamodel consists of subdomains corresponding to the ship subsystems (cargo, voyage, propulsion and power), the ship’s operational requirements and environmental conditions, the optimisation targets set by the stakeholders (including environmental as well as financial targets), and the decarbonisation technologies (currently installed or planned for installation).

More specifically, the performance subdomains considered are as follows:

- Propulsion Performance containing main engine, shafting, hull, propeller, rudder, and propulsion performance enhancement devices;
- Cargo Management focusing on loading and unloading processes to optimise energy consumption for running pumps and/or cargo equipment.
- Voyage Optimisation (VO) against economic, operational and environmental criteria (route, speed, trim, draft, Just in Time (JIT) arrivals, weather) using real-time predictive wind and wave energy spectra analysis, just-in-time port arrival

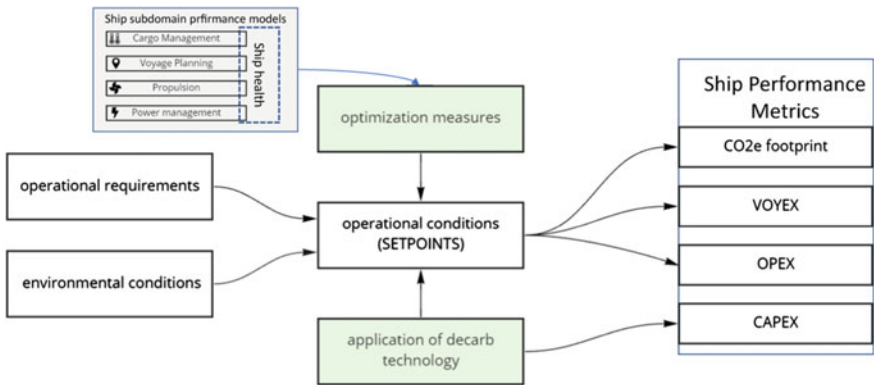


Fig. 3 Integrated ship performance management framework

scenarios and interaction with Fleet Management, Charter Party, Bunkering, Carbon Intensity Indicator (CII) Analysis. The VO models need to be able to utilise a variety of inputs such as location, destination, wave and current direction and intensity, fuel cost, local weather information, and other ship and voyage-specific parameters to determine the optimal route for the ship, and optimal engine, trim and rudder settings at regular intervals.

- Power Management models for integrated ship energy production, distribution and optimisation incorporating optimisation models to configure the Heat Recovery (HR) system, and simulation models to determine the effects of implementing additional HR Units.
- Decision support tools for valuation, selection and deployment of suitable decarbonisation solutions.
- Ship health monitoring and prediction models, utilising sensors, supporting accurate estimation of ship condition and updating of maintenance plans.

The parameters within a subdomain and across subdomains are related via causal relations that constitute an overall ship performance Knowledge Graph. The parameters are defined in terms of mathematical equations, data tables or other mathematical models and formulae. The values of output parameters related to the ship subdomains are obtained through the ship's digital twin, using a combination of white box models, manufacturer data sheets or black box techniques (for instance machine learning) that act on the digital twin provided data.

### ***3.2 Application of the Management Framework***

The above-described ship framework is used for the optimisation of the ship's performance [11], including environmental performance, by optimisation trade-offs between the parameters of different subdomains. This constitutes a complex multi-faceted problem, particularly in view of the dynamic nature of the environment of a ship's operation, as well as the changing landscape of decarbonisation solutions.

In our approach, optimisation refers to finding the operational conditions, including use of decarbonisation technologies as appropriate, that optimise a set of performance metrics, including CO<sub>2</sub> g/ton-mile function, as well as financial parameters such as VOYEX, OPEX and CAPEX.

As an example, the dependency between engine condition and hull efficiency has a strong bearing on the optimum ship speed for reduced fuel consumption and therefore CO<sub>2</sub>. This is, however, additionally subject to wind and sea conditions (e.g. wind speed and direction, as well as current strength and direction). Further, if a decarbonisation solution is installed such as wind assist propulsion, the operational optimisation of such a device will depend on predicting its effect on the ship hydrodynamics, and on controlling parameters such as sail angle of attack, accordingly. Additional ship-specific hydrodynamic enhancements may be installed, which in turn further complicate the matter of determining the effects of the sail.

Thus, environmental conditions exert an influence on the ship's resistance and therefore on the ship's power requirements through complex interactions with other ship subsystems. Other conditions such as the degradation of ship's subsystems are inextricably linked to performance in a physical manner that is not immediately clear, due to the often nonlinear relationships between the various elements and components. The above-described performance management framework addresses the entire lifecycle of the ship, i.e. from requirements specification and design to decommissioning. Concerning ship operations, real-time optimisation of performance can be achieved by altering controllable variables such as trim, ballast, speed or time between maintenance events according to the uncontrollable conditions, either environmental, economic or both. Weather routing takes advantage of weather and currents in order to optimise voyage distance or time travelled and thereby minimise voyage costs (or maximise profits) and maximise safety. Additional attributes of voyage planning susceptible to optimisation, include planning bunkering, supplying, and crewing frequency and location taking into account operational and economic factors as well as green fuels availability.

## 4 Discussion

### ***4.1 Expected Benefits of the Ship Decarbonisation Management Approach***

Overall, digital twinning helps to manage complexity when dealing with the interdependencies between technical, operations, chartering, safety, regulations perspectives that affect all performance metrics for ship operators. Large reductions in ship operating expenditure, up to 40%, are possible and reduced port time up to 30% due to improved streamlined ship operations, supported by reliable and timely digital twin data. Equally important, the enhanced shipping company capabilities are expected to produce increased volumes handled, revenue and profitability. Shipbuilding costs are expected to decrease by 15–20%.

Further, potential advantages of digital twinning include:

- Respond immediately to external changes, e.g. environment, fuel prices, competition, etc.
- Anticipate and prevent problems (ship preventive maintenance)
- Save time and money in simulation, testing, analysis
- Improve the vessel's performance, compliance and sustainability, having access to reliable information of technical characteristics of all components
- Manage ship-related documentation (certification, operating manuals, tech. documents)
- Link and synchronisation between the ship owners and customer (cargo owners) document management system, integrated with real-time process data.



## **4.2 *Costs Versus Benefits of the Proposed Approach***

Creating and maintaining digital twins can be an expensive endeavour. Suitable computer infrastructure (sensors, databases and platforms) is required to collect and integrate the data models and data that constitute the digital twin. Hence, it is important to weigh the digital twinning benefits against costs. Digital Twinning adoption will likely be dependent on establishing trusted and convincing Digital Twin ship applications based on cross-domain modelling standards and innovative architectures, ensuring that ship operators and other industry stakeholders can set up their own ship-specific digital twins, leveraging their own business models and building their own confidential knowledge, at a reasonable cost.

## **5 Conclusions and Recommendations**

### **5.1 *Digital Twin Framework Impact***

The proposed framework enacts the potential of digital twins for green shipping. The EU shipping sector providing a valuable and collective asset and intelligence, that not only unifies all actors to the common 2030/2040/2050 EU emissions reductions objective, but also takes the knowledge and interventions on one vessel to help inform the optimal and time sequenced interventions on similar vessels and increase investment confidence across the sector.

The impact of decarbonising the shipping sector to the total global GHG emissions using renewable energy sources, although small relative to other sectors, is nevertheless worthwhile. There is however considerable ground to cover both in the technological/research point of view and from the economic point of view in order to have a viable and efficient application of the renewable energy solutions to the decarbonisation of the entire shipping industry.

A digital twin-based system can record the current environmental performance of the ship, which then can act as a benchmark from which newly installed technology is evaluated, or from which expected performance can be predicted prior to installation/retrofitting in order to quantify cost versus benefit and the economic risk of the required investment. Post analysis of digital twins can be used for validating the claims of decarbonisation solution providers. For instance, assessing hull coating efficiency by use of digital twin can be made possible by comparing the rate of change of fuel consumption over time before and after the application of the hull coating.

From a longer term and fleet wide perspective optimising operational parameters of a heterogeneous fleet can be achieved with combination and interlinking of digital twins.

## 5.2 Further Research

A key principle underlying successful, advanced digital twinning is that there need only be strictly one digital twin for a given ship. What we learn from the DT of one ship may be transferrable to the DTs of other ships and indeed, sharing data and knowledge is a key aspect of creating a strategic industry capability to support green shipping. Future research on ship energy efficiency and emission reduction needs to be informed by advances in Big Data [12], and machine learning (ML), as these technologies can uncover patterns in the vast volumes of data collected from the ship and could potentially assist reduction of shipping emissions. Research in this paper is hoped to accelerate the trend towards shipping decarbonisation and contribute towards a global consensus and concerted action.

**Acknowledgements** Research described in this report has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101056799 (Project 'DT4GS').

## References

1. Europa web site. [https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal\\_en](https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en). Accessed 12 Jan 2022
2. Europa web site. [https://ec.europa.eu/clima/policies/international/negotiations/paris\\_en](https://ec.europa.eu/clima/policies/international/negotiations/paris_en). Accessed 12 Jan 2022
3. IMO web site. Reducing greenhouse gas emissions from ships. <https://www.imo.org/en/MediaCentre/HotTopics/Pages/Reducing-greenhouse-gas-emissions-from-ships.aspx>. Accessed 12 Jan 2022
4. Pfeifer A et al (2020) Challenges and opportunities of zero emission shipping in smart islands: a study of zero emission ferry lines. *eTransportation* 3
5. IMO web site. Energy Efficiency Measures. <https://www.imo.org/en/OurWork/Environment/Pages/Technical-and-Operational-Measures.aspx>. Accessed 12 Jan 2022
6. Agarwala P et al (2021) Using digitalisation to achieve decarbonisation in the shipping industry. *J Int Marit Saf Environ Aff Ship* 5(4) (2021)
7. Kshetri N (2021) The economics of digital twins. *IEEE Comput* 54(4):86–90
8. DT4GS Project web site. <https://dt4gs.eu/the-project/>. Accessed 12 Jan 2022
9. Tao et al (2018) Digital twin-driven product design framework. *J Prod Res*
10. Aldous LG (2015) Ship operational efficiency: performance models and uncertainty analysis. PhD thesis, UCL
11. Armstrong V (2013) Vessel optimisation for low carbon shipping. *Ocean Eng* 73(15):195–207
12. Qi Q, Tao F (2018) Digital twin and big data towards smart manufacturing and Industry 4.0: 360 degree comparison. *IEEE Access* 6:3585–3593

# Financial Sustainability of Automotive Software Compliance and Industry Quality Standards



Pavle Dakić<sup>ID</sup>, Vladimir Todorović<sup>ID</sup>, and Valentino Vranić<sup>ID</sup>

**Abstract** Achieving a high level of quality in the AV industry involves understanding the complexity of the production cycle with the constant challenges of applying appropriate standards. The need for compliance based on relevant certificates implies the creation of an environment of trust between manufacturers, component suppliers, and end customers. Research should sublimate the most current standards from various fields related to data protection, security, system sustainability, environmental management, financial sustainability, and business transparency. The success of a complex parts supply system is impossible without a stable approach to IT security. The control of production processes sometimes involves using non-standardized techniques, and one of the applicable ones is work-in-progress (WIP), which describes the use of partially finished raw materials in the stages of obtaining final products. Key points and the contribution of the current study are in getting a better overview of the situation within this area related to standards and the financial problems that can arise due to a lack of understanding of standards compliance. The main limitations within the study itself refer to the consideration of the theoretical foundations that should be used during future research related to the practical part during implementation.

**Keywords** Automotive industry and software compliance · Ems and sustainability standards · Sasb standards · Quality standards · Standards under development status · Work in progress (wip) · Standards list

---

P. Dakić

Faculty of Informatics and Computing, Singidunum University, Belgrade, Serbia

P. Dakić (✉) · V. Vranić

Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Bratislava, Slovakia  
e-mail: [pavle.dakic@stuba.sk](mailto:pavle.dakic@stuba.sk)

V. Vranić

e-mail: [vranic@stuba.sk](mailto:vranic@stuba.sk)

V. Todorović

Faculty of Business Studies and Law, MB University, Belgrade, Serbia  
e-mail: [vladimir.todorovic@ppf.edu.rs](mailto:vladimir.todorovic@ppf.edu.rs)

## 1 Introduction

In the era of technological progress and growing human needs, the standards of the automotive industry are being raised. There is a need for sustainable environmental measures that reflect the direction of development of this branch of industry. Therefore, companies invest financial and research efforts in the competitive market, complying with numerous regulations and standards at the global level [4, 17].

In today's time frame of complex and demanding market, vehicle manufacturers adapt to the needs related to the consistency and constant improvement of the quality of the adopted standards. They do this while respecting the high expectations of the consumer society. The resulting challenges are reflected in the application of functional elements to support the use of automated vehicles, electrification, and their integration.

Improving the vehicle based on simulation ensures an evident reduction in production costs and the necessary time to bring the final product to the market [26]. With the adoption of communications based on the Ethernet network system, the possibility of synergy appeared, i.e., the need to reuse protocol procedures that have been perfected and tested in various applications, and it is necessary to carefully analyze, adapt, and adopt [5, 8, 22, 28].

The regulations set in the automotive industry support companies in terms of finding answers to changing requirements and improving processes and production lines. One of the applicable processes during the creation of appropriate harmonized software is a continuous integration and continuous delivery (CI/CD) [15]. It represents support to companies in preserving safety, creating efficiency, mitigating bad impacts on the environment, and a better integrative attitude toward new technologies.

There are standards that refer to several industries and those that, with their specificities, are intended for automotive companies [4–7]. International monitoring organizations have implemented many of these regulations, and companies in the countries where they are established receive third-party certifications related to process compliance [20, 21, 30, 31].

## 2 Paper Organization

Below are the key contributions and a brief organization of this work, which is composed of the following units: 1. overview and understanding of the differences in the use of standards and their importance, 2. an understanding of the division and types of standards, 3. review of relevant literature in the field of automotive industry standards and understanding their significance.

### 3 Automotive Industry Quality Standards

The branch of the automotive industry that includes suppliers of automotive components, marketing companies, and other involved parties implies a long and complex supply chain that is supported by an approach to IT security.

The trends that are represented in the global market are stimulated by new forms of demand from demanding consumers. Modern markets dictate both producers and buyers of goods based on changing needs and awareness and based on ecological standards related to the environment.

The strongest automotive companies realized by analyzing the global circumstances that the direction of further improvement of this industry is related to sustainability and the green agenda, security, and data protection. This refers to the complete production cycle of production and sale of products and an adequate marketing approach. This is the reason for summarizing the essential standards at the world level aimed at preserving security, effectively controlling compliance, and strengthening security information [19, 24, 29].

#### 3.1 Industry Quality Standards—Work in Progress (WIP)

Production management often involves the use of non-standardized and unscripted production techniques and methods. This may be a trade secret for most manufacturers, but the basis for gaining a market advantage. One such application is the term work-in-progress (WIP), which describes the use of partially finished raw materials in the process of creating a final product.

The main reason for applying this technique is the lack of sufficient information on the quality level and response from end customers, which implies that the first version will contain possible defects. The process itself reduces the creation of scrap, increasing the degree of production quality required by the application of industrial quality standards.

Considering that the overhead costs of installed capacities and products that are in the stages of final creation are created, it is necessary to obtain a confirmation of the possibility of using the appropriate degree of quality of the standard prescribed by WIP. Transforming small businesses into corporate systems is made possible by financial resources invested in strategies, tools, and industry quality standards.

The use of WIP allows excluding the value of raw materials from the company's balance sheet until the appropriate quality appears during testing. It provides companies with a degree of protection and disclaimer because each product contains a WIP tag.

Based on this, the user can better understand the expectations about the quality of the product. Marking is done in specific steps from the prototype to the final product [14].

### 3.2 *Automotive Quality Standards*

Software standards research related to the sustainable development and financing of successful software represents a relationship to understanding the contribution of AV. The most comprehensive overview of research trends within computing is covered by the IEEE organization, which deals with various technological procedures specifically related to electric vehicles (EV) and the possibility of wireless charging [25].

Within the available literature [1], one can see the most current intersection of wireless charging technologies, standard characteristics, and direct implementations related to the application of safety standards and ISO 14000 characteristics [1, 27].

### 3.3 *Compliance Standards List*

Preservation of high quality and ethical attitude in the processes from design to final product is ensured by various standards in the automotive industry. Companies in a competitive market try to secure their place by seeking certifications to conform to the appropriate standards [31].

Company officials perform safety checks and compliance with standards using car safety checklists. They represent a group of reminders and instructions whose task is to ensure that there are no omissions in safety checks and other activities and to preserve trust in auto components.

By understanding the basic four divisions of standards into: sustainability, quality, security, and data protection, there is a basis for business success. On the division shown in the attachment Fig. 1. We can clearly see the software perspective and the list of necessary international standards that should be applied when designing the software.

#### IATF 16949

The certificate called IATF 16949 represents a certain set of methods that should be implemented with the aim of obtaining safe and reliable production. Within the framework of the standard, the method of joint production and processes at the global level of the automotive industry is regulated. Its main task is to support the constant improvement and efficient production of products [17].

#### ISO 9001

One of the most current standards in terms of quality management is ISO 9001. It is important to mention the above-mentioned IATF 16949 standard, which is bundled with ISO 9001 and is one of the mandatory standards for the automotive industry.

Application of this standard enables effective management of processes and requirements related to key domain knowledge. It can also be used for training employees and effectively meeting standardization [17].



Fig. 1 Automotive standards checklist. Source [17]

ASPICE

The development of software quality requires an understanding of the basic principles of design and quality assurance—QA. By applying certain questions from the QA area, users are enabled to evaluate the process, and necessary implementations and define further software development needs within the automotive industry within the ASPICE framework [17].

ASPICE was developed on the basis of the ISO/IEC 15504 standard, relying on the V-model of development with a focus on testing each of the development phases of the software [17].

3.4 Data Protection Management System

Automotive companies work diligently to comply with various standards and use a wide range of software tools for this. The complexity of reviewing the current volume of documents requires an efficient organization of regulations and standards with the aim of achieving an appropriate degree of alignment.

The most important aspects within the industry are related to information security and the possibility of applying appropriate data protection techniques.

The success of managers is precisely reflected in a systematic approach to a large number of different standards and regulations, providing a great contribution to this branch of the industry. This includes the quality and appropriate use of data protection, a sustainable environmental direction, and the sublimated implementation of ISO 27001, TISAKS, ISO 9001, ASPICE, ISO 140001, and GDPR in all periods of the vehicle production cycle [19, 24, 29].

#### GDPR

The AV industry could not survive without numerous component suppliers and end consumers. Given that manufacturers cooperate with countries within the EU and process a large amount of information from partners and customers, it is necessary to apply the general data protection regulation (GDPR), which should be harmonized with the needs of The importance of data protection management system (DPMS) [17].

### 3.5 *EMS and Sustainability Standards*

Managing potential environmental risks requires having an appropriate policy within the company that enforces respect and compliance with government regulations. One of the essential standards is the environmental management system (EMS). It provides proactive management capabilities within specific tools, enabling companies to summarize and prevent potential risks in cases of environmental non-compliance.

#### ISO 14001

Control and review of all possible impacts of the automotive industry on the environment imply the use of the ISO 14001 standard, which is the basis of the EMS certificate, for all manufacturers in the production of automobiles. ISO 14001 was created with the aim of global regulation of sustainable standards and environmental protection, providing everyone with an equal chance on the market.

#### ISO 45001

Companies regulate the business environment by, among other things, applying certificates in the fields of health care and protection related to employees and users. Reducing work-related injuries and illnesses will contribute to the higher work performance of each employee, and that is why companies are introducing the ISO 45001 standard.

### 3.6 *Automotive Cybersecurity Standards*

Code review is performed using certain tools that do not always have the appropriate background in cybersecurity standardization, which is why certain unforeseen omissions may occur [23]. Most companies are trying to develop an awareness of the effective design of secure software and the use of new technological solutions [15, 16].



Technological progress in the automotive industry integrates safer and more efficient experiences, related to better communication between control units (ECU) and CAN-Bus buses. Obtaining a higher degree of stability increases the complexity of the used electronic components because a higher degree of connected devices is required.

Safety assessments require higher financial costs due to the reason of having the appropriate simulation and emulation model of vehicles and hardware. The challenges are reflected in the requirements for the application of automotive protocols that do not contain certain security flaws in cases of cyber attacks. This unknowingly exposes vehicles to a higher level of attack, which puts additional pressure on the appropriate security assessments and models used within the lab [4, 12].

## 4 Financial Sustainable of Compliance with Standards

The goals of achieving higher profits based on the production and quality of the goods are related to the organization of workers in order to promote personal and consumer satisfaction and protection from unfortunate circumstances during work tasks. The need for faster and more efficient mobilization of the process of standards, quality, and necessity of legal services is analyzed, which includes: timely recognition of errors, focus on work standardization, peer review by colleagues and partners, continuous improvement based on employee performance measurement [18, 32].

Building a software project purposefully involves three important factors: quality, improvement, and price. It is necessary that the base state is based on three scopes that will be realized with the help of project management. Managed activities and production processes within the automotive industry are calculated into the cost prices of final high-tech products, taking into account the additional values of standardization and the impact of production functions.

The result of system standardization and economic profit is influenced by the analysis of engineering projects. The sublimation of confirmations about the sustainability of standards and their possible implementations will have an impact on smaller or larger technological progress and further directions in the automotive industry [9, 32].

### 4.1 SASB Standards

The necessity of the company's financial stability was confirmed by familiarization with organizations dealing with materiality standards. Transparency is a factor for companies' compliance and is implemented and refined by the Sustainability Accounting Standards Board (SASB). The estimated value of the company is influenced by risk-based sustainability indicators, all of which are provided by SASB standards [11].

The identification of different bodies and types of standards implies a division into certain industrial branches of application, where specifically the automotive industry raises a number of questions related to the problems of non-application of certain standards.

The main impact refers to machines and final products that have not yet been required to apply appropriate standards, which results in an increase in costs, the purchase of new machines/parts, and the occasional stoppage of food production using AV in the fields [3].

## 5 Related Work

Ecological and conventional criteria form the basis of the joint evaluation of component suppliers [13]. The system (WiP) being perfected tends to ensure optimal income. Expediency [2] is reflected inefficiently designed disassembly systems and dismantling lines, management of changing product quality at the end of the life cycle (EoL).

The improvement of the devices in the markets, including the implementation of parts of the instructions from the ISO standards (ISO 13485 and ISO 14971), is based on essential translational steps. These standards are necessary for the practical demonstration of solving the problems of regulatory inefficiency and liability integrated into the production of WiP autonomous medical devices (quality control systems, risk management, design control [10, 33, 34]).

Very important for the creation of highly sophisticated products in the electronics and auto industry are the proposed valid methods of repeating testing and involving standard procedures. For this, the moving protective tape tests were used, which enabled the quality of the test and the yield to be improved. These are means for achieving high-quality goals, but also for canceling possible failures [35].

## 6 Conclusion

By implementing the WIP standard, the progressive growth of complex safety-critical elements integrated with innovatively advanced systems to assist autonomous vehicles and drivers is enabled. As the evolution of the automotive industry has entered a more dynamic period of transformation of vehicle design and production methods, research has been carried out, and the effects of integration have been carried out, including modeling, analyzing, and refining a strategy based on software compliance.

Based on increasing demands, manufacturers are rolling out testing solutions that focus on sustainable development, software challenges, and electronics. It is necessary to implement WIP solutions that must adhere to long-term cybersecurity functionality and reliability driven by high-quality standards.

On the basis of the research carried out, there is a need for further research into the relationship between problems with harmonization within the framework of standards and software design management. A key question arises for the future perspectives of the technological direction, how to use the techniques for solving the software compliance problem in the automotive industry, given the lack of chips on the market and numerous logistical problems.

**Acknowledgements** The work reported here was supported by the Scientific Grant Agency of Slovak Republic (VEGA) under grant No. VG 1/0759/19 and the Operational Program Integrated Infrastructure for the project: Advancing University Capacity and Competence in Research, Development and Innovation (ACCORD), ITMS code 313021X329, co-funded by the European Regional Development Fund (ERDF) Grant Number: 313011W988.

## References

1. Ahmad A, Alam MS, Chabaa R (2018) A comprehensive review of wireless charging technologies for electric vehicles. *IEEE Trans Transp Electrification* 4(1):38–63 (Mar 2018). <https://doi.org/10.1109/tte.2017.2771619>
2. Bentaha ML, Moalla N, Ouzrout Y (2020) A disassembly line design approach for management of end-of-life product quality. In: *Product lifecycle management enabling smart X*. Springer International Publishing, pp 460–472. [https://doi.org/10.1007/978-3-030-62807-9\\_37](https://doi.org/10.1007/978-3-030-62807-9_37)
3. Ćurčić M, Todorović V, Dakić P, Ristić K, Bogavac M, Špiler M, Rosić M (2021) Economic potential of agro-food production in the republic of serbia 68(3):687–700. <https://doi.org/10.5937/ekopolj2103687c>
4. Dakić P, Živković M (2021) An overview of the challenges for developing software within the field of autonomous vehicles. In: 7th conference on the engineering of computer based systems. ECBS 2021, Association for Computing Machinery, New York. <https://doi.org/10.1145/3459960.3459972>
5. Dakić P, Savić J, Todorović V (2021) Software quality control management using black-box testing on an existing webshop trinitishop. *FBIM Trans* 9(1) (May 2021). <https://doi.org/10.12709/fbim.09.09.01.03>, <https://www.meste.org/ojs/index.php/fbim/article/view/1137>
6. Dakić P, Todorović V (2021) Cost-effectiveness and energy efficiency of autonomous vehicles in the eu. *FBIM Trans* 9(2) (10 2021). <https://doi.org/10.12709/fbim.09.09.02.03>, <https://www.meste.org/ojs/index.php/fbim/article/view/1198>
7. Dakić P, Todorović V, Biljana P (2021) Investment reasons for using standards compliance in autonomous vehicles. In: *ESD conference, Belgrade 75th international scientific conference on economic and social development, ESD conference Belgrade, 02-03 December, 2021* MB University, Teodora Drajzera 27, 11000 Belgrade, Serbia. <https://www.shorturl.at/diMRS>
8. Dakić P, Todorović V, Vranić V (2022) Financial justification for using ci/cd and code analysis for software quality improvement in the automotive industry. In: *2022 IEEE zooming innovation in consumer technologies conference (ZINC)*, pp 149–154. <https://doi.org/10.1109/ZINC55034.2022.9840702>
9. Dakić P, Todosijević A, Pavlović M (2016) The importance of business intelligence for business in marketing agency. In: *International scientific conference ERAZ 2016 knowledge based sustainable (2016)*. <https://doi.org/10.13140/RG.2.1.2490.3289>, značaj poslovne inteligencije za poslovanje marketinške agencije
10. Diga D, Severin I (2021) Key life test process optimization using six-sigma approach. *J Innov Bus Best Pract* 1–17 (Feb 2021). <https://doi.org/10.5171/2021.536861>

11. Foundation TI (2022) Sasb standards overview (Nov 2022). <https://www.sasb.org/standards/>
12. Granata D, Rak M, Salzillo G (2021) Towards HybridgeCAN, a hybrid bridged CAN platform for automotive security testing. In: 2021 IEEE international conference on cyber security and resilience (CSR). IEEE (Jul 2021). <https://doi.org/10.1109/csr51186.2021.9527969>
13. Gupta S, Soni U, Kumar G (2019) Green supplier selection using multi-criterion decision making under fuzzy environment: a case study in automotive industry. *Comput Ind Eng* 136:663–680 (Oct 2019). <https://doi.org/10.1016/j.cie.2019.07.038>
14. Hayes A (2022) Work-in-progress (WIP) definition with examples (May 2022). <https://www.investopedia.com/terms/w/workinprogress.asp>
15. Hroncova N, Dakic P (2022) Research study on the use of CI/CD among slovak students. In: 2022 12th international conference on advanced computer information technologies (ACIT). IEEE (Sep 2022). <https://doi.org/10.1109/acit54803.2022.9913113>
16. Indriasari TD, Luxton-Reilly A, Denny P (2020) A review of peer code review in higher education. *ACM Trans Comput Educ* 20(3):1–25 (sep 2020). <https://doi.org/10.1145/3403935>
17. InfopulseSCM: Primary Standards for automotive Industry Compliance (Jun 2021). <https://compliance-aspekte.de/en/articles/checklist-of-mandatory-standards-for-automotive-industry/>
18. Jr DWL (2021) Evaluating legal services: the need for a quality movement and standard measures of quality and value. In: *Research handbook on big data Law*. Edward Elgar Publishing, pp 404–431 (2021). <https://doi.org/10.4337/9781788972826.00027>
19. Kiener C, Merklein M (2019) Research of adapted tool design in cold forging of gears. *Int J Mater Form* 13(6):873–883 (Sep 2019). <https://doi.org/10.1007/s12289-019-01508-0>
20. Krocka M, Dakic P, Vranic V (2022) Automatic license plate recognition using OpenCV. In: 2022 12th international conference on advanced computer information technologies (ACIT). IEEE (Sep 2022). <https://doi.org/10.1109/acit54803.2022.9913168>
21. Kročka M, Dakić P, Vranić V (2022) Extending parking occupancy detection model for night lighting and snowy weather conditions. In: 2022 IEEE zooming innovation in consumer technologies conference (ZINC). pp 203–208. <https://doi.org/10.1109/ZINC55034.2022.9840556>
22. Königseder KMT (2021) Protocols for automotive ethernet. In: *Automotive ethernet*. Cambridge University Press, pp 247–314 (Apr 2021). <https://doi.org/10.1017/9781108895248.010>
23. Lee W, Kim D, Park Y, Huh KY (2020) Development of a web-based open source CAE platform for simulation of IC engines. *Int J Automot Technol* 21(1):169–179 (Jan 2020). <https://doi.org/10.1007/s12239-020-0017-8>
24. Liu Y, Mukherjee N, Rajski J, Reddy SM, Tyszer J (2020) Deterministic stellar BIST for automotive ICs. *IEEE Trans Comput-Aided Des Integr Circ Syst* 39(8):1699–1710 (Aug 2020). <https://doi.org/10.1109/tcad.2019.2925353>
25. Malik MN, Khan HH (2018) Investigating software standards: a lens of sustainability for software crowdsourcing. *IEEE Access* 6:5139–5150. <https://doi.org/10.1109/access.2018.2791843>
26. Mantilla-Perez P, Perez-Rua JA, Millan MAD, Dominguez X, Arboleya P (2020) Power flow simulation in the product development process of modern vehicular DC distribution systems. *IEEE Trans Veh Technol* 69(5):5025–5040 (May 2020). <https://doi.org/10.1109/tvt.2020.2983288>
27. Patón-Romero JD, Baldassarre MT, Rodríguez M, Piattini M (2019) Application of ISO 14000 to information technology governance and management. *Comput Stand Interfaces* 65:180–202 (Jul 2019). <https://doi.org/10.1016/j.csi.2019.03.007>
28. Petričko A, Dakić P, Vranić V (2022) Comparison of visual occupancy detection approaches for parking lots and dedicated containerized rest-api server application. 3237. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139820286&partnerID=40&md5=039c8261aaa114d8a048afb907c06e5>
29. Pinto R, Pereira J, da Rocha H, Martin RI, Espirito-Santo A (2019) A discussion about the implementation of a WSN to industry 4.0 based on the IEEE 1451 standard. In: 2019 IEEE 17th international conference on industrial informatics (INDIN). IEEE (Jul 2019). <https://doi.org/10.1109/indin41052.2019.8972222>

30. Szarka R, Dakic P, Vranic V (2022) Cost-effective real-time parking space occupancy detection system. In: 2022 12th international conference on advanced computer information technologies (ACIT). IEEE (Sep 2022). <https://doi.org/10.1109/acit54803.2022.9913171>
31. Team IE (2021) 7 Important certifications for automotive companies (Nov 2021). <https://www.indeed.com/career-advice/career-development/automotive-industry-certifications>
32. Todorović V, Dakić P, Aleksić M (2021) Company management using managerial dashboards and analytical software. In: ESD conference, Belgrade 75th international scientific conference on economic and social development, ESD conference Belgrade, 02–03 Dec 2021, MB University, Teodora Drajzera 27, 11000 Belgrade, Serbia. <https://shorturl.at/diMRS>
33. Vieira A, Silva F, Campilho R, Ferreira L, Sá J, Pereira T (2020) SMED methodology applied to the deep drawing process in the automotive industry. *Procedia Manuf* 51:1416–1422. <https://doi.org/10.1016/j.promfg.2020.10.197>
34. Wegener K, Bleicher F, Heisel U, Hoffmeister HW, Möhring HC (2021) Noise and vibrations in machine tools. *CIRP Ann* 70(2):611–633. <https://doi.org/10.1016/j.cirp.2021.05.010>
35. Yeh CH, Chen JE (2019) Repeated testing applications for improving the IC test quality to achieve zero defect product requirements. *J Electr Test* 35(4):459–472 (Jul 2019). <https://doi.org/10.1007/s10836-019-05812-0>

# A Novel Multiband Patch Antenna Based on the Modification of a Rectangular Design



Rafael B. Méndez-Vásquez, Marcelo D. Lojano-Angamarca,  
Luis F. Guerrero-Vásquez, Jorge O. Ordoñez-Ordoñez,  
and Paul A. Chasi-Pesantez

**Abstract** In this paper, we propose a novel multiband patch antenna design. To achieve these characteristics, a conventional rectangular patch antenna was used. Modification process consisted of changing, on the one hand, the ground plane for a semicircular one, and on the other hand, in resonant plane components such as clamps and slots were added, with which multiband was achieved. In first instance, antenna simulation process was carried out in HFSS software. After that, antenna was built, to check if simulated data with measured results are similar. For measurement, a Virtual Network Analyzer (VNA) and a transceiver were used. Finally, when analyzing results, it was possible to define that antenna complies with expected parameters.

**Keywords** Multiresonant antenna · Patch antenna · Microstrip · Slot insertion · Notch filters

## 1 Introduction

Wireless communications advancement has caused an increase in technological demand for antennas [1], because requirements of miniaturizing its dimensions and improving electromagnetic characteristics are increasing [2, 3]. Currently, there

---

R. B. Méndez-Vásquez · M. D. Lojano-Angamarca · L. F. Guerrero-Vásquez (✉) ·  
J. O. Ordoñez-Ordoñez · P. A. Chasi-Pesantez  
Universidad Politécnica Salesiana, Cuenca, Ecuador  
e-mail: [lguerrero@ups.edu.ec](mailto:lguerrero@ups.edu.ec)

R. B. Méndez-Vásquez  
e-mail: [rmendezv@est.ups.edu.ec](mailto:rmendezv@est.ups.edu.ec)

M. D. Lojano-Angamarca  
e-mail: [mlojanao@est.ups.edu.ec](mailto:mlojanao@est.ups.edu.ec)

J. O. Ordoñez-Ordoñez  
e-mail: [jordonezo@ups.edu.ec](mailto:jordonezo@ups.edu.ec)

P. A. Chasi-Pesantez  
e-mail: [pchasi@ups.edu.ec](mailto:pchasi@ups.edu.ec)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_40](https://doi.org/10.1007/978-981-99-3091-3_40)

is a clear interest in microstrip and patch antennas development, due to their low production cost, small sizes, and with bandwidths that meet requirements of current telecommunications. In addition, these antennas provide support for different technologies, including fifth generation of mobile networks (5G).

Microstrip or patch antennas have their main application in mobile phones, drones, radars, and Wi-Fi networks; the 802.11b standard gives us operating frequencies from 2.4 and 5.8 GHz. One telecommunications challenge is antennas reduction for different applications like mobile networks [4–7], biomedicine, space suits [8], telemedicine with ultra-wideband textile antennas [9], agriculture, aerospace applications, GPS, textile antennas which are part of the next generation of wearable electronic antennas [10], among others [11]. A microstrip antenna is made up of two sides of conductive material, on a dielectric substrate. One of these faces contains the element that radiates power, while the other face is the ground plane (GND). Feeding is usually done by a coupled microstrip line [12]. Design approaches for these antennas have been explored since the 1970s [13], focusing mainly on two objectives: ultra-wideband (UWB) and multiband. To achieve these objectives, there are several proven techniques, which can be classified into four groups. The first technique, known as slot insertion, is generally used in mobile antennas, whose objective is to achieve more compact dimensions, improve performance, and achieve multiband operation [14, 15]. The second technique consists of making metallic through holes in order to generate additional resonant frequencies, in addition, when applying these holes, in some cases displacements of initial resonant frequencies are produced [16]. Main feature of the third technique lies in the use of disturbances or parasitic elements in slots, to reduce the size of the antenna [13, 17]. Finally, the fourth technique is based on the construction material, this being able to be a meta-material. The latter have the ability to alter material electromagnetic properties; thus, by entering time-varying fields, generated dipoles provide negative permeability and permittivity [18, 19].

In this context, our paper presents the design of a multiresonant patch antenna, based on basic principles and initial modeling of a rectangular patch antenna. Modifications were applied to initial design, in order to achieve multiple resonances in a range from 2.5 to 12.5 GHz. Antenna operation results were obtained through simulation in HFSS and, subsequently, they were verified with antenna construction and characterization.

The document is organized as follows. In Sect. 2, initial antenna design and modifications implemented to achieve multiband are presented. In Sect. 4, we include measurement results obtained in laboratory. These results are analyzed by comparing them with those obtained from simulations. Finally, Sect. 5 collects main conclusions generated with our work development.

## 2 Antenna Design

### 2.1 Initial Model

We started from a rectangular patch antenna model. This model served as an initial design to start making modifications to patch geometry, with the aim of achieving multiband characteristics. In Fig. 1a, patch layout is shown, while in Table 1, values of antenna dimensions are detailed.

### 2.2 Modified Model

First modification was made to ground plane. First, size was reduced by half and a circumference was placed on it, generating the shape shown in Fig. 2a. Dimensions of these modifications can be seen in Table 2.

Fig. 2b compares the values of initial antenna  $S_{11}$  and antenna with ground plane geometry modification. In general, the curve of  $S_{11}$  decreased, and a resonance frequency of 6.8 GHz was obtained.

Next modification was made in the feeder, with 0.25 mm insertions, to obtain a decrease in  $S_{11}$  parameter at frequencies close to 2.4–2.8 GHz. Dimensions of this geometry are specified in Table 3. In addition, four upper clamps were placed with alternate symmetries as shown in Fig. 3. Slots have also been added to patch's rectangular geometry sides (Fig. 4).

Figure 5 shows the simulated values of  $S_{11}$  parameter comparing initial with final antenna. Initial antenna, which was the starting point of the work, values of  $S_{11}$  can be seen from  $-17$  to  $-24.95$  dB, at frequencies of 3.4, 6.82, 9.44, 10, and 13.68 GHz.

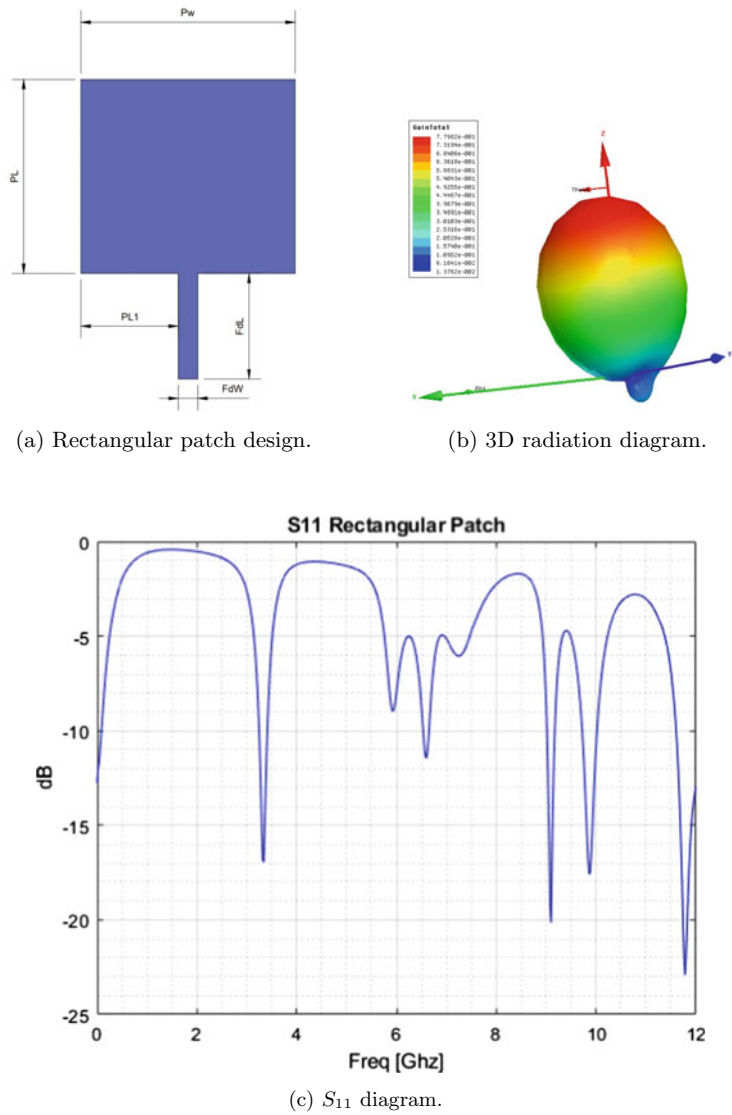
What was sought with modifications is to obtain acceptable values of  $S_{11}$  in frequencies of the ISM bands, that is, close to 2.4 and 5.1–8 GHz. Through modifications, lower values of  $S_{11}$  were obtained, as well as ISM resonance frequencies at 2.4 and 5.1 GHz.

$S_{11}$  parameter presents a great improvement compared to normal patch antenna frequencies. Resonant frequencies are found at 2.58, 3.54, 4.14, 5.34, 6.86, 7.68, and 9.92 GHz on the power scale in dB can be seen in Fig. 5, initial values of  $-20.90$  dB reaching a minimum of  $-30.42$  dB.

## 3 Antenna Manufacturing

Modified antenna was built in a 1.6 mm thick copper bakelite, FR4 epoxy material. A laser cut was needed for smaller slots that are located in the feeder and rectangular geometry sides. Antenna total dimensions are 45 mm  $\times$  45 mm, as shown in Fig. 5a; patch size is compared to US 1 dollar coin.

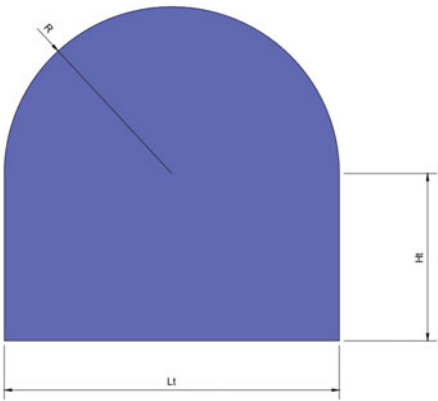




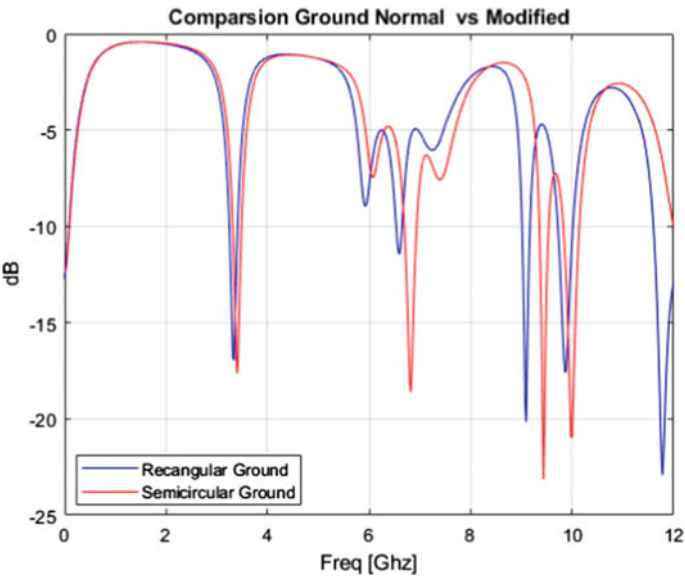
**Fig. 1** Initial design before modifications

**Table 1** Initial model dimensions

Patch [mm]		Feed	
$P_L$	20.6	$F_{dW}$	2.05
$P_W$	22.43	$F_{dL}$	11.07
$P_{L1}$	10.19		



(a) Semicircular modified ground plane.



(b)  $S_{11}$  diagram comparison between original and modified.

**Fig. 2** Modified model ground plane and  $S_{11}$  diagram comparison

**Table 2** Ground plane dimensions

Ground	[mm]
$H_t$	22.5
$L_t$	45
$R$	22.5



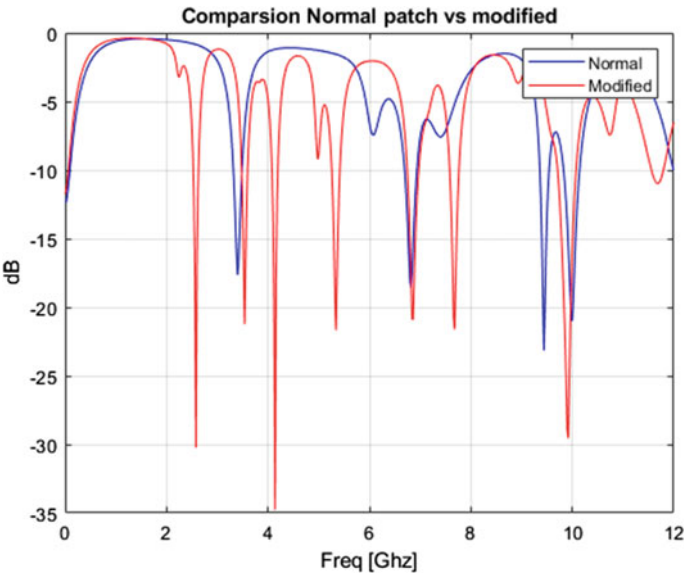


Fig. 4  $S_{11}$  diagram comparison between original and modified

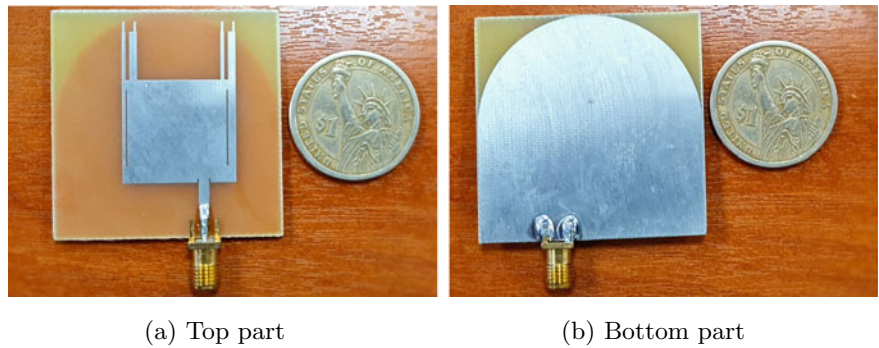
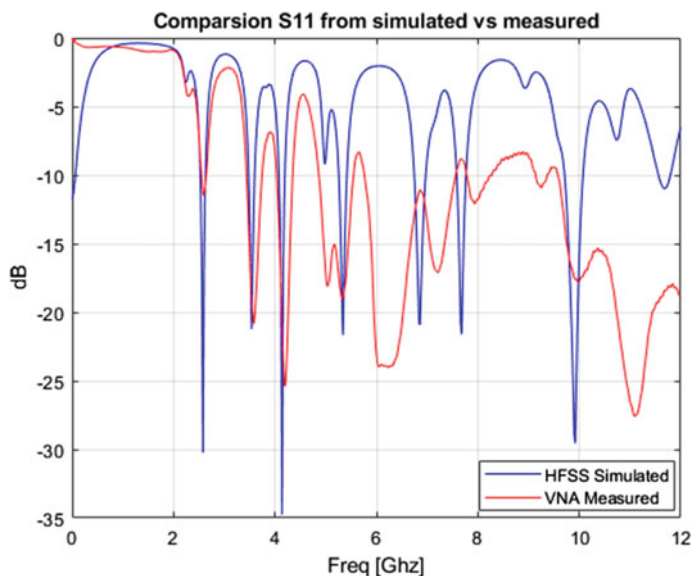


Fig. 5 Manufactured antenna

Frequency for radiation pattern measurement, for the different tests, was established at 4.2GHz, because this frequency was the one that had the best  $S_{11}$  at the time of analyzing in the VNA.



**Fig. 6** Comparison between simulation and measurement  $S_{11}$

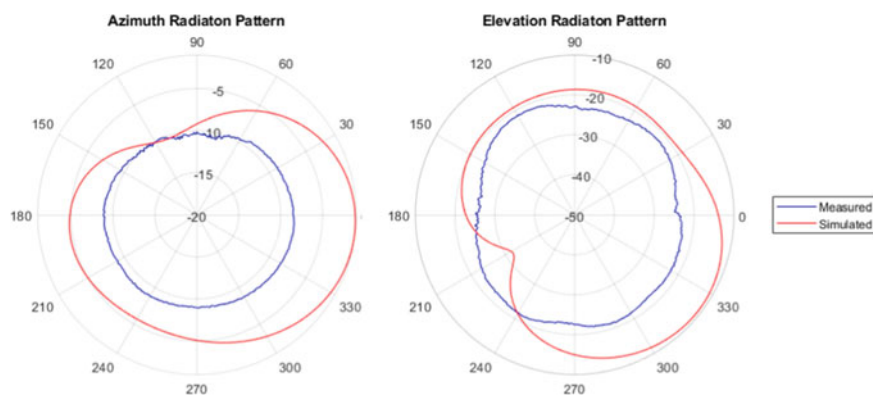
## 5 Conclusions

Starting from a basic rectangular patch antenna modeling, it has been possible to optimize and obtain multiband results, as can be seen in Fig. 2a. In addition, multiple resonances are observed starting from 2.5 to 12.5 GHz.

Antenna modifications are the product of many iterations made in HFSS simulator. Clamps placed on antenna top are the ones that allow the generation of the multiband, in combination with slots. They are fundamental to obtain values of  $S_{11}$  below  $-15$  dB. In this case, they have generated that the antenna has optimal results. When modifying slots, we can say that the smaller they are, the better results we will obtain. Regarding manufacture of this antenna, slots were made with a resolution of 0.25 mm, built with a precision laser.

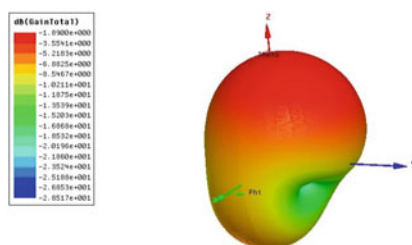
With this work we seek to demonstrate that any basic antenna can become a multiresonant antenna, with simple modifications such as changing feeder position. In initial model, it has a feeder placed in center, but once it is moved we see better results. Slots make  $S_{11}$  go down to values from  $-20$  to  $-35$  dB. When comparing final results with initial results of a rectangular patch antenna, it is observed that they are significantly better. For antenna characterization, a VNA with a measurement range from 5 KHz to 15 GHz and a National Instrument PXI with measurement ranges from 5 KHz to 6.5 GHz were used.

Tests were obtained in a space free of noise and interference. Results of characterization are consistent with respect to the simulation, in terms of  $S_{11}$  and radiation patterns. Stable radiation patterns are observed at the main resonance frequencies.



(a) Azimuth plane.

(b) Elevation plane.



(c) 3D radiation pattern.

**Fig. 7** Simulated and measured

## References

1. Song L, Rahmat-Samii Y (2018) A systematic investigation of rectangular patch antenna bending effects for wearable applications. *IEEE Trans Antennas Propag* 66(5):2219–2228
2. Matinmikko M, Mustonen M, Höyhty M, Rauma T, Sarvanko H, Mämmelä A (2010) Distributed and directional spectrum occupancy measurements in the 2.4 GHz ISM band. In: 7th international symposium on wireless communication systems. IEEE, pp 676–980
3. Benavides JB, Lituma RA, Chasi PA, Guerrero LF (2018) A novel modified hexagonal shaped fractal antenna with multi band notch characteristics for UWB applications. In: IEEE-APS topical conference on antennas and propagation in wireless communications (APWC). IEEE, pp 830–833
4. Gurjar R, Upadhyay DK, Kanaujia BK, Kumar A (2020) A compact modified sierpinski carpet fractal UWB MIMO antenna with square-shaped funnel-like ground stub. *AEU-Int J Electron Commun* 117:153126
5. Tiwari D, Ansari JA, Saroj AK, Kumar M (2020) Analysis of a miniaturized hexagonal sierpinski gasket fractal microstrip antenna for modern wireless communications. *AEU-Int J Electron Commun* 123:153288
6. Balanis CA (2015) *Antenna theory: analysis and design*. Wiley

7. Jara-Quito TA, Guerrero-Vásquez LF, Ordoñez-Ordoñez JO, Chasi-Pesantez PA, Morocho-Maita LA, Peralta LB (2020) Design of a flame fractal patch antenna for UWB applications. In: IEEE ANDESCON. IEEE, pp 1–5
8. Haagensohn T, Noghmanian S, de Leon P, Chang Y (2015) Textile antennas for spacesuit applications: design, simulation, manufacturing, and testing of textile patch antennas for spacesuit applications. *IEEE Antennas Propag Mag* 57(4):64–73
9. Yadav A, Kumar Singh V, Kumar Bhoi A, Marques G, Garcia-Zapirain B, de la Torre Díez I (2020) Wireless body area networks: UWB wearable textile antenna for telemedicine and mobile health systems. *Micromachines* 11(6):558
10. Shahariar H, Soewardiman H, Muchler CA, Adams JJ, Jur JS (2018) Porous textile antenna designs for improved wearability. *Smart Mater Struct* 27(4):045008
11. Rajan SP, Vivek C (2019) Analysis and design of microstrip patch antenna for radar communication. *J Electr Eng Technol* 14(2):923–929
12. Khidre A, Lee KF, Yang F, Elsherbeni AZ (2012) Circular polarization reconfigurable wideband e-shaped patch antenna for wireless applications. *IEEE Trans Antennas Propag* 61(2):960–964
13. Lu WJ, Li XQ, Li Q, Zhu L (2018) Generalized design approach to compact wideband multi-resonant patch antennas. *Int J RF Microwave Comput-Aided Eng* 28(8):e21481
14. Chen Q, Lin H, Wang J, Ge L, Li Y, Pei T et al (2018) Single ring slot-based antennas for metal-rimmed 4G/5G smartphones. *IEEE Trans Antennas Propag* 67(3):1476–1487
15. Guerrero-Vásquez LF, Jara-Quito TA, Ordoñez-Ordoñez JO, Cevallos-Gonzalez MK, Chasi-Pesantez PA (2019) A novel multi band patch antenna based on plotting exponential partial sums in the complex plane. In: 2019 international conference on electromagnetics in advanced applications (ICEAA). IEEE, pp 869–872
16. Arora C, Pattnaik SS, Baral R (2018) Dual band microstrip patch antenna array loaded with split ring resonators and via holes. *AEU-Int J Electron Commun* 93:253–260
17. Lituma-Guarta RA, Benavides-Aucapiña JB, Poveda-Pulla DF, Guerrero-Vásquez LF, Chasi-Pesantez PA (2018) A novel hybrid fractal antenna design for ultra-wideband application. In: 2018 IEEE 10th Latin-American conference on communications (LATINCOM). IEEE, pp 1–5
18. Desai A, Upadhyaya T (2018) Transparent dual band antenna with  $\mu$ -negative material loading for smart devices. *Microwave Opt Technol Lett* 60(11):2805–2811
19. Morocho-Maita LA, Peralta-Peralta LB, Bermeo-Moyano JP, Guerrero-Vásquez LF, Jara-Quito TA, Ordoñez-Ordoñez JO (2022) Reconfigurable antenna proposal based on origami techniques. In: Proceedings of sixth international congress on information and communication technology. Springer, pp 609–617

# Labour Conditions and Their Impact on the Development of Green Economies in 2020



María Fernanda Romo-Fuentes, Francisco J. Cantú-Ortiz,  
and Héctor G. Ceballos-Cancino

**Abstract** Nowadays, there is a global interest in making economies green, but how can we make them so? The economic agents, meaning, families, business and the state participate as a whole by performing activities that reduce their ecological footprint. For this purpose, research has been done in several areas to determine the impact certain factors have on ecology; thus, we propose a study in this paper that aims to show the impact that labour conditions have in green economies; specifically, a multivariate regression analysis is made between the CO<sub>2</sub> emissions per capita and five variables representing the labour conditions, which are average monthly earnings of employees, level of national compliance to labour rights, percentage of population covered by at least one social protection benefit, unemployment annual rate age 15+ and working poverty rate. This analysis is made employing data from 196 countries in 2020. The results show that of all the variables the ones that are significant for the regression model are the percentage of population covered by at least one social protection benefit and the working poverty rate, where the first is positively related while the second is negatively related to the level of CO<sub>2</sub> emissions, which indicates that the actual labour conditions are not appropriate for the development of green economies, and thus, changes must be made so that employees have better labour conditions, and this does not translate into damaging the environment.

**Keywords** Labour conditions · Green economies · Multivariate regression analysis

---

M. F. Romo-Fuentes (✉)  
Tecnológico de Monterrey, Estado de México, Mexico  
e-mail: [mferomof@gmail.com](mailto:mferomof@gmail.com)

F. J. Cantú-Ortiz · H. G. Ceballos-Cancino  
Tecnológico de Monterrey, Monterrey, Mexico  
e-mail: [fcantu@tec.mx](mailto:fcantu@tec.mx)

H. G. Ceballos-Cancino  
e-mail: [ceballos@tec.mx](mailto:ceballos@tec.mx)



# 1 Introduction

Much has been said about green economies and their importance to the future quality of human life; for this purpose, many studies have been made to recognize the impact that different human activities have on the development of green economies. Dating from several years back, it has made known that to achieve green economies a change in paradigms must be achieved by the population as a whole. In 2015, Poschen in his book “Decent work, green jobs and sustainable economy: solutions for climate change and sustainable development” [1] notes that climate change is a great opportunity to change the way jobs are done, since “the transition to environmentally and socially sustainable economies can become a strong driver of job creation, job upgrading, social justice and poverty eradication” [1]; moreover because of the fast pace at which climate change is advancing, we do not have enough time to solve each problem at a time, but to work with all aspects to obtain the best possible outcome.

Some researchers have chosen to identify the way in which green jobs, which are defined as “employment created in economic sectors and activities, which reduces their environmental impact and ultimately brings it down to levels that are sustainable” [2], are heavily influenced by factors such as education. In an article from 2019, Lee and van der Heijden [3] made a study to determine how institutions of higher education impact the creation of green jobs, which, in turn, impact circular economies; this was done by assessing “the impact that the presence of institutions of higher education has on the presence of green jobs in the 100 largest metropolitan regions in the United States” [3]; the data was analyzed with a multivariate regression analysis and two-stage least square regressions to conclude that “enhanced higher education and sustainability oriented departments and centers have a positive impact on green job development in urban regions” [3].

Others have chosen to study the development of green jobs from the policies perspective, which is the way governments support activities so that the transition from normal to green can be achieved, because it is easy to see that this change cannot be done by businesses alone. In [4] Bluedorn and Hansen made an analysis of advanced economies by model simulations to understand how to make a transformation to zero emissions economies. With this analysis they determined that it is important that governments implement “stronger environmental policies [...] a policy package incorporating a green infrastructure push, phased-in carbon prices, and targeted training and an earned income tax credit to provide income support and incentivize labour supply” [4].

Now, it seems clear that, even though green economies and their relation to jobs have been studied, it has not been yet completely understood how the current conditions of jobs around the world impact how green economies are; thus, for this paper, a multivariate regression analysis between CO<sub>2</sub> emissions per capita (dependent variable) and five variables representing the labour conditions (independent variables), which are average monthly earnings of employees, level of national compliance to labour rights, percentage of population covered by at least one social protection benefit, unemployment annual rate age 15+ and working poverty rate, is made, to

determine if the hypothesis: “there is a positive relation between labour conditions and the development of green economies” is valid.

The paper is presented as follows: in Sect. 2 a brief description of the data employed in the analysis is given, followed by a description of the applied methodology. In Sect. 3 the obtained results are presented followed by their analysis shown in Sect. 4. And, in Sect. 5 the conclusions to which we arrived after the data processing are presented.

## 2 Method and Data

In this section we give a brief description of the data employed in this research as well as the method in which this data is analyzed.

### 2.1 Data

The data used in this project comes from two databases:

- data corresponding to the CO<sub>2</sub> emissions per capita comes from the Emissions Database for Global Atmospheric Research, developed by the Joint Research Centre in the European Commission [5]
- data corresponding to the five variables with which we consider the labour conditions come from the database from the International Labour Organization (ILO-STAT) [6]. This data was, in turn, collected from different labour surveys applied by the countries' governments.

In Table 1 specifications for each variable can be seen. The year for which this data is considered is 2020 since it is the latest year for which this information is available.

It is important to note that we do not have the information for all the variables for all the countries. For the CO<sub>2</sub> emissions per capita we have the data for 196 countries, while for the independent variables, in the same order shown in the table, we have data for 80, 126, 124, 106 and 114 countries, respectively, which means that before analyzing the data we have to process it so that we only consider in each test the countries for which we do have information for the variables considered in said test. In other words, before making tests between the dependent and different combinations of the independent variables, for each test we only consider the countries for which we have information regarding the variable to be analyzed, for example, if we make a test between the CO<sub>2</sub> emissions per capita and average monthly earnings of employees plus the percentage of population covered by at least one social protection benefit, we check for the rows in which we have *NaN*, which indicates there's no data for that variable for that country, and do not consider this country for that particular test.

**Table 1** Variables employed in the multivariate linear regression

Variable	Measurement unit	Symbol
CO <sub>2</sub> emissions per capita	Million tonnes	MT
Average monthly earnings of employees	US dollars	USD
Level of national compliance to labour rights (freedom of association and collective bargaining (FACB))	This variable is measured in a scale from 1 to 10, “with 0 being the best possible score (indicating higher levels of compliance with FACB rights) and 10 the worst (indicating lower levels of compliance with FACB rights)” [6]	
Percentage of population covered by at least one social protection benefit	Percentage	%
Unemployment annual rate age 15+ (unemployed persons as percentage of the labour force)	Percentage	%
Working poverty rate (percentage of employed people living in poverty)	Percentage	%

## 2.2 Method

After processing the data so that we do not consider the countries for which we do not have values for those variables, as was stated in the previous subsection, and before making the tests, two things were done. The first one was a statistical analysis; in this analysis the mean, median, minimum value, maximum value, standard deviation, kurtosis and skewness were calculated for each variable, and this is important because in linear regressions the variables whose values are bigger and tend to have more influence in the result, and therefore, with the statistical analysis it is possible to determine if, before the linear regression, the variables must be normalized or standardized.

After the statistical analysis, all the independent variables were plotted against the dependent variable; this is done to identify the type of relation the variables have. Employing the results of the statistical analysis shows that all the variables are located within different ranges of values, and the graphs show that not all the independent variables have a linear relation to the dependent variable; it is chosen to apply a natural logarithmic transformation to the data, as this assures that the range of values for each independent variable is almost the same and the relation between dependent and independent variables is close to linear.

Now, after transforming the data, different tests were done to determine which independent variables explain the dependent variable the best, these tests were done by employing the *statsmodel* library in Python, and by employing this library it is possible to develop a model that fits the data through ordinary least squares (OLS),

reminding that this method “requires a straight line to be fitted to a set of data points such that the sum of the squares of the deviations from the observed data points to the assumed line is minimized” [7] As a result of applying the OLS commands we can get the regression results in a table, from which we can check if the variables are statistically significant and the evaluation parameters of the model as a whole, the first parameters analyzed for each model were the *F-statistic* and the *P value*, since we had to check if the variables considered in each test were significant; moreover, by using these two values we choose the variables for the final model that will be shown in the following section; the chosen variables pass the *F-statistic* and the *P value* tests at 95% significance.

After checking the *F-statistic* and the *P value* we check other important values, which are the  $R^2$  and adjusted  $R^2$  parameters, since they are used to evaluate the final model, as  $R^2$  gives a quantitative measure of how good a given regression is to a set of data points. The value ranges from 0 to 1, where 0 means the fit is terrible and 1 means we have a perfect fit; thus, a low  $R^2$  means it's a poor fit and a high  $R^2$  value means it's a good fit [8]. And, finally, to assess if the developed models are valid, we check for the values corresponding to the Durbin–Watson (DW) test, to test for multicollinearity and the Jarque–Bera (JB) test, to determine if the residuals show a normal distribution, and by applying the Breusch–Pagan test found in the same library, we test for heteroscedasticity.

The tests were conducted in the following way, first one per each independent variable against the dependent variable, then one employing all the independent variables, and after that, by checking if the variables passed the *F-statistic* and the *P value* tests, as was previously stated, we eliminated variables, one at a time, until we get the final model, shown in the following section.

### 3 Results

After making the tests described in the previous section, it was found that the independent variables that describe best the changes in the dependent variable are the percentage of population covered by at least one social protection benefit ( $x_1$ ) and the working poverty rate ( $x_2$ ), since the models that include these variables pass the *F-statistic* test, and individually, they pass the *P-value* test; moreover, with just these two values  $R^2$  and *Adj. R*<sup>2</sup> have the highest evaluation, these two are close to 0.7, which indicates that the model is good but can be definitely improved, which, in turn, means that there are variables that are missing from the model. Employing the mentioned variables we can analyze the data for 86 out of the 196 countries, for which we obtain the OLS regression results shown in Table 2

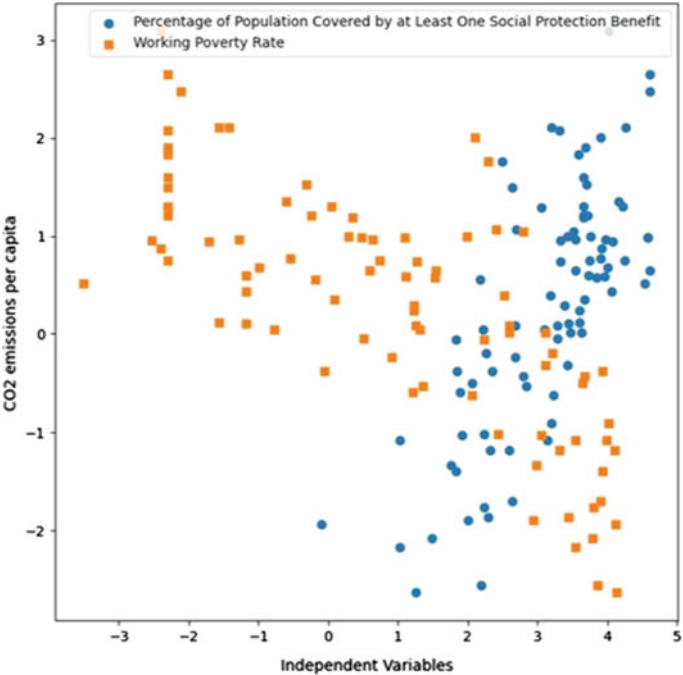
This means that the regression model is represented as follows:

$$\log(\text{CO}_2) = -1.236 + 0.5588 \log(x_1) - 0.2722 \log(x_2) \quad (1)$$

**Table 2** OLS regression results

Variable	Coefficient	Std error	t-statistic	Prob
Constant	−1.2360	0.361	−3.422	0.001
$x_1$	0.5588	0.104	5.366	0.000
$x_2$	−0.2722	0.044	−6.136	0.000

$R^2 = 0.689$ ,  $Adj.R^2 = 0.681$ ,  $Prob(F - statistic) = -9.28e - 22$ ,  $DW = 1.673$ ,  $JB = 3.013$ ,  $Prob(JB) = 0.222$



**Fig. 1** CO<sub>2</sub> emissions per capita versus significant dependent variables

Figure 1 shows a scatter plot in which the relation between the two independent and dependent variables can be seen. By analyzing the graph it is clear that the logarithmic transformation made the relation between the variables an almost linear one, which was what we were aiming for and in accordance with the regression results in Table 2 and Eq. 1, there's a positive and a negative relation between the percentage of population covered by at least one social protection benefit, the working poverty rate and the CO<sub>2</sub> emissions per capita, respectively.

Now, regarding the validity of the model, we know that there must not be significant levels of autocorrelation in the residuals of the data after the fitting of the linear regression model since the opposite could mean that the validity of the model is questionable, because of this, we use the DW test, as was stated in the methodol-

ogy section. For this, we first obtain the critical values for our dataset from the table shown in [9], considering we have two independent variables and 86 elements in the data, at 5% significance level, we have a lower limit of 1.600 and an upper limit of 1.696, comparing these values with the one obtained from our dataset, which is 1.673, and we can see that it is within the interval limited by the upper and lower limit. This indicates that there is no autocorrelation.

On the other hand, it is important to test for the normal distribution of the residuals because if this condition is not satisfied the computed estimators lose efficiency, and the F-statistic and P-value tests lose validity [10]. Thus, we employ the JB test, since the P value of the JB test is 0.222, which is higher than 0.05, the significance value, and we cannot reject the null hypothesis which means that the residuals do have a normal distribution.

Finally, homoscedasticity, this property relates to similar variance in the data residuals, this is important because if the contrary is true, that is, if the variances are not similar, which is called heteroscedasticity, the test results could be biased or skewed [11]. To test for this property, we employ the Breusch–Pagan test, which gives a p-value of 0.9605, since this value is higher than 0.05, which is the significance level, we accept the null hypothesis, that is, the residuals of the regression show homoscedasticity.

## 4 Discussion

Given the results shown in the previous section, it can be seen that the variables which remain as part of the model are the percentage of population covered by at least one social protection benefit and the working poverty rate, and since the developed model is expressed as logarithms, this means that all changes are expressed as percentages. Thus, an increase of 1% of the percentage of population covered by at least one social protection benefit produces and increase in approximately 0.558% in the CO<sub>2</sub> emissions per capita, while an increase in 1% of the working poverty rate decreases the CO<sub>2</sub> emissions per capita in about 0.272%.

Now, what does this mean? First, the fact that the CO<sub>2</sub> emissions diminish as the working poverty rate increases means that as employees have a low-level income they have less possibilities of consumption, and thus generate less pollution, while the fact that the level of CO<sub>2</sub> emissions is positively related to the percentage of population covered by at least one social protection benefit means that as people have better labour conditions the level of pollution increases, and this could also be related to the level of consumption of employees as the jobs that give better social benefits are the high-level income ones.

These results highlight that the actual labour conditions are not positive for the development of green economies; thus, we must consider how to improve these conditions while, also, improving the results. In related works to this research, we can see that the development of green economies is significantly affected by labour as well as policy making and education, as was stated in the introduction. Moreover,

the impact all these factors would have combined is what people should actually aim for. For example, in the work by Wang and Shao [12] it was found that environmental regulations represented by environmental-related technologies and education levels show positive and significant impact on green growth, which also sheds light into the importance of Research and Development (R&D) investment, or in the work of Burger et al. [13] an analysis to determine which work skills and type of education are related to progress of circular economies is described; at the end it was found that there is no cohesion between the types of skills and education needed for a sustainable economy, but given the education and skills demand for this type of economic activities, special education and training programs must be developed.

## 5 Conclusion

After the development of this research, we can see that our initial hypothesis stated that there is a positive relation between labour conditions and the development of green economies is valid, as the actual labour conditions do have an impact in the level of CO<sub>2</sub> emissions per capita; this relation shows that the actual labour conditions are not appropriate for the development of green economies, and thus, we must attend to modifying how work and consumption are done, so that employees can have better labour conditions and this does not translate into increasing the level of pollution.

For this purpose job policies must be appropriately modified and education with a green perspective must be given. It is suggested that further research is done to determine and measure how different factors affect the development of green economies, as we know that labour, policies and education are important determinants but knowing for sure would make policy making and the allocation of resources easier and more efficient.

**Acknowledgements** We thank Tecnológico de Monterrey and Conacyt for the resources and support given for the development and publication of this project.

## References

1. Poschen P (2015) Decent work, green jobs and the sustainable economy: solutions for climate change and sustainable development. Greenleaf Publishing [u.a.], Sheffield
2. Fernando M (2011) Introduction to green jobs. Key concepts. [https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---ilo-jakarta/documents/presentation/wcms\\_164514.pdf](https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---ilo-jakarta/documents/presentation/wcms_164514.pdf)
3. Lee T, van der Heijden J (2019) *Energ Environ* 30(1):141. <https://doi.org/10.1177/0958305X18787300>
4. Bluedorn J, Hansen NJ (2022) World economic outlook 2022. International Monetary Fund
5. Crippa M et al (2021) Emissions database for global atmospheric research, version v6.0\_ft\_2020 (GHG time-series). <https://data.jrc.ec.europa.eu/dataset/2f134209-21d9-4b42-871c-58c3bdcfb549>

6. International Labour Organization. ILOSTAT (2022). <https://ilostat.ilo.org/>
7. Kashyap R, Kumar AVS (eds) (2019) Challenges and applications for implementing machine learning in computer vision. IGI Global, Hershey, Pennsylvania. OCLC: 1154511974
8. Kane F (2017) Hands-on data science and Python machine learning: perform data mining and machine learning efficiently using Python and Spark. Packt Publishing, Birmingham, UK. OCLC: 1001346998
9. Mooi E, Sarstedt M (2010) A concise guide to market research. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 161–200. [https://doi.org/10.1007/978-3-642-12541-6\\_7](https://doi.org/10.1007/978-3-642-12541-6_7)
10. Chica Olmo J (2014) Problema sobre normalidad 1. Tech. rep., Facultad de Ciencias Económicas y Empresariales. Universidad de Granada, Granada, España. <https://www.ugr.es/jchica/Pagina2/GUIME/Otrosn%20ejercicios/Normalidadn%201.pdf>
11. Scribbr (2022) What is homoscedasticity? <https://www.scribbr.com/frequently-asked-questions/what-is-homoscedasticity/>
12. Wang X, Shao Q (2019) Sci Total Environ 660:1346. <https://doi.org/10.1016/j.scitotenv.2019.01.094>, <https://linkinghub.elsevier.com/retrieve/pii/S0048969719301056>
13. Burger M et al (2019) Res Policy 48(1):248. <https://doi.org/10.1016/j.respol.2018.08.015>. <https://linkinghub.elsevier.com/retrieve/pii/S0048733318302026>



# A Review of Deep Learning Techniques of Chest X-ray Analysis for Thoracic Disorders



Pawan Sharma, S. Gurunarayanan, and Anupama Karuppiiah

**Abstract** ML and DL algorithms are increasingly used in medical applications for monitoring and diagnosis of various health conditions. DL algorithms are increasingly used in image-based diagnosis. DL algorithms require a GPU cluster and a large amount of data to be trained. Even for testing, a GPU cluster is needed. The objective of our research work is to build accelerators to implement DL algorithms to diagnose thoracic disorders using chest X-rays. Hardware accelerators are easily interfaceable to existing CPU or SoC-based systems. This makes the system more portable, accessible, and cheaper. In this paper, we present a review of various DL algorithms and how they are used in medical diagnosis. We also have reviewed existing deep learning accelerators in medical applications.

**Keywords** Deep learning · Hardware accelerators · Medical imaging · Convolutional neural networks · Thoracic disorders

## 1 Introduction

Thoracic disorders include disorders such as achalasia, chronic obstructive pulmonary disease, emphysema, lung cancer, pneumothorax, and others that affect the heart, lungs, mediastinum, esophagus, chest wall, diaphragm, and major vessels [1]. We need a team of cardiologists, vascular experts, and radiologists to undertake cutting-edge diagnostic heart and vascular tests in order to find these disorders.

---

P. Sharma (✉)

Birla Institute of Technology and Science, Pilani, Pilani Campus, Pilani, India  
e-mail: [ps@pilani.bits-pilani.ac.in](mailto:ps@pilani.bits-pilani.ac.in)

S. Gurunarayanan

Birla Institute of Technology and Science, Pilani, Hyderabad Campus, Telangana, India  
e-mail: [sguru@hyderabad.bits-pilani.ac.in](mailto:sguru@hyderabad.bits-pilani.ac.in)

A. Karuppiiah

Birla Institute of Technology and Science, Pilani, K. K. Birla Goa Campus, Goa, India  
e-mail: [anupkr@goa.bits-pilani.ac.in](mailto:anupkr@goa.bits-pilani.ac.in)

A group of imaging specialists examines high-resolution radiological imaging to discover specifics and find issues that could otherwise go undetected. Due to the high cost and limited accessibility of radiological facilities, this type of medical care may not be universally accessible to everyone. Medical professionals have been using this approach for many years to investigate and detect fractures or abnormalities in bodily organs. Diagnosis on sensitive body areas such as bones, chest, teeth, etc., is done using X-rays. This is because X-rays are non-invasive and extremely effective diagnostic tools for detecting abnormal alterations. Chest X-rays can detect chest problems such as cavitation, consolidations, infiltrates, blunted costophrenic angles, and widely dispersed small nodules [2]. The radiologist can identify many of the disorders and diseases indicated above by analyzing chest X-rays. Multiple chest X-ray images must be laboriously classified by radiologists in order to identify anomalies. Even when one has access to medical services, getting a second opinion in cases of acute illnesses or unusual disorders is frequently necessary. Thus, machine learning and deep learning approaches are crucial for evaluating thoracic images.

The rest of the paper is organized as follows. In Sect. 2, we provide a brief overview of the various deep learning algorithms used in the literature review for analyzing X-ray images. We provide an overview of the major public datasets for thoracic disorders in Sect. 3. Numerous analysis techniques, such as localization, segmentation, and annotations, are the main emphasis of dataset analysis. We delve deeper into each of these image analysis methods in Sect. 4. The next section, Sect. 5, gives an overview of the various hardware accelerators that can be used to increase the system's computational capability and analyze medical images considerably more quickly than using traditional computing methods. The conclusion is in Sect. 6.

## 2 Overview of Deep Learning Methods

ML and DL algorithms have become more prevalent recently in a variety of medical specialties. Although the idea behind deep learning is not new [3], the recent explosive growth in computing power and the accessibility of digital data has made it possible for it to be successfully applied in a variety of applications, including NLP, speech recognition, self-driving cars, and most notably health care [4]. CNN, one of the common DL algorithms has been applied successfully in the area of computer vision. The ability of deep learning-based risk values derived from CXR images to predict long-term all-cause mortality was recently demonstrated [5].

### 2.1 Deep Learning Algorithms

ML algorithms are categorized as either supervised or unsupervised. The primary objective of supervised learning is to label unlabeled data by training a model with known data. A model is presented with the dataset,  $D = [x, y]_{n=1}^N$  of input features  $\mathbf{x}$

and label  $\mathbf{y}$  pairs where  $\mathbf{y}$  represents an instance of a fixed set of classes. In a regression scenario,  $\mathbf{y}$  can alternatively take the form of a vector of continuous values. The objective of supervised training is to determine model parameters,  $\Theta$  that, according to a loss function,  $L(y, \hat{y})$ , best characterize the data. The output of the model,  $\hat{y}$  is known as the loss function, and it is obtained by inputting a data point, denoted by  $\mathbf{x}$ , through the function  $f(x; \Theta)$  which is used to describe the model.

Unsupervised machine learning involves analyzing data in the absence of labels. Unsupervised algorithms are trained to search for patterns. Some of the common unsupervised algorithms include (a) Principal component analysis (PCA) (b) Clustering. Unsupervised learning uses a set of loss functions. An example of this is reconstruction loss defined by  $L(x, \hat{x})$ ; in this case, the model learns to reconstruct its input from a lesser dimension or noisy input.

## 2.2 Neural Networks

The vast majority of deep learning techniques are based on neural networks. A neural network consists of “neurons or units with a certain activation  $\mathbf{a}$  and parameters  $\Theta = \{\mathbf{W}, \mathbf{B}\}$  where  $\mathbf{W}$  is a set of weights and  $\mathbf{B}$  is a set of biases”. The activation is a linear representation of the combination of neuron’s input  $\mathbf{x}$  and the parameters and defines the transfer function, also called element-wise linearity  $\sigma(\cdot)$  as:

$$a = \sigma(\mathbf{w}^T \mathbf{x} + b)$$

Transfer functions for traditional neural networks are usually sigmoid or hyperbolic tangent functions. The most well-known conventional neural network, the multi-layered perceptron (MLP), has several layers of transformations:

$$f(\mathbf{x}; \Theta) = \sigma(\mathbf{W}^T \sigma(\mathbf{W}^T \dots \sigma(\mathbf{W}^T \mathbf{x} + b)) + b)$$

where  $\mathbf{W}$  is a matrix comprising of columns  $\mathbf{w}_k$ , connected to output activation  $k$ . The term “hidden layer” is frequently used to describe the layers between input and output. The activations are mapped to the distribution over classes  $P(y|\mathbf{x}; \Theta)$  through the network’s last layer at this point:

$$P(y|\mathbf{x}; \Theta) = \text{softmax}(\mathbf{x}; \Theta) = \frac{e^{\mathbf{w}_i^T \mathbf{x} + b_i}}{\sum_{k=1}^K e^{\mathbf{w}_k^T \mathbf{x} + b_k}}$$

where  $\mathbf{w}_i$  gives the weight factor that leads to a class-related output node,  $i$ .

Today’s most popular models are entirely supervised. CNN and RNN are popular architectures. Though CNN is commonly used for image analysis, RNNs are also gaining popularity.

## 2.3 Convolutional Neural Networks

CNNs form the basis for all image processing tasks involving deep learning. Neurons in convolutional layers only link to a small portion of the previous layer's "receptive field." By acting as a sliding window over all regions and efficiently detecting the same local pattern everywhere, these neurons are applied to various parts of the previous layer. In this manner, both the learnt weights and spatial information are transmitted. The CNN weights are assigned in order for the network to be able to perform convolutional operations on images. Since the model does not require the learning of separate detectors for the same object occurring in different places in an image, this renders the network equivariant with regard to input translations. The input image is convolved by each layer with a set of  $\mathbf{k}$  kernels,  $W = (W_1, W_2, W_3, \dots, W_k)$  and added biases,  $B = (b_1, b_2, b_3, \dots, b_k)$  which result in the creation of a new feature map  $X_k$  for that layer.

## 2.4 Deep CNN Architectures

The common architecture and architectural variations of the prevalent CNN architectures are described in the ensuing subsections.

### 2.4.1 General Classification Architectures

LeNet [6] and AlexNet [7] are shallow networks that use kernels with big receptive fields in the layers that are near the input and kernels with reduced receptive fields in the layers that are close to the output of the network. In place of hyperbolic tangents in the AlexNet activation functions, rectified linear units were used instead. Since 2001, deeper models have been preferred. Instead of using a single layer of kernels with a wide receptive field, it is possible to describe a similar function with fewer parameters by stacking smaller kernels. This enables a more precise representation. Deeper designs typically use less memory during inference, enabling the use of mobile computing platforms like smartphones. The first study to investigate considerably deeper networks with small fixed kernels in each layer was [8]. The 2014 ImageNet challenge was won by the VGG19 or OxfordNet model, a 19-layer architecture. To increase the effectiveness of the training process and decrease the number of parameters, more complicated building blocks have been built on top of these networks. A 22-layer network was introduced by Stollenga et al. [9]. Due to its use of inception blocks, GoogLeNet also referred to as Inception defines a [10] that replaces the module in defining the equation

$$X_k^l = \sigma [W_k^{l-1} * X^{l-1} + b_k^{l-1}]$$

that has a collection of varying-size convolutions. This makes it possible to define a similar function with fewer parameters, which is analogous to how the stacking of small kernels works.

### 2.4.2 Recurrent Neural Network

RNNs were initially created for discrete sequence analysis. Because both the input and output can be of various lengths, RNNs can be viewed as a generalization of MLPs.

Hence, RNNs are appropriate for applications such as machine translation. In this case, the input and the output are as a source of target languages. A latent or hidden state  $\mathbf{h}$  is maintained by the ordinary RNN at a time ‘ $t$ ’ as the consequence of a nonlinear mapping that is derived from the input  $\mathbf{x}_{(t)}$  and the previous state  $\mathbf{h}_{(t-1)}$ :

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{R}\mathbf{h}_{t-1} + b),$$

where  $\mathbf{W}$  and  $\mathbf{R}$  are time-shared weight matrices. One or more fully connected layers and a softmax are used to map the sequence to the posterior over the classes for classification.

$$P(y|x_1, x_2 \dots, x_T; \Theta) = \text{softmax}(\mathbf{h}_T; \mathbf{W}_{\text{out}}\mathbf{b}_{\text{out}})$$

Since RNNs are intrinsically deep, they experience the same issues as training with a standard deep neural network since gradient methods must be back propagated from the output over time. Several memory-focused devices have been designed to address this issue. The Long Short Term Memory [LSTM] is the oldest and most well-liked. A recent and widely used LSTM simplification is the Gated Recurrent Unit. RNNs were initially proposed for one-dimensional input, but they are now frequently used with images. They have been successfully applied to segmentation issues in medical applications [11].

## 3 Public Datasets

For automatic diagnosis, it is essential to identify thoracic disorders from X-ray images, either individually or collectively. Chest X-ray images in large quantities, are saved in image archiving and communication systems along with radiological reports (PACS). It can be difficult to manage and retrieve patient abnormalities data from these massive datasets for deep learning algorithms to use when creating computer-aided systems. The dataset is typically partitioned into small subsets called mini-batches, and a gradient step is carried out for each mini-batch in order to speed up the training of these big datasets. The term “iteration” is typically eschewed in favor

**Table 1** ChestX-ray public datasets

Dataset	Image/Patient	Format	Labels	Labeling method	Location
PadChest	I:160 K P:67 K	DICOM	174	RIP RP	<a href="https://bimcv.cipf.es/bimcv-projects/padchest/">https://bimcv.cipf.es/bimcv-projects/padchest/</a>
ChestX-ray8	I:109 K P:33 K	DICOM	8	NLP RR	<a href="https://www.kaggle.com/datasets/nih-chest-xrays/data">https://www.kaggle.com/datasets/nih-chest-xrays/data</a>
ChestX-ray14	I:112 K P:31 K	PNG	14	RP RI	<a href="https://nihcc.app.box.com/v/ChestX-ray-NIHCC">https://nihcc.app.box.com/v/ChestX-ray-NIHCC</a>
ChestX-Det10	I:3 K	PNG	13	RI	<a href="https://github.com/Deepwise-AILab/ChestX-Det10-Dataset">https://github.com/Deepwise-AILab/ChestX-Det10-Dataset</a>
MIMIC-CXR	I:372 K P:65 K	JPEG DICOM	14	RP	<a href="https://physionet.org/content/mimic-cxr-jpg/">https://physionet.org/content/mimic-cxr-jpg/</a>
CheXpert	I:224 K P:65 K	JPEG	14	RCI RP	<a href="https://stanfordmlgroup.github.io/computations/chexpert/">https://stanfordmlgroup.github.io/computations/chexpert/</a>
SIIM-ACR	I:16 K P:16 K	DICOM	1	RI	<a href="https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation">https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation</a>

Values above 10,000 are rounded and expressed as *K*  
*RP* Report parsing, *RI* Radiologist Interpretation, *RR* Radiologist Report, *CT* Computer Topography, *LT* Laboratory Tests, *NLP* Natural Language Processing

of the term “epoch” to denote a sweep through the complete dataset in order to avoid confusion with a gradient step (Table 1).

## 4 Deep Learning for Chest Radiography

We give a review of the literature on deep learning in medical imaging, particularly chest radiography, in this section.

### 4.1 Image Classification

Image or exam classification was one of the first uses of deep learning for medical image analysis. In exam classification, one or more photographs are provided as inputs, and an output diagnostic variable indicates whether or not the disease is

present. The majority of the research uses a pre-built dataset comprised of deep learning models for predicting a pathology, metadata, or a set of labels. Wang et al. [12] evaluates the efficiency of several classification algorithms for the 14 diagnosed illnesses in the ChestX-ray 14 dataset. Resnet, DenseNet, Inception, VGG, and AlexNet are the most often used architectures. The behavior of deep learning algorithms is also affected by image size [13]. By examining a whole image, image-level prediction can also be utilized for regression. Regression values are typically used to calculate a severity score for a specific pathology using data like the subject's age, gender, and medical history.

## 4.2 Segmentation

The quantitative analysis of clinical features that are linked to volume and form is made feasible by the segmentation of organs and other substructures in medical images. This makes it possible to examine these characteristics quantitatively. In the context of thoracic diseases, segmentation has been the subject of most research in the field of chest X-ray interpretation. A significant topic that is explored in written works is the localization of unusual or abnormal anatomical features. Before any image analysis is performed, the initial step in such pipelines is often the definition of the region of interest, abbreviated as ROI. In order to improve accuracy or make more accurate predictions, the segmentation process itself may be beneficial in quantifying clinical parameters based on form or area assessment [14]. Organ segmentation is one of the most researched chest X-ray segmentation tasks, as it is used clinically to determine heart enlargement (cardiomegaly) and is directly estimated from heart and lung segmentation [15]. Another use is called foreign object segmentation, and it involves separating foreign objects like catheters, tubes, and lines.

## 4.3 Localization

An essential pre-processing step in segmentation tasks is the localization of anatomical objects for organs or landmarks [16] (in space or time). Localization, which is the process of identifying a region of interest in medical imaging, typically necessitates the parsing of 3D volumes. The majority of research on chest X-ray analysis focuses on locating anomalies rather than objects or anatomy. Pneumonia, TB, and localization of nodules are often explored applications in the literature. In order to produce more accurate and quick localization algorithms, specialized architectures including YOLO, Mask RCNN, and Faster RCNN have been developed recently in computer vision research. These state-of-the-art designs have been rapidly improved in performance for CXR analysis. As an example, [17] demonstrated that pneumothorax detection on chest radiographs was successful using the first YOLO architecture. AUC values of 0.898 and 0.905 on 3-h and 1-day follow-up chest radiographs,

respectively, were attained by the model on chest X-rays collected after percutaneous transthoracic needle biopsy (PTNB) for pulmonary lesions in 1319 patients in an external dataset. Other works have used RetinaNet, Mask R-CNN, and RCNN architectures for the localization of nodules and masses [18].

#### **4.4 Image Generation**

Clinical workflow optimization of image quality to reduce multiple radiation dose exposures is a significant problem. The use of an iterative reconstruction technique results in notable improvements and reduces noise in CT images. Iterative reconstruction does, however, have some drawbacks. (1) Vendor-specific methods that demand CT scanner data for image smoothing that lose anatomical features such as interlobar fissures, and (2) the creation of novel image textures. By generating images with less noise but ones that are familiar to radiologists, deep learning-based image production might overcome these restrictions. The primary focus of many different research activities in this field is image generation. There are several applications for image production beyond denoising, such as data augmentation, visualization, abnormality detection through reconstruction, and domain adaptability. The preferred technique for producing chest X-rays is the Generative Adversarial Technique [19]. The majority of research is focused on CXR generation, which synthesizes images from random noise using unconditional GANs to augment training datasets [20].

### **5 Hardware Accelerators**

The development of specialized hardware to process deep learning algorithms has not yet kept pace with the field's rapid growth in terms of performance, network, size, and data size. A device that can train and infer deep neural networks quickly is referred to as an "accelerator." Few organizations, like Nvidia with its GPUs and Google with its Tensor Processing Units, along with relatively new startups and research groups trying to produce ASICs for deep learning training and acceleration, are now demonstrating progress in hardware accelerators. Specialized embedded deep learning accelerators such as Nvidia's Jetson and Xavier series and Movidius Neural Compute Stick [21] are available. A more recent edge accelerator designed specifically for the healthcare industry is called NVidia Clara embedded. According to the literature, deep learning accelerators are currently implemented using three different technologies: CMOS, FPGA, and Memristors.



## 5.1 CMOS DNN Accelerators

The healthcare IoT industry is made possible by CMOS-driven acceleration and accelerator chips. Massive parallelism and restricted memory access are used to accelerate MAC processes, which is beneficial for offline data centre scale acceleration as well as edge-enabled AI devices. While certain accelerators are used for recurrent connections in RNN accelerations, most are used for CNN inferences. With a few exceptions that use power in the range of a few watts, the total power consumption per chip for these accelerators is discovered to be in the mW range [22]. In the case of accelerators with high power requirements, which limit battery life otherwise, large heat sinks are required. Furthermore, it is demonstrated that a number of these accelerators can efficiently compute massive and deep CNNs, such as VGG and ResNet, while operating within a constrained power budget. Eyeriss, for example, has shown that it can be used as a mobile diagnostic tool that can be built into portable medical imaging systems and run AlexNet for a variety of medical imaging applications. The Origami chip is a potential CNN accelerator that is commonly utilized in CNN-based deep learning applications [23].

## 5.2 FPGA-Based Deep Neural Networks

Field-Programmable Gate Arrays are the first truly programmable logic devices that could be set up after manufacturing (FPGAs). They are made up of programmable logic, programmable interconnect, and programmable I/O blocks, as well as embedded processor cores, AI engines, high-density memory units, DSP cores, and other IP blocks needed to create a reprogrammable system-on-chip solution. They focus on applications where high parallelism and concurrency are desired due to their spatial structure and device fabric architectures. FPGAs provide a flexible framework that allows researchers to adapt to HDL synthesis flow or high-level synthesis flow by accepting logic input in both hardware description languages, such as VHDL or Verilog, and high level languages, such as System C, C++, and System Verilog [24, 25]. Stone et al. [25] also provides a thorough explanation of FPGA-based CNN accelerators. Researchers have built and shown fixed-point parameter representations using OpenCL on a Starter Platform for the OpenVINO Toolkit FPGA. It has been demonstrated that using particular algorithms and the hardware-software co-design process, FPGAs may offer  $\times 10$  energy-delay efficiency when compared to cutting-edge GPUs for deep learning accelerators [26].

### 5.3 *Memristive Deep Neural Network*

With memories that retain data even when the power is off, thanks to the memristor, also known as a memory resistor, it will be possible to create much more energy-efficient computing systems that can be turned on immediately. Memristor provides several advantages over traditional computing systems like CPU and GPU. First, memristors have low power consumption considering that they are two-terminal nonvolatile devices [27]. Second, memristors can be integrated with larger densities and are compatible with CMOS. Third, because of the memristor's nanoscale size, its switching speed is extremely quick. These characteristics make memristor an attractive choice for neuromorphic computing. As the fourth essential component of a circuit, memristors can adjust their resistance (conductance) in response to variations in the applied current or voltage. Similar to how brain synapses adjust to their surroundings through learning. DNN learning and inference will be much improved by the concurrent MAC operations that can be done out inside memory using this in-situ processing. Memtorch is a system that [28] have created that performs this conversion by using an unrolling operation to change DNNS into Memristor-DNNs.

## 6 Conclusion

In this paper, we have presented a review of various deep learning algorithms used in health care, specifically in the detection of thoracic diseases. We also presented a review of various hardware accelerators used in health care specifically with respect to thoracic diseases. From our survey, we can conclude that there are several advantages to using hardware accelerators that run DL algorithms that are configured to give the maximum accuracy when processing images from chest X-rays. Some of the advantages are that the device will have a small factor hence making in portable. Our research work will also concentrate on reduction of power consumed, which is a necessity if the device is to be made portable. Latency of accelerators is higher than in case of GPU cluster. In many cases we may not require low latencies as the image analysis need not be run real-time. The X-rays will be taken, and then the resultant image (filtered to remove noise) will be passed through an algorithm that breaks up the image into segments and feeds only the relevant segment to the DL algorithm. The diagnosis need not be done while the X-rays are being taken so latency may not be a major issue.

The primary issues that will be dealt with why building the hardware accelerators will be (a) processing power requirements (since we have to run DL algorithms even though the latency may be more; they still need the processing power to converge (b) storage requirements (c) power consumption.

Another issue when designing the system will be the large number of images that will be needed to train the algorithm. We will need a large number of healthy and infected chest X-rays to train the DL algorithm. Hardware accelerators that are

specifically designed for diagnosing thoracic diseases are still to be researched. Only when an accelerator can be designed the performance of accelerators can be analyzed and implemented.

## References

1. Loyolamedicine Homepage. <https://www.loyolamedicine.org/find-a-condition-or-service/heart-and-vascular/heart-vascular-conditions/thoracic-disorders/>. Last accessed 20 Oct 2022
2. Abiyev RH, Ma'aitah MKS (2018) Deep convolutional neural networks for chest diseases detection. *J Healthcare Eng* 2018(4168538):11
3. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*
4. LeCun Y, Jackel LD, Bottou L, Cortes C, Denker JS, Drucker H, Guyon I, Muller UA, Sackinger E, Simard P et al (1995) Learning algorithms for classification: a comparison on handwritten digit recognition. *Neural Netw Stat Mech Perspect* 261:276
5. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. *Commun ACM* 60(6):84–90
6. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
7. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. [arXiv:1409.4842](https://arxiv.org/abs/1409.4842)
8. Lin M, Chen Q, Yan S (2013) Network in network. [arXiv:1312.4400](https://arxiv.org/abs/1312.4400)
9. Stollenga MF, Byeon W, Liwicki M, Schmidhuber J (2015) Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: *Advances in neural information processing systems*. pp 2998–3006
10. Baltruschat I, Steinmeister L, Nickisch H, Saalbach A, Grass M, Adam G, Knopp T, Ittrich H (2020) Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *Eur Radiol*
11. DSouza AM, Abidin AZ, Wismüller A (2019) Automated identification of thoracic pathology from chest radiographs with enhanced training pipeline. In: *Medical imaging 2019: computer-aided diagnosis*. SPIE, p 123
12. Wang W, Feng H, Bu Q, Cui L, Xie Y, Zhang A, Feng J, Zhu Z, Chen Z (2020) MDU-Net: a convolutional network for clavicle and rib segmentation from a chest radiograph. *J Healthcare Eng* 2020:1–9
13. Sogancioglu E, Murphy K, Calli E, Scholten ET, Schalekamp S, Ginneken BV (2020) Cardiomegaly detection on chest radiographs: segmentation versus classification. *IEEE Access* 8:94631–94642
14. Yang D, Zhang S, Yan Z, Tan C, Li K, Metaxas D (2015) Automated anatomical landmark detection on distal femur surface using convolutional neural network. In: *IEEE international symposium on biomedical imaging*. pp 17–21
15. Park S, Lee SM, Kim N, Choe J, Cho Y, Do K-H, Seo JB (2019) Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 29(10):5341–5348
16. Kim Y-G, Lee SM, Lee KH, Jang R, Seo JB, Kim N (2020) Optimal matrix size of chest radiographs for computer-aided detection on lung nodule or mass with deep learning. *Eur Radiol* 30(9):4943–4951
17. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Proceedings of the 27th international conference on neural information processing systems*, vol 2. MIT Press, pp 2672–2680

18. Madani A, Moradi M, Karargyris A, Syeda-Mahmood T (2018) Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, pp 1038–1042
19. Gopalan R, Li R, Chellappa R (2011) Domain adaptation for object recognition: an unsupervised approach. In: International conference on computer vision. pp 999–1006
20. Azimi I et al (2017) Hich: hierarchical fog-assisted computing architecture for healthcare IoT. *ACM Trans Embedded Comput Syst* 16(5s):1–20
21. Sahu P, Yu D, Qin H (2018) Apply lightweight deep learning on internet of things for low-cost and easy-to-access skin cancer detection. In: Proceedings of the medical imaging: imaging informatics for healthcare, research and applications, vol 10579, Houston, TX, USA, Art no 1057912
22. Kennedy P (2019) Huawei Ascend 910 provides a NVIDIA AI training alternative. Serve The Home. [Online]. Available: <https://www.servethehome.com/huawei-ascend-910-provides-anv-idia-ai-training-alternative/>
23. Cavigelli L, Benini L (2017) Origami: a 803-GOp/s/W convolutional network accelerator. *IEEE Trans Circuits Syst Video Technol* 27(11):2461–2475
24. Abts D et al (2020) Think fast: a tensor streaming processor (TSP) for accelerating deep learning workloads. In: Proceedings of the ACM/IEEE 47th annual international symposium on computer architecture. pp 3347–3357
25. Stone JE, Gohara D, Shi G (2010) OpenCL: a parallel programming standard for heterogeneous computing systems. *Comput Sci Eng* 12(3):66–73
26. Guo K, Zeng S, Yu J, Wang Y, Yang H (2019) A survey of FPGA based neural network inference accelerator. *ACM Trans Reconfigurable Technol Syst* 12(1):1–26
27. Valentian A et al (2019) Fully integrated spiking neural network with analog neurons and RRAM synapses. In: Proceedings of the IEEE international electron devices meeting, San Francisco, CA, USA, pp 14.13.1–14.13.4
28. Lammie C, Xiang W, Linares-Barranco B, Azghadi MR (2020) MemTorch: an open-source simulation framework for memristive deep learning systems. [arXiv:2004.10971](https://arxiv.org/abs/2004.10971)

# Multi-task Learning Method Using Emoji Prediction as Auxiliary Task for Sentiment Analysis



Haruki Asano and Masafumi Matsuhara

**Abstract** There are text sentiment analysis methods that target a variety of sentiments, such as joy, anger and sadness. This method takes work to perform a highly accurate analysis because creating the dataset is more costly than negative/positive binary targets. Therefore, there is a need for data augmentation in sentiment analysis. Data augmentation in sentiment analysis exists using emojis. By defining the sentiment expressed by emojis, it is possible to automatically annotate a sentiment label to text with emoji. However, conventional methods have a problem in that the definition and weighting of sentiments are subjective. Our method automatically generates training data using emojis and learns it by multi-task learning. This method aims to improve the performance of sentiment analysis by transmitting only helpful information for sentiment analysis.

**Keywords** Sentiment analysis · Emoji · Multi-task learning

## 1 Introduction

Text sentiment analysis is a technique for mechanically identifying sentiments from the text. There are text sentiment analysis methods that target a variety of sentiments, such as joy, anger and sadness. This method takes work to perform a highly accurate analysis because creating the dataset is more costly than negative/positive binary targets. In Japanese, most conventional sentiment analysis methods that target a variety of sentiments are rule-based methods using sentiment word dictionaries. However, the rule-based methods have a problem that it is difficult to identify the context due to the influence of grammatical rules and polysemous words. Since texts contain a variety of sentiments, there are cases in which negative/positive analysis is not enough [7]. Improving the accuracy of sentiment analysis for various sentiments

---

H. Asano (✉) · M. Matsuhara  
Iwate Prefectural University, Iwate, Japan  
e-mail: [g231t002@s.iwate-pu.ac.jp](mailto:g231t002@s.iwate-pu.ac.jp)

M. Matsuhara  
e-mail: [masafumi@iwate-pu.ac.jp](mailto:masafumi@iwate-pu.ac.jp)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_43](https://doi.org/10.1007/978-981-99-3091-3_43)

521

is essential, which requires many datasets and deep learning techniques. However, sentiment is an ambiguous concept. Therefore, sentiment analysis is considered more costly in creating a dataset.

There is some research on the automatic construction of training data for sentiment analysis using emoji-based distant supervision. Distant supervision automatically generates training data by extracting relationships from raw data based on cues obtained from a knowledge base. Emoji-based distant supervision defines sentiments expected to be represented by the emojis. This method then annotates sentiment labels to the text based on the assumption that text with a particular emoji represents the defined sentiment [12]. This method is able to generate many training data at a low cost automatically. However, it is a problem because of noise in the data, subjective elements and ambiguity in interpreting the relationship between emojis and sentiments.

In this paper, we propose a method to improve the accuracy of analyzing various sentiments by appropriately learning automatically generated training data using emojis. Our method generates good-quality training data by defining the sentiments that emojis are supposed to represent using objective indices. In addition, we use multi-task learning to learn the training data appropriately. Multi-task learning is a method for learning multiple tasks jointly. This method improves the accuracy of a particular task or multiple tasks by learning-related tasks jointly.

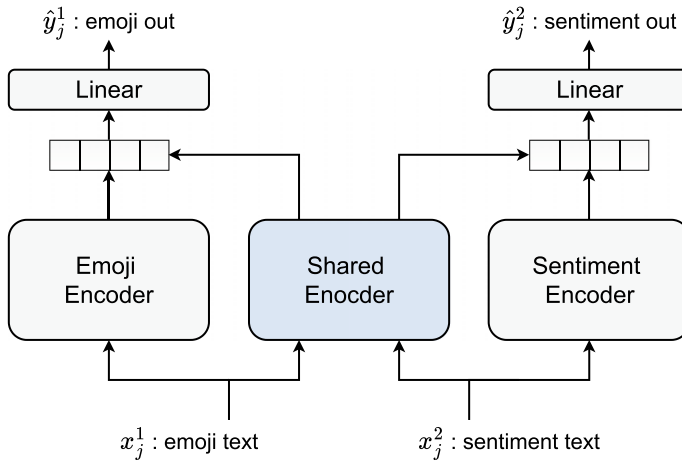
The proposed method defines the sentiments of emoji based on a dataset called Emotag1200 [20]. Emotag1200 was annotated with associations between 150 emojis and eight sentiments (joy, sadness, anticipation, surprise, anger, fear, disgust and trust) by nine raters. We then define the task of predicting the value of Emotag1200 from the text as an emoji prediction task and use this as an auxiliary task for the sentiment prediction task. This process aims to improve the performance by transmitting only helpful information to the sentiment analysis. This method is not affected by the weights of subjectivity or ambiguity of the defined sentiment.

## 2 Related Work

### 2.1 Sentiment Analysis

In Japanese, sentiment word dictionaries have been used for text sentiment analysis targeting various sentiments. There are pattern matching methods based on “感情表現辞典(Emotive expression dictionary)”, and ML-Ask is frequently used [5, 18, 23]. On the other hand, deep learning methods such as transformer has been used for sentiment analysis in English [7, 24, 25]. In English, Plutchik’s eight sentiments have been used as the target of sentiment analysis [9, 19].

Emoji-based distant supervision has been conducted to reduce the cost of labeling. Generally, this method defines the sentiment expressed by an emoji and annotates the sentiment label to text with the emoji [6, 12, 22]. Some research has used transfer



**Fig. 1** Multi-task learning model architecture

learning of weights obtained in predicting emojis without defining the sentiment [2, 4].

## 2.2 Multi-task Learning

Due to the recent spread of deep learning, multi-task learning has been actively used. There are many examples of its application in text classification and sentiment analysis [1, 26]. There are multi-task learning models that separate task-specific space from task-invariant space by adding a task-specific layer in addition to the shared layer [14]. There is also a method called adversarial multi-task learning [15]. It is possible to separate spaces more accurately by adding adversarial loss to the above model. Recently, adversarial multi-task learning has been applied to aspect-based sentiment analysis and humor detection [11, 21].

## 3 Proposed Method

### 3.1 Collecting Data with Emojis

Since emojis are defined as Unicode characters, it is possible to do pattern matching using emojis as queries. Therefore, collecting a large number of texts with emojis

is an easy process. The Twitter API<sup>1</sup> is available to extract text with emojis from a large number of raw text resources. In this paper, we use the Twitter API to collect tweets with the target emojis and use them for learning emoji prediction.

### 3.2 Multi-tasking Learning Setup

Multi-task learning using deep learning is used for emoji and sentiment prediction tasks in our method. The architecture of the proposed model is shown in Fig. 1.

Each task has a training dataset  $\mathcal{D}_t$  consisting of  $n_t$  data samples.  $t$  is the task identifier, 1 for emoji and 2 for sentiment prediction, i.e.,  $\mathcal{D}_t = \{x_j^t, y_j^t\}_{j=1}^{n_t}$ , where  $x_j^t$  is the  $j$ th training instance and  $y_j^t$  is its label. The input shape of the text is  $x_j^t \in \mathbb{R}^l$ ,  $l$  is the sequence length, and each element of the vector is the word identifier. The sequence is padded with arbitrary constants and has a fixed length.

After the text is tokenized, it is input into the shared and task-specific encoder. The output of each task is obtained by fusing the two features and fully connecting them. In sentiment analysis, the meaning of the same word or sentence may differ significantly depending on the domain. For these reasons, it degrades performance if collected data with emojis from a domain different from sentiment prediction. In our method, a task-specific encoder is added to the shared encoder to prevent accuracy degradation due to negative knowledge transfer.

### 3.3 Text Encoding by BERT

Bidirectional Encoder Representations from Transformers (BERT) is used for the encoder [3]. First, the input  $x_j^t$  is embedded into the dimension  $d_e$  and transformed into  $H \in \mathbb{R}^{l \times d_e}$ . After performing positional encoding and sentence type discrimination, calculate query, key and value vectors using Eq. (1).

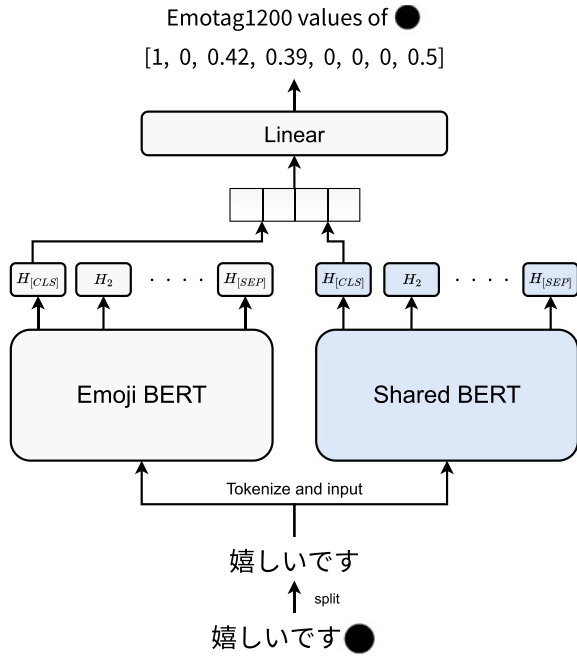
$$\begin{aligned} Q_i &= HW_i^Q \\ K_i &= HW_i^K \\ V_i &= HW_i^V \end{aligned} \tag{1}$$

$i$  represents the index of Attention head,  $W_i^Q \in \mathbb{R}^{d_e \times d_q}$ ,  $W_i^K \in \mathbb{R}^{d_e \times d_k}$ , and  $W_i^V \in \mathbb{R}^{d_e \times d_v}$  is a weight matrix that transforms the dimensions of features. Then, Multi-head Attention is calculated by Eq. (2).

---

<sup>1</sup> <https://developer.twitter.com/en/docs/twitter-api>.



**Fig. 2** Emoji prediction learning

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}\left(\frac{\mathbf{Q}_i^T \mathbf{K}_i}{\sqrt{d_k}}\right) \mathbf{V}_i$$

$$\text{Multi-head}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O \quad (2)$$

where  $\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$

Multiple heads are concatenated and return to the original embedding dimension  $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_e}$ . Multiple regularizations are applied, and the final output is obtained by applying the GELU function.

For the sake of simplicity, we define that the encoding of the above process is repeated for the number of layers as Eq. (3).

$$\mathbf{H} = \text{BERT}(x_j^t) \quad (3)$$

### 3.4 Emoji Prediction Learning

Emojis often have the character of expressing sentiments. Therefore, a task that takes text as input and predicts the contained emoji is expected to provide a knowledge representation of the relationship between text and sentiment.

The architecture of emoji prediction is shown in Fig. 2.

Since we used BERT for the encoder, the beginning of the input text is the [CLS] token and the end of the [SEP] token.

The predicted label is the Emotag1200 value of the contained emoji in the text. Although we define it as emoji prediction, it is a task to predict the sentiment values generated from emojis. The aim is to clarify the mapping between emojis and sentiments by converting emojis to sentiments.

Text  $x_j^1$  is input to the emoji prediction specific encoder and the encoder shared with the sentiment prediction task, respectively, and the hidden states  $\mathbf{H}_{\text{emoji}} \in \mathbb{R}^{l \times d_e}$  and  $\mathbf{H}_{\text{shared}} \in \mathbb{R}^{l \times d_e}$  are calculated.

$$\mathbf{H}_{\text{emoji}} = \text{BERT}_{\text{emoji}}(x_j^1) \quad (4)$$

$$\mathbf{H}_{\text{shared}} = \text{BERT}_{\text{shared}}(x_j^1) \quad (5)$$

After that, the hidden states corresponding to each [CLS] are concatenated to calculate the text feature  $C_{\text{emoji}} \in \mathbb{R}^{2d_e}$ , and the output is obtained by fully connecting text feature.

$$C_{\text{emoji}} = \text{Concat}(\mathbf{H}_{\text{emoji}, [\text{CLS}]}, \mathbf{H}_{\text{shared}, [\text{CLS}]}) \quad (6)$$

$$\hat{y}_j^1 = C_{\text{emoji}} W^1 + b^1 \quad (7)$$

### 3.5 Sentiment Prediction Learning

Sentiment prediction is performed in the same way as emoji prediction. Text  $x_j^2$  is input to the sentiment prediction specific encoder and the encoder shared with the emoji prediction task, respectively, and the hidden states  $\mathbf{H}_{\text{sent}} \in \mathbb{R}^{l \times d_e}$  and  $\mathbf{H}_{\text{shared}} \in \mathbb{R}^{l \times d_e}$  are calculated.

$$\mathbf{H}_{\text{sent}} = \text{BERT}_{\text{sent}}(x_j^2) \quad (8)$$

$$\mathbf{H}_{\text{shared}} = \text{BERT}_{\text{shared}}(x_j^2) \quad (9)$$

After that, the hidden states corresponding to each [CLS] are concatenated to calculate the text feature  $C_{\text{sent}} \in \mathbb{R}^{2d_e}$ , and the output is obtained by fully connecting this.

$$C_{\text{sent}} = \text{Concat}(\mathbf{H}_{\text{shared}, [\text{CLS}]}, \mathbf{H}_{\text{sent}, [\text{CLS}]}) \quad (10)$$

$$\hat{y}_j^2 = C_{\text{sent}} W^2 + b^2 \quad (11)$$

**Table 1** Selected emojis and number of samples

Sentiment	Emoji	Name	Number of samples
Joy	😊	smiling face	531
Sadness	😭	crying face	2,389
Surprise	❗	exclamation mark	2,517
Anger	😡	pouting face	10,252
Fear	😱	fearful face	4,034
Disgust	👎	thumbs down	5,361
Trust	😘	kissing face with smiling eyes	9,799

3.6 Loss Optimization

In multi-task learning, the loss for each task is calculated, and the final loss  $L$  is obtained by adding them together. In our method, the final loss is calculated by Eq. (12).

$$L = \lambda_1 L_e + \lambda_2 L_s \tag{12}$$

Emoji prediction loss  $L_e$  and sentiment prediction loss  $L_s$  are adjusted by the weight parameter of  $\lambda_i$ . Our method transfers the knowledge representation obtained in the emoji prediction task to the sentiment prediction by optimizing the above losses jointly. It is expected that the obtained knowledge will be beneficial to sentiment analysis and improve its performance.

4 Experiment

4.1 Dataset

WRIME [8] is used as the dataset for sentiment prediction. WRIME is a text labeled with eight sentiments, joy, sadness, anticipation, surprise, anger, fear, disgust and trust, in 4 levels (0: none, 1: weak, 2: medium, 3: strong) from subjective and objective viewpoints. In this experiment, we used the average of the three labels annotated objectively and normalized the data to a minimum value of 0 and a maximum value of 1. The training, validation and test sets are those assigned to WRIME, and their numbers are 30,000, 2500 and 2500, respectively.

The emoji prediction dataset is generated by extracting tweets with the target emoji. We selected the emoji with the highest value in each sentiment in Emotag1200 as the representative emoji for this sentiment. Also, we used oversampling for the

sample size to prevent an imbalance in the sentiment categories. Here, “anticipation” was the sentiment with the highest number of labels of 1 or more in the WRIME. Since the number was 11,391, texts with representative emojis except “anticipation” were extracted until 11,391. The selected emojis and the number of samples are shown in Table 1. The number of labels and values for each sentiment is not perfectly equal. This is because the strength of labels is ignored and emojis have values other than representative sentiment.

## 4.2 Comparison Models

In this experiment, the performance of the following models was compared.

- BERT<sub>BASE</sub>: Learning to predict sentiment using WRIME.
- BERT<sub>AUG</sub>: Learning to predict sentiment using WRIME and emoji prediction dataset combined dataset.
- BERT<sub>MTL</sub>: Multi-task learning with a shared layer using WRIME and emoji prediction dataset.
- BERT<sub>SP-MTL</sub>: Multi-task learning with shared and task-specific layers using WRIME and emoji prediction dataset (proposed method).

BERT<sub>AUG</sub> is a model trained to predict sentiment values by combining WRIME and an emoji prediction dataset. Since the emoji prediction dataset is converted to the same category of sentiment values as WRIME by Emotag1200, it is possible to be combined. Unlike models with multi-task learning, this model learns the labels of the emoji prediction dataset as if they were WRIME. In WRIME, the label  $y$  in  $[0, 3]$  is normalized to  $[0, 1]$ , and Emotag1200 is similarly  $[0, 1]$ . Therefore, the correspondence between the label  $y$  and  $y'$  in Emotag1200 is  $y = 3y'$ .

We use the Japanese BERT pre-training model<sup>2</sup> published by Inui Laboratory at Tohoku University for all models.

## 4.3 Learning Setup

The BERT used for the encoder has 12 layers, 768 hidden state dimensions, 12 attention heads and a sequence length of 64. AdaBound is used for the optimizer [16]. Its initial learning rate is  $1 \times 10^{-5}$ , and the final learning rate is  $1 \times 10^{-2}$ . The weight parameters  $\lambda_1$  and  $\lambda_2$  used for the learning loss of BERT<sub>MTL</sub> and BERT<sub>SP-MTL</sub> are set to 0.4 and 0.6, respectively, to ensure uniform learning convergence. The output is an identity function for emoji prediction and a sigmoid function for sentiment

<sup>2</sup> <https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking>.

prediction. The loss functions for emoji and sentiment prediction are mean squared error and binary cross-entropy, respectively.

4.4 Evaluation Metrics

Mean absolute error (MAE), P@k, R@k and F@k, are used for evaluation. In multi-label classification, each instance is often assigned only a small number of labels, in contrast to the dimensionality of the labels. Therefore, a rank evaluation is used for the top few instances [13].

In the WRIME, essential sentiments are considered the top one to three. The MAE is used to measure the overall prediction error and the rank metrics to measure the discriminative performance of the essential sentiments.

4.5 Performance Evaluation

The MAE for each model is shown in Table 2. Table 2 shows that BERT<sub>MTL</sub> has the lowest MAE on average. BERT<sub>SP-MTL</sub> is slightly inferior to BERT<sub>MTL</sub> but has a minor error than BERT<sub>BASE</sub>. The results suggest that the knowledge representation obtained from emoji prediction contributes to sentiment analysis in multi-task learning.

On the other hand, BERT<sub>AUG</sub> showed a significant increase in error compared to the other models. This result indicated the possibility of inappropriate label weight between WRIME and emoji prediction datasets and over-fitting to noise in the emoji prediction data. Although tuning the label weights may improve performance, multi-task learning automatically performs noise filtering and scaling. Multi-task learning transfers only helpful knowledge for sentiment analysis and is more effective than tuning the label.

The P@k, R@k and F@k for each model are shown in Table 3. Table 3 shows that the rank metrics of the models with multi-task learning have all improved compared to BERT<sub>BASE</sub> and BERT<sub>AUG</sub>. Not only the overall prediction error but also the ability to identify essential sentiments was improved by multi-task learning. The performance of BERT<sub>AUG</sub> also decreased in the rank evaluation, which may be the exact reason for the increase in MAE.

Table 2 MAE for each sentiment category

Model	Joy	Sadness	Anticipation	Surprise	Anger	Fear	Disgust	Trust	Ave
BERT <sub>BASE</sub>	0.152	0.109	0.144	0.135	<b>0.017</b>	0.098	0.074	0.041	0.096
BERT <sub>AUG</sub>	0.162	0.124	0.150	0.147	0.058	0.117	0.099	0.071	0.116
BERT <sub>MTL</sub>	<b>0.149</b>	<b>0.098</b>	0.141	<b>0.129</b>	0.018	<b>0.095</b>	0.075	0.040	<b>0.091</b>
BERT <sub>SP-MTL</sub>	0.150	0.100	<b>0.136</b>	0.130	<b>0.017</b>	0.098	<b>0.073</b>	<b>0.038</b>	0.093

いろいろ言っちゃう時あるけど根底にあるのはえびの素晴らしさ知らないの勿体ないっていう  
(Sometimes I say many things but at the root of it all,  
the idea is that it would be a waste not to know the wonders of shrimp.)

Fig. 3 Texts with the most significant differences in F@2 between the models

Table 3 The P@k, R@k, and F@k for each model is shown

model	P@1	P@2	P@3	R@1	R@2	R@3	F@1	F@2	F@3
BERT <sub>BASE</sub>	0.698	0.572	0.467	0.443	0.672	0.790	0.542	0.618	0.587
BERT <sub>AUG</sub>	0.678	0.535	0.446	0.427	0.627	0.756	0.524	0.577	0.561
BERT <sub>MTL</sub>	<b>0.734</b>	<b>0.585</b>	<b>0.476</b>	<b>0.465</b>	<b>0.688</b>	<b>0.807</b>	<b>0.569</b>	<b>0.633</b>	<b>0.599</b>
BERT <sub>SP-MTL</sub>	0.724	0.581	0.471	0.458	0.680	0.799	0.561	0.626	0.593

In this experiment, BERT<sub>MTL</sub> performed better overall than BERT<sub>SP-MTL</sub>. The dataset used in this experiment consisted of text generated from SNS for both emoji and sentiment predictions. Therefore, the model with shared hard parameters is more suitable. However, MAE decreased for anticipation, anger, disgust and trust compared to BERT<sub>MTL</sub>. It is suggested that the proposed model performed better for the specific sentiments.

It is assumed that emoji data from other domains will be used when collecting them from the same domain is impossible. In the case of multi-task learning with data from different domains, BERT<sub>SP-MTL</sub> is expected to perform more appropriate learning and improve performance. In the future, we need to evaluate the proposed method in case where the domains of emoji data and sentiment analysis data are different.

4.6 Analysis of Predicted Values

Predictions were analyzed to examine changes in features due to multi-task learning. In order to make it easier to understand the feature values, we extracted the results for BERT<sub>MTL</sub> instead of BERT<sub>SP-MTL</sub> and compared them with those of BERT<sub>BASE</sub>. The texts with the most significant differences in F@2 between the models are shown in Fig. 3. Figure 3 shows the texts in the WRIME test data where the F@2 of BERT<sub>MTL</sub> was higher than that of BERT<sub>BASE</sub> and where the difference was the largest. Prediction results for the text of Fig. 3 of each model are shown in Table 4. The true set is {anticipation, fear}, and the top two predicted sentiments are {sadness, disgust} for BERT<sub>BASE</sub> and {anticipation, fear} for BERT<sub>MTL</sub>. The result confirms that there are samples in the test data for which multi-task learning is able to improve the F@k.

The Self Attention of each model when predicting the text is shown in Fig. 4. The top one is for BERT<sub>BASE</sub>, and the bottom one is for BERT<sub>MTL</sub>, showing the Self Attention calculated using [CLS] as the query vector when predicting the text in Fig. 3. Since the hidden states corresponding to [CLS] are used for text features, the

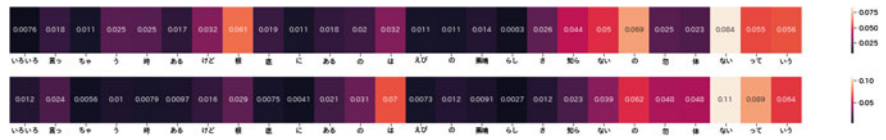


Fig. 4 Self Attention for each model when predicting Fig. 3 text

Table 4 Prediction results for text of Fig. 3

Model	True set	Prediction set	F@2
BERT <sub>BASE</sub>	{anticipation, fear}	{sadness, disgust}	0
BERT <sub>MTL</sub>		{anticipation, fear}	1.0

結局昨日はジムで汗を流し、そのまま飲みに行き、カラオケで暴れ、精も根も尽き果てて眠りに落ちた。喉がいたい...  
(After all, yesterday, I worked out at the gym, went straight to the bar, went on a karaoke binge, and fell asleep after running out of energy. My throat is aching.)

久々真面目にお稽古に行こうと思ったら、道具が根こそぎない。  
久々でそれはダメだろうってと躍起になってさがしてみたが、ない。  
気づけばもう30分遅刻している。ほんと、どーしょーもねーなあ  
(After a long time, when I thought about going to practice seriously, the tools were not by the roots. I thought that was a bad idea and searched diligently for it, but it was not there.  
I realized that I was already 30 minutes late. I really can't help it.)

Fig. 5 Negative samples with “ 根 ” exist in the WRIME training set

above values are considered to represent the contribution of words to the prediction. Figure 4 shows that the word “ 根 ” (root) has a high Attention in BERT<sub>BASE</sub>, while it decreases in BERT<sub>MTL</sub>. Negative samples with “ 根 ” shown in Fig. 5 were also founded from the WRIME training set. This result suggests that BERT<sub>BASE</sub> may have interpreted the word “ 根 ” as a word with strong negative sentiment due to overtraining on the dataset, resulting in enhanced sadness and disgust. On the other hand, the multi-task learning process allowed BERT<sub>MTL</sub> to learn a wide variety of sentiment expressions and to broaden its interpretation of the word “ 根 ”. As a result, BERT<sub>MTL</sub> was able to output positive sentiments such as “anticipation”. We confirm the benefit of the change in features due to multi-task learning, and BERT<sub>SP-MTL</sub> also benefits from the same. However, since BERT<sub>SP-MTL</sub> have feature changes due to adding task-specific layers, this should also be considered in the future.

5 Conclusion

In sentiment analysis, creating a dataset is costly due to ambiguity in interpretation. In this paper, we proposed a method that automatically generates training data using

emojis and learns it by multi-task learning. Specifically, we established task-specific BERT and BERT shared among tasks and combined the features obtained from each encoder to construct a model that predicts sentiment values and emojis.

Experimental results showed a decrease in MAE and an increase in F@k for the proposed method. They indicated that emoji prediction is a practical auxiliary task for improving the performance of sentiment analysis. The performance of the proposed method and a simple multi-task learning model using only the shared layer was almost the same. However, the MAE of anticipation, anger, disgust and trust were reduced, and the performance of the proposed method was improved for specific sentiments. The advantages of the proposed method will be beneficial when emoji prediction and sentiment prediction are in different domains from each other.

We plan to reevaluate the proposed method in case where the domains of emoji data and sentiment analysis data are different.

**Acknowledgements** This work was supported by JSPS KAKENHI Grant Number 21K12611.

## References

1. Bingel J, Søgaaard A (2017) Identifying beneficial task relations for multi-task learning in deep neural networks. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics, Valencia, Spain, pp 164–169
2. Boy S, Ruiter D, Klakow D (2021) Emoji-based transfer learning for sentiment tasks. In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: student research workshop, pp 103–110. <https://doi.org/10.18653/v1/2021.eacl-srw.15>
3. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
4. Felbo B, Mislove A, Søgaaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, Denmark, pp 1615–1625. <https://doi.org/10.18653/v1/D17-1169>
5. Fujita S, Takano K (2020) A method for content recommendation using user's emotions derived from contents. *J Inform Process (JIP)* 61(6):1200–1209
6. Go A, Bhayani R, Huang L (2019) Twitter sentiment classification using distant supervision, vol 1(12). CS224N project report
7. Irtiza Tripto N, Eunus Ali M (2021) Detecting multilabel sentiment and emotions from Bangla YouTube comments. In: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pp 1160–1168. <https://doi.org/10.18653/v1/2021.semeval-1.163>
8. Kajiwara T, Chu C, Takemura N, Nakashima Y, Nagahara H (2021) WRIME: a new dataset for emotional intensity estimation with subjective and objective annotations. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 2095–2104. <https://doi.org/10.18653/v1/2021.naacl-main.169>
9. Kant N, Puri R, Yakovenko N, Catanzaro B (2018) Practical text classification with large pre-trained language models. arXiv



10. van der Laurens M, Geoffrey H (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(86):2579–2605
11. Liang B, Yin R, Gui L, Du J, He Y, Xu R (2020) Aspect-invariant sentiment features learning: adversarial multi-task learning for aspect-based sentiment analysis. In: *Proceedings of the 29th ACM international conference on information and knowledge management*, New York, USA, pp 825–834
12. Lin Y, Cui H, Utsuro T (2020) Utilizing emoji in collecting training instances of a model for sentiment analysis of tweets. *Theor Intell Inform* 32(5):923–933. [https://doi.org/10.3156/jsoft.32.5\\_923](https://doi.org/10.3156/jsoft.32.5_923)
13. Liu J, Chang W, Wu Y, Yang Y (2017) Deep learning for extreme multi-label text classification. In: *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, Tokyo, Japan, pp 115–124. <https://doi.org/10.1145/3077136.3080834>
14. Liu P, Qiu X, Huang X (2016) Recurrent neural network for text classification with multi-task learning. In: *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, pp 2873–2879. <https://doi.org/10.5555/3060832.3061023>
15. Liu P, Qiu X, Huang X (2017) Adversarial multi-task learning for text classification. In: *Proceedings of the 55th annual meeting of the association for computational linguistics*, Vancouver, Canada, pp 1–10. <https://doi.org/10.18653/v1/P17-1001>
16. Luo L, Xiong Y, Liu Y, Sun X (2019) Adaptive gradient methods with dynamic bound of learning rate. In: *Proceedings of the 7th international conference on learning representations*, New Orleans, Louisiana
17. Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant supervision for relation extraction without labeled data. In: *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP*, Suntec, Singapore, pp 1003–1011
18. Ptaszynski M, Dybala P, Rzepka R, Araki K, Masui F (2017) ML-Ask: open source affect analysis software for textual input in Japanese. *J Open Res Softw* 5(1):16
19. Robert P, Henry K (1980) Chapter 1—A general psychoevolutionary theory of emotion. In: *Theories of emotion*. Academic Press
20. Shueb AAM, de Melo G (2020) EmoTag1200: understanding the association between emojis and emotions. In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp 8957–8967. <https://doi.org/10.18653/v1/2020.emnlp-main.720>
21. Smádu R, Cercel D, Dascalu M (2021) UPB at SemEval-2021 Task 7: adversarial multi-task learning for detecting and rating humor and offense. In: *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pp 1160–1168. <https://doi.org/10.18653/v1/2021.semeval-1.163>
22. Suttles J, Ide N (2013) Distant supervision for emotion classification with discrete binary values. In: *International conference on intelligent text processing and computational linguistics*, pp 121–136
23. Toriumi F, Sakaki T, Yoshida M (2020) Social emotions under the spread of COVID-19 using social media. *Trans Jpn Soc Artif Intell AI* 35(4):F-K45-1-7. <https://doi.org/10.1527/tjsai.F-K45>
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser U, Polosukhin I (2017) Attention is all you need. In: Guyon I et al (eds) *Advances in neural information processing systems*, vol 30(NIPS 2017). Curran Associates, Inc
25. Ying W, Xiang R, Lu Q (2019) Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In: *Proceedings of the 5th workshop on noisy user-generated text (W-NUT 2019)*, Hong Kong, China, pp 316–321. <https://doi.org/10.18653/v1/D19-5541>
26. Zhang J, Yan K, Mo Y (2021) Multi-task learning for sentiment analysis with hard-sharing and task recognition mechanisms. *Information* 12(5):2078–2489

# Smart Cities Improving Government Management Systems with Blockchain Technology



Marciele Berger Bernardes, Francisco Pacheco de Andrade,  
and Lucas Cortizo

**Abstract** The confluence of two phenomena, accelerated dissemination of information and communication technologies (ICTs)—digital transformation, associated with the increase of population in large urban centers converge to the emergence of the phenomenon that is conventionally called smart cities. Smart cities are understood as those that employ ICTs to optimize their services, ensuring not only a more efficient and effective governance, but also an improvement in the quality of life of their citizens. And, in this ICT-dominated environment, a digital interconnection of everyday objects with the Internet (“Internet of Things” or IoT) is created. In this sense, from the literature review, this paper seeks to explore to what extent the use of blockchain technology would succeed in being the secure governance infrastructure in the development of smart cities filled by the connected devices. The study explains how blockchain creates an information registration system which allows network participants to exchange data with a high degree of reliability and transparency without the need for a centralized management. And given the characteristics of blockchain, this study aims to analyze how this technology can contribute to improving governance in smart cities, toward efficient urban management, and namely to improvement the Sustainable Development Goals (SDGs). The results achieved with the case studies allowed us to identify the limits and possibilities concerning the use of blockchain for the governance of cities. Thus, it can be concluded that despite the associated limits, such as ethical use problems and sustainability issues, blockchain has the potential to function as the integral infrastructure of digital ecosystems that may be employed in smart cities.

**Keywords** Smart cities and governance · Blockchain · Limits and possibilities

---

M. B. Bernardes (✉) · F. P. de Andrade · L. Cortizo  
Escola de Direito, Universidade do Minho, Braga, Portugal  
e-mail: [marcieleberger@gmail.com](mailto:marcieleberger@gmail.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_44](https://doi.org/10.1007/978-981-99-3091-3_44)

535

## 1 Introduction

Society is still experiencing another of its successive revolutions, the so-called 4th industrial revolution [1], marked by the intensive use of the Internet, in a wide spectrum of areas: artificial intelligence, robotics, Internet of Things, 3D printing, nanotechnology, and quantum computing, to name just a few. We live in the data era, insofar as the most valuable companies in the world are those related to the technology sector, such as Facebook and Google [2]. But the question that remains is how do they manage to be the most valuable companies in the world without charging for their services? Precisely, because data is the fuel, or the new oil of this era [3], the use of data by companies and governments allows them to act as platforms, benefiting life in cities, in terms of efficiency, sustainability, and quality of life. Moreover, the United Nations Organization has issued the warning that by 2050, more than 70% of the world's population will live in cities [4].

This is the scenario in which the smart cities paradigm develops. In this paradigm, the various stakeholders (government, business, society, and universities), when dialoguing regarding the use of the Internet to promote the sustainable development of cities, propel the helixes of innovation. But how to use technologies in a safe way to really deliver value services to people?

Seeking answers to this question, the focus of this paper is on the study of blockchain (an information registration system in which network participants exchange data securely without the need for a centralized administrator), which in return promises to create solutions for the enabling pillar of smart cities—connectivity infrastructure—to deliver secure services, enabling smart governance, in attention to the guidelines of sustainable urban development.

This work presents the reasons why blockchain allows for efficient and secure operations over the internet, which are essential for smart city projects.

Therefore, the main objective of this work is to perform a literature review aiming to analyze how the blockchain technology can contribute to improve governance in smart cities, toward an efficient urban management. To this end, the study was structured in four topics: the first presents the concept of urban digitization—within the Smart Cities paradigm elucidating among the multiple dimensions why smart governance is a transversal dimension responsible for orchestrating the others; the second section, the core of the matter, presents blockchain technology (which can ensure transparency and security in data recording); in the third section the possibilities and limits of blockchain use are raised; finally, it presents a synthesis and a strategic vision on the theme proving that it is possible to use technological tools and create relationships based on transparency and security, rescuing the citizen's trust in public management.

## 2 Smart Cities and Governance

Defining smart cities is a complex task, since there is no consensus about their conceptual core, being defined by a list of characteristics extracted from the elements that compose them [5]. For this reason, it is possible to find in the literature several denominations, such as Digital Cities [6], Knowledge-Based Urban Development [7], Smart Cities [8, 9], Smart and Human Cities [10], Responsive Cities [11], Distributed City [12], Smart and Sustainable Cities [13].

It is important to mention that the first milestone to address the issue of smart cities dates to the Kyoto Protocol, which, with the aim of promoting sustainable development, stimulated several governmental actions toward smart growth [14].

Since then, the matter has evolved from the initial stage, associating the term with the intensive use of technological systems by cities (in the transportation, lighting, and security sectors, for example); until reaching a current stage in which technologies are employed as tools to optimize systems and deliver valuable services to citizens, focusing on improving people's quality of life [15].

In this perspective, [16] identifies what smart cities are from six “model situations”. The first focuses on the concept of public safety; the second places new technologies at the center of responses to social problems. The third is linked to the expansion of digital platforms. On the other hand, the fourth is articulated with ‘civilizational’ agendas, making a greater call for citizen participation. In the fifth situation type, the idea of smart city corresponds to the importation of ‘smart solutions’ (one size fits all) without strategic framing or adequacy. Finally, the sixth “situation-type” focuses on cities that, faced with a lack of resources, bet on low-tech solutions, software, and digital applications to the detriment of large investments in infrastructure.

It can be observed that the common feature of the situations above is the use of technologies as a tool to improve management; however, each state will adopt a “type” of smart city, sometimes more technological, sometimes more anthropological.

It should also be noted that the construction of smart city projects involves key elements. In this sense, the Quadruple Helix (QH) innovation approach—“Technology and Data”; “Government”; “Society” and “Physical Environment”—can be used as a basis for co-production of smart city projects to better capture their impacts [17].

Depending on the meaning of the word “smart”, its dimensions are also extracted in the academic discourse and international documents, from five to eight dimensions of Smart Cities are identified [18]. One of the pioneer studies on the subject smart cities understands them as a paradigm composed of six major dimensions: transport and mobility, life, government, economy, people, and development [9]. Considering the objectives of this work, the dimension “Smart Governance” will be approached as a prominent element.

In the study of [19], identify defining elements, fundamental to smart governance, which are grouped into three major groups: (1) the use of a technology (smart ICT); (2) organizational processes (smart collaboration and participation, smart internal

management, smart decision-making, and smart administration); and (3) desired outcomes (smart results). Thus, following the group that adopts “organizational processes” as the defining element of governance, smart governance is believed to be that which seeks effective, efficient, transparent, and more collaborative management.

As can be seen, in the context of smart cities, the “smart governance” dimension unveils itself as a transversal element in all governmental actions and involves gains, especially in actions related to transparency/access to information; open data; data protection; and popular participation.

However, one cannot forget the risks associated with the insertion of technologies in the governance of cities, namely treating the city as a cognizable machine, rather than a complex system; promoting technical solutions (topdown), rather than citizen-centric solutions; treating cities as generic markets, promoting (one-size-fits-all) solutions, rather than recognizing the needs of each city; portraying deployed technologies as being a goal of cities and politically benign, rather than unveiling that they reflect the views and values of their developers and stakeholders; promoting privatization of city services; prioritizing investments in vested interests, reinforcing inequalities and levels of control, rather than creating a more socially just and equal society [20].

For all that, [21] argue that the rise of blockchain technology, as a transparent and accountable mechanism to protect data, is paving the way to solve serious challenges associated with the employment of ICTs in city governance, among them privacy, security, and data integrity. This is the approach that follows.

### 3 Blockchain Technology

In this step, it is necessary to develop the concept of blockchain, to understand how its use can offer more transparency and responsibility in the use of ICTs, based on the technological particularities that compose it. Considering that the processing of personal data is part of the dynamics of smart cities and that governance is a fundamental part of the process, blockchain offers specific designs that can be used in the design of smart city infrastructure.

As mentioned, there are serious challenges associated with the use of ICTs in the governance of smart cities, including the privacy of citizens inserted in this context, the security of the information exchanged, and the integrity of the data that constitute the various systems.

And in this sense, blockchain offers an information registration system, whose fundamental characteristic is to base data exchanges on a distributed ledger among several users so that decentralization eliminates the need for a trusted third party, generating more reliability and transparency in the operations to be performed.

In the study of Buterin et al. [22], explain that the very structure of the blockchain has already been developed to create an intrinsic kinship between all the blocks in the chain. There are functions in the block with the purpose of proving that a given recorded transaction is part of that specific block and simultaneously marking the

block header with the hash code of the previous block, interconnecting the blocks with each other, and creating the kinship relationship.

Not enough, there will be a timestamp: stamping time of when the block was validated [23] elucidates that timestamps serve to increase the difficulty of the validation process, preventing an attacker from manipulating the blockchain, given that records will necessarily be made in a chronological manner.

Without prejudice to other particularities of this technology [24] summarize the blockchain to a temporal and authenticated record of the activity of a network, after all the participants of this network follow predefined steps and will enter data in a linearly organized manner.

That is why blockchain can be understood as a means of offering security in the registration of data, coming from its own structure and not coming from trust in a third entity. And in the dynamic nature of smart cities, this structure is deemed to offer a prosperous prospect, as will be detailed below.

### ***3.1 Blockchain and Smart Cities***

Blockchain can contribute in the development of smart cities or smart towns, mainly to develop urban mobility [25]. According to these authors, advances in distributed ledger technologies can assist in the development of various industries, such as autonomous cars, smart urbanization, and smart mobility.

In fact, challenges emerge in this context and blockchain needs to be envisaged from the most diverse perspectives, since its implementation in smart cities is not limited to the everyday use of ICTs. The technology should also be thought of to tackle broader urban problems, such as climate change and social development [25] follow an interesting reasoning regarding a specific use of blockchain to generate more integration of the various elements that make up the complexity of smart city systems and processes. However, it cannot be forgotten that mobility is only one aspect of smart cities.

As explained earlier, promoting sustainable development is the initial milestone of the smart city theme [14]. Therefore, the adoption of blockchain may bring several benefits regarding transparency and security, also in addition to the concept of sustainable development that should be considered by governmental actions.

It is known that some systems that use blockchain—for example, the cryptocurrency industry—abuse energy expenditure. Truby [26] explains that energy waste is a relevant problem, but there are solutions to “decarbonize” the use of blockchain. According to the author, the process depends heavily on the adoption of laws and policies aimed at sustainable development in the various uses of blockchain. The problem of waste is also discussed in [27], whose contribution is to understand the change in block insertion protocols as one of the causes to avoid energy expenses.

Furthermore, smart cities create true decentralized ecosystems, as explained by Pustišek et al. [28]. In this context of decentralized environments, it is possible to

create strategic priorities to prioritize the overall benefit of the community over the inordinate profit of a few participants.

For example, smart cities that use interconnected devices (the Internet of Things, or IoT). According to Pustišek et al. [28], the IoT market is multifaceted. Different entities are involved and can be both service providers and consumers at the same time. And in the face of this complexity, blockchain-based systems can promote more transparency and security in the exchange of information as, for instance, it is possible to impose authentication and authorization on participants via the blockchain. This works as a technique to prevent attacks and unwanted access in this hyper-connected environment.

When thinking that such smart environments—replete with IoT devices may be embedded in hyperconnectivity, there is yet another meeting point between smart cities and a possible practical implication of the use of blockchain. For Calvo [29] has a work focused on the ethics of smart cities, the author intently glimpses moral implications of “hyperconnectivity, algorithmization, and the datafication of urban digital Society”. Such author also sees smart cities as the result of a confluence of various institutes that include artificial intelligence, IoT, and Big Data. And in this context, there will be a dichotomy as hyperconnectivity increases information exchange, but also increases risks.

Therefore, hyperconnectivity favors unregulated IoT use and may cause irreversible social problems in smart cities. That is why the adoption of blockchain in smart cities should also consider ethical questions about hyperconnectivity, responsible use of algorithms, and monetization of the use of personal data. In fact, blockchain offers a relevant flexibility that can be adjusted to the most diverse uses and multifaceted environments [30].

In addition, the attention should be drawn to the adoption of the 2030 Agenda for Sustainable Development [31]. In the UN report overview for “Blockchain for cities support to the SDGs” [32; pg. 64], Blockchain for Cities-B4C can contribute projects to the SDGs, for example, to provide objectives to monitor sustainable development goals.

In this sense, SDG11 has the potential to transform cities into more inclusive, safe, resilient, and sustainable cities. The use of blockchain for these cases can also consolidate the idea of sustainable cities to reduce poverty (SDG1), can contribute to the effort toward good health and well-being (SDG3), decent work and economic growth (SDG8), peace, justice and strong institutions (SDG16), and the partnership for the goals (SDG17).

In addition, blockchain can also support improvement in terms of SDG12 through product origin tracking application, which is a critical solution for responsible consumption. On the other hand, probably the use of blockchain for smart cities may not reduce inequalities (SDG 10), as well as it probably will not work against energy consumption (SDG 7).

For the UN study, it is not clear whether the use of blockchain can contribute positively to (SDG 9) associated with industry, innovation, and infrastructure. Finally, it should be borne in mind that other challenges persist across the world, as there is no one-size-fits-all approach when it comes to the practical use of such technology.

## 4 Final Considerations

This paper presented possible ways to improve the administration of cities by delivering efficient and secure services, enabling smart governance, in attention to the guidelines of sustainable urban development, and mainly to improvement the Sustainable Development Goals (SDGs).

It was clear that, in an urban scenario where governance is fundamental for the management of the various challenges to be faced in the daily life of cities, it is of importance to invest in the implementation of an infrastructure such as blockchain for the delivery of innovative public services, within the Smart City paradigm. The implementation of this type of infrastructure is relevant as much as investments are needed in other infrastructures (transportation, energy, sanitation, and security) that enable a balanced urban life and a competitive and sustainable economy.

The survey found that the perpetuity of information in blockchain and the unlikelihood of undue changes or deletions make blockchain to offer good solutions to the problems of smart cities, both with regard to secure storage of information and data exchange, and also in the authentication of users. Furthermore, blockchain has the potential to function as the integral infrastructure of the digital ecosystems that will be created.

The potential benefits of this kind of investment reach the public power—having an effective and efficient public governance—blockchain has worked in the complexity of uses related to cryptocurrencies, in addition to the use in several sectors (such as supply chain and data storage) and being deployed as the basis of the future “metaverse”. Hence, it would not be far off to think that blockchain technology can be adapted for various smart city challenges.

Whatsoever the use may be, broader issues emerge as smart city challenges that must be considered when implementing blockchain in this arena. From ethical issues to excessive energy consumption need to be discussed, since smart cities need to have sustainable development as the basic scope of their development.

**Acknowledgements** Our thanks to the JUSGOV—Research Centre for Justice and Governance, at University of Minho, project Smart Cities and Law, E.Governance and Rights: Contributing to the definition and implementation of a Global Strategy for Smart Cities (NORTE-01-0145-FEDER-000063), IP Isabel Fonseca, Centre for Justice and Governance (JusGov), Law School of University of Minho, Portugal.

## References

1. Schwab K (2017) A quarta revolução industrial. Levoir, Portugal
2. Brandfinance. Global 500 2022 Ranking (2022) The annual report on the world's most valuable and strongest brands. Available in: <https://brandirectory.com/download-report/brand-finance-global-500-2022-preview.pdf>. Accessed on: 2022 Apr
3. Mayer Schönberger V, Cukier K (2013) Big data: como extrair volume, variedade, velocidade e valor da avalanche bde informação cotidiana. Elsevier, São Paulo, p 130



4. United Nations—UN, World Urbanization Prospects—The 2018 Revision (2018). Available in: <https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf>. Access on: 2022 Apr
5. Van Bastelaer B (1998) Digital cities and transferability of results. pp 1–15
6. Lemos A (2007) Cidade digital. Portais, inclusão e redes no Brasil. EDUFBA, Salvador
7. Yigitcanlar T, Velibeyoglu K (2008) Knowledge-based Urban development: the local economic development path of Brisbane, Australia. *Local Econ* 23(3):195–207
8. Cohen B What exactly is a smart city? Available in: <https://www.fastcompany.com/1680538/what-exactly-is-a-smart-city>
9. Giffinger R et al (2007) City-ranking of European medium-sized cities. Centre of Regional Science, Vienna UT
10. Prado A et al (2016) Smartness that matter: towards a comprehensive and human-centred definition of smart cities. *J Open Innovation Technol Market Complex* 1–13
11. Goldsmith S, Crawford S The responsive city: engaging communities through data-smart governance. New York, [s.n.]
12. Vassão CA (2010) Metadesign: ferramentas, estratégias e ética para a complexidade. Bluche, São Paulo
13. Estevez E, Lopes NV, Janowski T (2016) Smart sustainable cities. Reconnaissance study. [s.l.] International Development Research Center (IDRC)
14. United Nations (1997) Protocolo de Kyoto de la convención marco de las naciones unidas sobre el cambio climático
15. Bernardes MB (2019) Cidades Inteligentes: Proposta de modelagem regulatória para a governança participativa, no contexto lusobrasileiro. Universidade do Minho
16. Ferrão JA (2021) Cidade contra a Pandemia: mais smart ou pós-smart? Smart cities, cidades sustentáveis
17. Gil-Garcia JR, Pardo TA, Nam T (2015) What makes a city smart? Identifying core components and proposing an integrative and comprehensive conceptualization. *Inf Polity* 20(1):61–87
18. Anthopoulos LG (2015) Understanding the smart city domain: a literature review. In: Transforming city governments for successful smart cities, pp 9–18
19. Bolívar MPR, Meijer AJ (2016) Smart governance: using a literature review and empirical analysis to build a research model. *Soc Sci Comput Rev* 34(6):673–692. <https://doi.org/10.1177/0894439315611088>
20. Kitchin R (2014) The real-time city? Big data and smart urbanism. *GeoJournal* 79(1):1–14
21. Sun J, Yan J, Zhang KZK (2016) Blockchain-based sharing services: what blockchain technology can contribute to smart cities. *Financ Innovation* 2(1):1–9
22. Buterin V et al (2014) A next-generation smart contract and decentralized application platform. *Ethereum White Paper* 3(37)
23. Szalachowski P (2018) Towards more reliable bitcoin timestamps. In: Crypto valley conference on blockchain technology (CVCBT). IEEE, pp 101–104
24. Christidis K, Devetsikiotis M (2016) Blockchains and smart contracts for the Internet of Things. *IEEE Access* 4:2292–2303
25. Cvar N, Trilar J et al (2020) The use of IoT technology in smart cities and smart villages: similarities, differences, and future prospects. *Sensors* 20(14)
26. Truby J (2018) Decarbonizing bitcoin: law and policy choices for reducing the energy consumption of blockchain technologies and digital currencies. *Energy Res Soc Sci* 44:399–410
27. Saleh F (2020) Blockchain without waste: proof-of-stake. SSRN 3183935
28. Pustišek M et al (2022) Technology and applications for industry 4.0, smart energy, and smart cities. Walter de Gruyter GmbH, Berlin/Boston
29. Calvo P (2020) The ethics of Smart City (EoSC): moral implications of hyperconnectivity, algorithmization and the datafication of urban digital society. *Ethics Inf Technol* 22(2):141–149
30. Suliman A (2019) Monetization of IoT data using smart contracts. *IET Netw* 8:32–37. <https://doi.org/10.1049/iet-net.2018.5026>

31. United Nations—UNSDG (2022) Data privacy, ethics and protection: guidance note on big data for achievement of the 2030 agenda. Available in: <https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda>. Access on: 2022 Nov
32. United For Smart Sustainable Cities-U4SSC (2020) Blockchain for smart sustainable cities. Geneva. Available in: [https://www.itu.int/dms\\_pub/itu-t/opb/tut/T-TUT-SMARTCITY-2020-54-PDF-E.pdf](https://www.itu.int/dms_pub/itu-t/opb/tut/T-TUT-SMARTCITY-2020-54-PDF-E.pdf). Access on: 2022 Nov

# Simple Moving Average (SMA) Investment Strategy During COVID-19 Pandemic



Juan P. Licona-Luque, Luis F. Brenes-García, Francisco J. Cantú-Ortiz,  
and Héctor G. Ceballos-Cancino

**Abstract** During COVID-19 pandemic, the financial market experienced an increase in its volatility, making it riskier for most investors. Although many conservative financial experts advise against trading, some investors trade using mathematical prediction models in an attempt to beat the market. A very popular forecasting model is the simple moving average (SMA). This paper intends to answer two main questions. The first one is whether using the SMA to determine when to sell and buy stock over an extended period of time brings overall positive returns to the investor. The second question would be: Is this method better than the buy and hold strategy? Which one brings the optimal returns? A computational implementation of the SMA is proposed to answer these questions by running a back-testing simulation of an investor using SMA against an investor using the buy and hold method. This simulation runs from February 2020 through March 2022, and it is constrained by only taking into account the companies that performed best and worst on the S&P 500 index the two previous years (2018–2020).

**Keywords** Simple moving average · Stock market · Trading · Indicator · Simple return

---

J. P. Licona-Luque (✉) · L. F. Brenes-García · F. J. Cantú-Ortiz · H. G. Ceballos-Cancino  
Monterrey Institute of Technology (ITESM), Monterrey, NL, Mexico  
e-mail: [juan.licona@exatec.tec.mx](mailto:juan.licona@exatec.tec.mx)

L. F. Brenes-García  
e-mail: [luisfe.brenes@exatec.tec.mx](mailto:luisfe.brenes@exatec.tec.mx)

F. J. Cantú-Ortiz  
e-mail: [fcantu@tec.mx](mailto:fcantu@tec.mx)

H. G. Ceballos-Cancino  
e-mail: [ceballos@tec.mx](mailto:ceballos@tec.mx)

# 1 Introduction

This past two years the world has experienced one of the largest health threats of modern history. The COVID-19 pandemic's effects were heavily felt not only in hospitals and health-related institutions but also in the economic sector. Governments have implemented policies in an effort to stop the spread of the virus, but this policies have important economic consequences that are being reflected worldwide in the stock markets.

Volatility and risk on the global stock market have increased substantially, and there is a high correlation between the reactions of individual markets and the severity of the outbreak in each country [1]. While policies are needed to stop the spread of the virus and level the stock markets are needed, Zhang [1] suggested that non-conventional policy interventions have caused further uncertainty on the short-term while potentially creating long-term problems.

Volatile economic periods, as the one caused by COVID-19, tend to be challenging for traders and investors as the market turns risky and unpredictable. Is therefore important to them to understand which is the optimal investment strategy to use in this periods of economic turmoil. In this research, we intend to test the effectiveness of the simple moving average strategy (SMA) by back-testing it against the traditional buy and hold method (B&H).

We chose the SMA because is one of the technical indicators most used by professional investors . Besides that, it is also interesting to evaluate this moving average because there are mixed opinions on the academic world regarding technical indicators. On one hand, SMA has proven to outperform the buy and hold method in other periods of economic crisis such as the 2001–2002 Dot-Com bubble crash and the 2007–2008 Global Financial Crisis [2], but even when there is evidence that prove SMA has been effective back-testing in some historic contexts, the use of technical indicators is relatively recent and has received some critics from financial academics. An example of this is Malkiel [3] stating that “technical analysis is an anathema to the academic world” and that, after transaction costs, technical methods “don’t do better than a buy and hold strategy for most investors”.

The goal of our research is to test whether the SMA outperforms the traditional buy and hold method in the specific context of the COVID-19 global pandemic investing on the S&P 500 index.

# 2 Method

To be able to earn money in the stock market, one must sell stocks at a higher price than the originally paid: “buy low, sell high”. There are two main methods to identify ideal times to buy and sell stock, these are fundamental analysis and technical analysis. Fundamental analysis centers on the idea that sometimes a stock deviates from its intrinsic value, thus being undervalued or overvalued, this indicating when

one should buy or sell a stock. This type of analysis studies available information of the company “fundamentals” to determine the intrinsic value of the stock. Technical analysis aims to forecast the future stock price through the study of past price data and the identification of patterns. It rests on the idea that prices move in trends; therefore, one should buy when the trend is going upwards and sell when it’s going downwards [2].

The buy and hold strategy is usually associated with a fundamental analysis method, while the SMA is an example of technical analysis and it’s used as a tool to identify trends.

## 2.1 Buy and Hold

The buy and hold rule, as opposed to trading, rests on the idea that one should buy a good stock and hold to it for a very long time. It is usually a long-term investment strategy based on the empirical observation that if one buys stock from a good company, good returns can be expected on the long run, even if the prices fluctuate on a daily basis (volatility). People investing on a buy and hold strategy don’t trade their stock on a regular basis according to market fluctuation. They don’t sell as a reaction to what happens on the market, and they sell when they consider appropriate, according to their financial needs [4].

This strategy relies in the simple return formula (1) explained by Watsham and Parramore [5], where  $P_{t_0}$  is the original price and  $P_{t_1}$  is the current price:

$$R(t_0, t_1) = \frac{P_{t_1} - P_{t_0}}{P_{t_0}} \quad (1)$$

There are several ideas supporting buy and hold over trading strategies. Barber and Odean (2002) state that excessive trading, often caused by over-confidence, tend to result on poor performance for individual investors. This same study found that from 1991 to 1996 investors that traded less obtained significantly better returns than those who traded more [6]. Also, transaction costs need to be taken into account [7], and when these are present, trading is usually reduced [4].

## 2.2 Simple Moving Average

Identifying trends in stock market data is complicated, mainly because of the wild fluctuation on the stock prices driven by supply and demand. Moving averages such as the SMA are used to “smooth” the stock price data, thus revealing the underlying trend [2].

The SMA is given by the following equation:

$$\text{SMA}_t(n) = \frac{1}{n} \sum_{i=0}^{n-1} P_{t-i} \quad (2)$$

This equation can also be formulated recursively to accelerate it's computation:

$$\text{SMA}_t(n) = \text{SMA}_{t-1}(n) + \frac{P_t - P_{t-n}}{n} \quad (3)$$

The two main characteristics of a moving average are its lag time and its smoothness. Investors ideally would like averages that have a short lag time (reflect better recent price changes) and high smoothness (less smooth averages lead to a bigger number of trades, increasing the transaction costs). The problem is that this two characteristics are directly related, so that the smoother an average is, the longer it is lag time [2].

- The SMA lag time is given by

$$\text{Lag}(\text{SMA}_n) = \frac{\sum_{i=0}^{n-1} i}{\sum_{i=0}^{n-1} 1} = \frac{n-1}{2} \quad (4)$$

- As the SMA Herfindahl index equals  $\frac{1}{n}$  it's smoothness is given by

$$\text{Smoothness}(\text{SMA}_n) = \left(\frac{1}{n}\right)^{-1} = n \quad (5)$$

- An SMA lag time can be written as a linear function of it's smoothness as follows:

$$\text{Lag}(\text{SMA}_n) = \frac{1}{2} \times \text{Smoothness}(\text{SMA}_n) - \frac{1}{2} \quad (6)$$

### 2.3 Back-Testing SMA Against B&H

This experiment will be conducted performing a two year back-test from February 2020 to March 2022 over the S&P 500 index. Data was obtained from Yahoo Finance API. The test will be performed over the top 5 best and worst performing companies from the index, using Python as the analysis tool.

- B&H: For the buy and hold method, we will “buy” stock on March 1st 2020, hold it, and then “sell” it on March 1st 2022.

- SMA: A Python implementation will be used over the two year period to automatically “buy” when the market is bearish and “sell” when it’s bullish according to the SMA analysis.

We will then proceed to analyze the returns of both investing strategies and determine whether the SMA would have been a better method than buy and hold investing on the S&P 500 during the COVID-19 pandemic [1].

## **2.4 Development**

Once all the data has been collected, we will proceed to develop our experiment in two different stages. The first stage consists of using data from February 1, 2018 to February 4, 2020, in order to select the stocks for our simulation and the best parameters for the SMA. Stocks selection will be done by analyzing with the buy and hold strategy, and selecting the 5 best and the 5 worst performing companies. Then, the SMA strategy will be computed using the chosen stocks, iterating a window length of range from 5 to 45 days. Each iteration will be runned for every window length possible of each 10 stocks previously selected. After the calculation has been completed, we will use the highest SMA return of each stock and its, respectively, window length for the next stage. The second stage uses data from March 12, 2020 to March 3, 2022, calculating the buy and hold return for 10 previous stocks used in stage one. Then, we will use the best window length performance of the past stage into this second stage. In this way, using the past parameters on the new evaluations will provide real and well sustained results for trading applications. Finally, we will compare both strategies of the second stage and present the results.

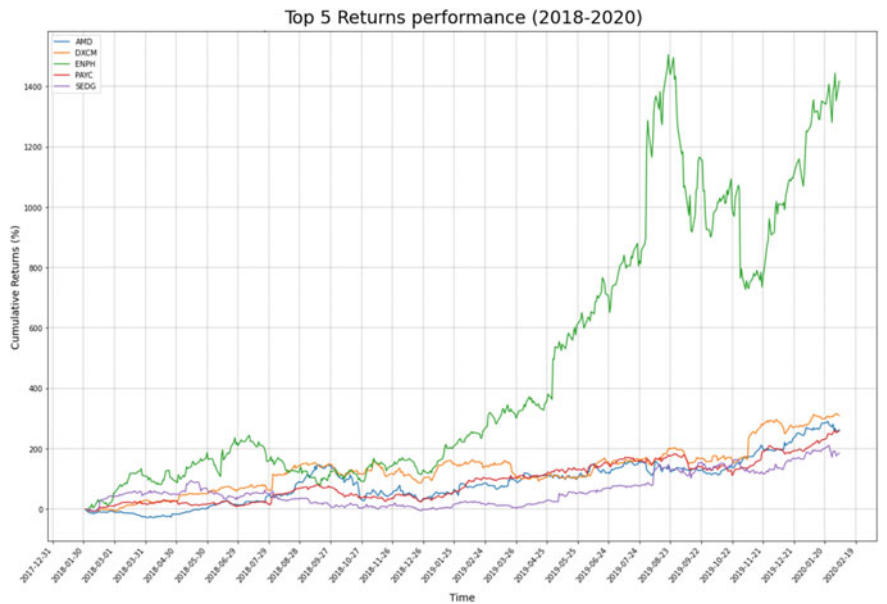
## **3 Results**

The goal of this research was to run a simulation of two hypothetical investors, one using a B&H strategy, the other one using SMA. We decided to run the simulation on the 5 companies from the S&P500 index that performed best and worst during the 2018–2020 period. The performance of this top 5 and bottom 5 companies on the time-frame 2018–2020 is shown in Figs. 1 and 2. Additionally, Table 1 shows the B&H returns of each company on 2018–2020.

Next step was to define the SMA parameters for our 2020–2022 simulation. The most relevant parameter to be defined was window length. The window length for each company was selected by running a simulation on each company through the 2018–2020 period trying different values from 1 to 100. Results from this simulation showing the window length that rendered best returns on this period are shown in Table 2. This window length was then used as a parameter for the simulation that we ran from 2020 to 2022.

**Table 1** Returns from 2018–2020

Category	Stock	BuyHold%
TOP	ENPH	1417.05
TOP	DXCM	309.69
TOP	AMD	262.41
TOP	PAYC	258.73
TOP	SEDG	185.25
BOTTOM	EQT	−79.43
BOTTOM	DXC	−61.53
BOTTOM	PCG	−59.43
BOTTOM	HAL	−58.51
BOTTOM	KHC	−59.13



**Fig. 1** Top 5 best performing companies 2018–2020

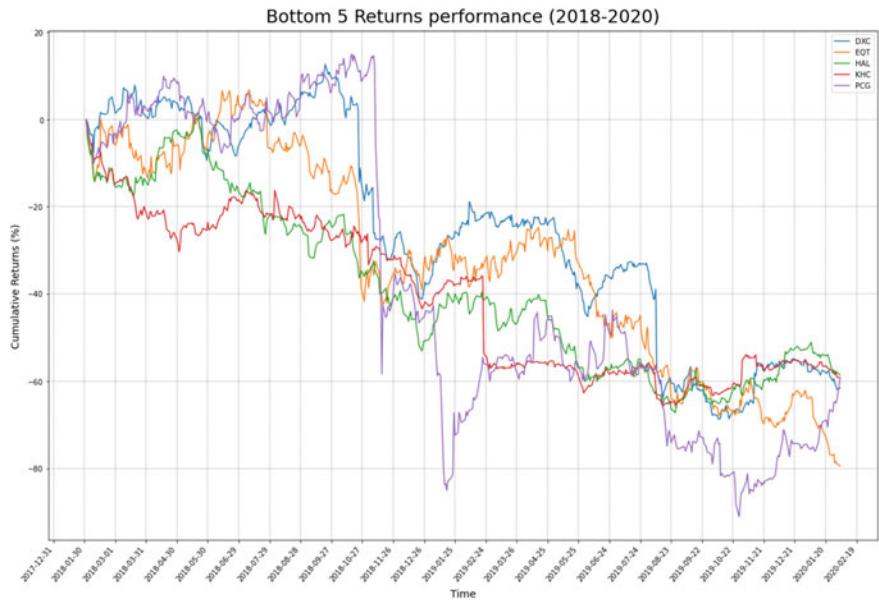
With the 5 best and worst companies selected and the parameters defined for SMA, we proceeded to do our simulation experiment. We recreate an hypothetical situation in which two investors, one using B&H and other using SMA, decide to invest during the COVID-19 period for two years. Investors are standing on February 2020 and have access to information and data from prior years (2018–2020), which is used to select the companies to invest in, and in the case of the SMA, the window length. Then, the simulation is ran from 2020 to 2022, with the results favoring B&H



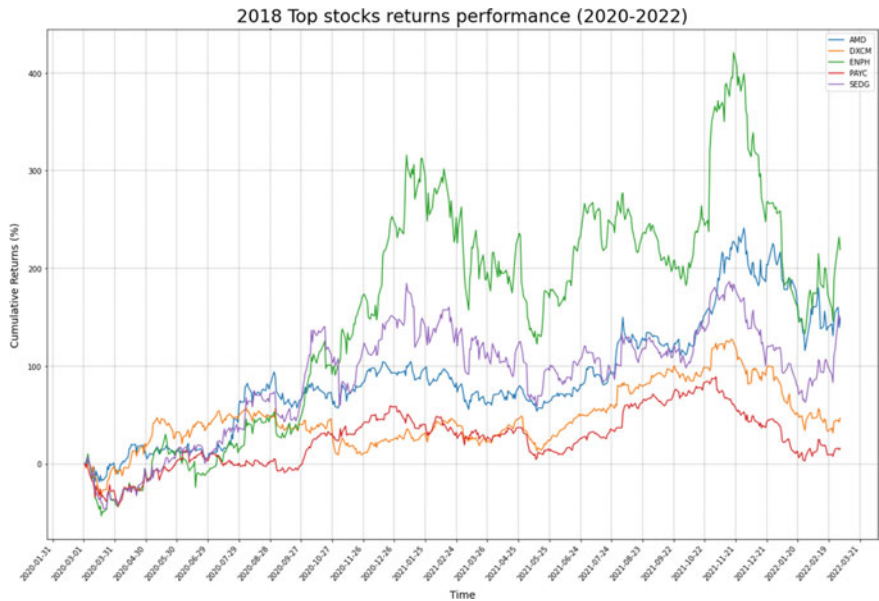
**Table 2** Optimal window length for top and bottom 5 companies selected after 2018–2020 evaluation

Category	Stock	WindowLen	SMA%
TOP	ENPH	21	1024.28
TOP	DXCM	6	198.33
TOP	AMD	37	184.52
TOP	PAYC	14	198.23
TOP	SEDG	42	134.14
BOTTOM	EQT	45	−33.98
BOTTOM	DXC	31	35.76
BOTTOM	PCG	37	42.61
BOTTOM	HAL	19	8.10
BOTTOM	KHC	14	−35.27

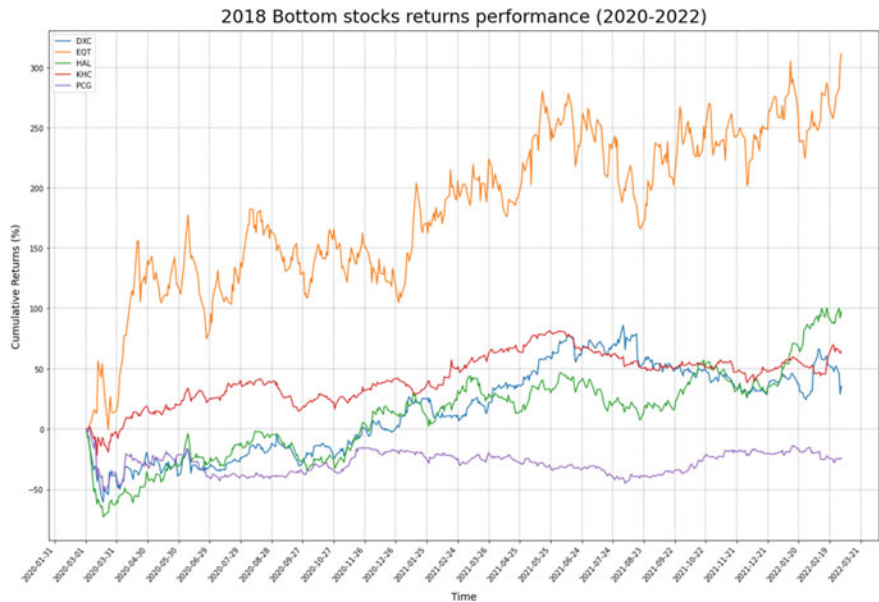
over SMA for 9 of the 10 companies as shown in Table 3. Additionally, Figs. 3 and 4 show the stock performance on any given date during a 2020–2022 time-frame.



**Fig. 2** Bottom 5 worst performing companies 2018-2020



**Fig. 3** Performance of the top 5 companies from 2018–2020 during the COVID-19 period 2020–2022



**Fig. 4** Performance of the bottom 5 companies from 2018–2020 during the COVID-19 period 2020–2022



**Fig. 5** KHC performance using SMA as indicator for when to buy stock (green arrow) and sell stock (red arrow)

**Table 3** Results of 2020–2022 simulation compare B&H and SMA performance

Category	Stock	B&H%	SMA%	SMA > B&H
TOP	ENPH	218.85	213.43	No
TOP	DXCM	47.09	17.85	No
TOP	AMD	149.22	21.86	No
TOP	PAYC	15.83	−14.02	No
TOP	SEDG	144.33	13.82	No
BOTTOM	EQT	311.58	95.67	No
BOTTOM	DXC	35.83	−44.02	No
BOTTOM	PCG	−23.85	−37.94	No
BOTTOM	HAL	97.32	46.23	No
BOTTOM	KHC	64.79	70.75	Yes

4 Discussion

Based on this results, it is clear that the SMA was not able to beat the market, even in the cases where the buy and hold method lost money, the SMA lost more. The only case where the SMA performed better was on stock from KHC but even then

the difference (5.96%) was not significant. The SMA indicator on the KHC stock showing where to buy and sell is shown on Fig. 5.

It is worth noting that this experiment is a simulation of an investor standing in March 2020 analyzing data from 2018 to 2020 to select the optimal investment strategy for the period 2020–2022, so the results from this last period are the ones we are focusing on. Data from 2018 to 2020 was analyzed to select the best and worst stock to invest in and the optimal window length for the SMA, but there was no forecasting. We can see that in the 2018–2020, buy and hold performs better on the TOP stock, and SMA performs better on the BOTTOM stock, as shown on Tables 1 and 2. However, this is rather an obvious consequence of running a simulation already knowing which companies performed best and worst in that time-frame. Again, this 2018–2020 simulation was just ran in an attempt to forecast which companies would perform best and worse in the next two years, assuming one was standing on March 2020 and had no information past that date. The relevant results are 2020–2022.

We can see that SMA beats buy and hold in 9 out of 10 companies. Why did this happen?

The SMA not out-performing the buy and hold method may be a result of the lag that is always present when forecasting using moving averages. The SMA gives equal weight to each price observation on the selected time window length, and there are other moving average methods that give more importance to the most recent stock price changes, thus reducing the lag time [2].

This results are in harmony with the suggestion of many financial experts who suggest that trading is a bad idea, and that the best investment strategy is to buy and hold [3, 4, 6]. Shiryayev et al. state that according to the efficient market hypothesis (EMH), if the market is efficient and the stock prices right there is no point to trade. One should sell a stock according to one's personal financial needs and not as a reaction to what is happening on the market [4].

#### ***4.1 Future Work***

For further research, it is recommended to repeat this experiment with different types of moving averages with a smaller lag time. We recommend experimenting with linear moving average (LMA), weighted moving average (WMA), and exponential moving average (EMA) which respond better to fast changes on the stock prices. It would also be interesting to repeat the experiment with more volatile stock than the S&P500 and compare how moving averages perform against the more traditional buy and hold strategy on high-risk and unpredictable markets like cryptocurrencies.

## 5 Conclusion

This paper presents an experiment that attempts to compare two common investment strategies, a traditional, and a technical one during the COVID-19 pandemic time-frame. Buy and hold is a traditional strategy that consists on buying a good stock and holding it for a determined period of time. SMA is a technical trading strategy that consists on buying and selling stock according to a moving average trying to beat the market.

The experiment was ran as a two year (2020–2022) simulation over data extracted from Yahoo Finance back-testing both strategies on the top 5 companies from the S&P500 index that performed best and worst the previous two years. The results were in favor the traditional buy and hold strategy which outperformed the SMA on 9 out of 10 stocks that were analyzed.

This research concludes that the SMA was not able to beat the S&P500 market on the 2020–2022 period, which confirms the opinion of several financial experts who recommend to buy and hold and avoid trading. An investor would have lost a lot of money opting for a SMA approach over the S&P500 this past two years.

We acknowledge that this experiment alone is not enough to determine whether the SMA would be a better strategy than the B&H in every possible scenario. For further work, we recommend repeating this experiment using different types of moving averages. We also recommend running the same simulation over more volatile and high-risk stock.

## References

1. Zhang D, Hu M, Ji Q (2020) Financial markets under the global pandemic of Covid-19. *Finan Res Lett* 36:10
2. Zakamulin V (2017) Market timing with moving averages: The anatomy and performance of trading rules
3. Malkiel BG (2007) A random walk down wall street: the time-tested strategy for successful investing
4. Shiryaev A, Xu Z, Zhou XY (2008) Thou shalt buy and hold. *Quant Finan* 8:765–776
5. Watsham TJ, Parramore K (1996) *Quantitative methods for finance*. Cengage Learning EMEA, Andover, England
6. Barber BM, Odean T (2000) Trading is hazardous to your wealth: the common stock investment performance of individual investors. *J Finan LV*
7. Glabadanidis P (2017) Timing the market with a combination of moving averages. *Int Rev Finan* 17:353–394

# Yield Prediction of Maize Using Random Forest Algorithm



Jane Kristine G. Suarez and Luisito Lolong Lacatan

**Abstract** This study provides a framework that can be used to advance the implementation of data mining in farming. The Department of Agriculture in the province of Bulacan, Philippines may create, incorporate, organize, manage, and carry out any policies, plans, programs, and activities relating to the study's findings. The output of the first phase is a prediction model. The second phase, will give rise into the development of a system that predict the future yield of maize with Geographical Information System for visualization. The researchers implemented a Cross-Industry Standard Process for Data Mining also known as CRISP-DM to successfully carry out the data mining process. Pre-processing was carried out using Waikato Environment for Knowledge Analysis using attribute selection to filter the data before subjecting it to an attribute evaluator. There are eight independent variables namely Municipality, Year, Month, Type of corn, Area, Production, Yield, Season, and one dependent variable called Status. In order to confirm the importance of the attribute, correlation feature-based selection subset evaluation was carried out which gives merit to attributes which have a higher correlation with the class or the dependent variable and low correlation to other independent variables. The algorithm used to predict the yield of maize as "high" or "low" got a computed kappa value of 0.9844, which is substantial enough to conclude that the algorithm used is reliable.

**Keywords** Agriculture · Geographical information system · Machine learning · Yield prediction

---

J. K. G. Suarez (✉)  
Bulacan State University, Malolos, Philippines  
e-mail: [jane kristine.suarez@bulsu.edu.ph](mailto:jane kristine.suarez@bulsu.edu.ph)

L. L. Lacatan  
Pamantasan ng Cabuyao, Cabuyao, Philippines

## 1 Introduction

Machine learning appears to hold promise for dealing with agricultural big data, but it will have to reinvent itself to meet current challenges [1]. With information technology, improvements and efficiency can be realized in practically any area of business, especially in agriculture. A farmer today harvests a growing amount of data in addition to crops.

In the province of Bulacan, Philippines the main agricultural and industrial products include bamboo, rice, sugar cane, maize, melons, and vegetables. It has three component cities and 569 barangays from 21 municipalities (Malolos the provincial capital, Meycauayan, and San Jose del Monte). North of Metro Manila is where Bulacan is situated. The provinces of Pampanga, Nueva Ecija, Aurora, and Quezon, as well as Metro Manila and Rizal, all border Bulacan to the west, north, east, and south, respectively. Additionally, Bulacan is located on Manila Bay's northeastern shore. With the increasing area of maize planted by farmers in the Philippines. The researchers focuses on maize which is another temporary dominant crop in the province of Bulacan.

In this study, the researchers utilized the data mining techniques, machine learning classification, and geographical information system-based model in the prediction of future yield. The data collected from the Provincial Agriculture will be the basis for training and test data for yield status. A spatial map was provided to indicate the location of high and low status of future yield. With the purpose of being of help the researchers selected the Province of Bulacan in partnership with the Provincial Agriculture Office of Bulacan with the aim to pursue sustainable and effective programs for farmers' and being true to its mission of being a responsive organization in the spirit of commitment and innovation for agricultural development. This study provides a framework that can be used to advance the implementation of data mining in farming and could serve as a basis for decision-making, policies, plans, projects, and activities relative to the yield prediction of crop.

## 2 Procedure

This section presents the methods in the chronological listing of steps and procedures utilized by the researchers. The methods used for gathering of data, as well as techniques employed in the analyses of data.

### 2.1 Study Area

The researchers collected the accomplishment report of maize planted and harvested from 12 municipalities such as Angat, Bustos, Norzagaray, San Miguel, Sta. Maria,

San Rafael, San Ildefonso, City of San Jose Del Monte, Doña Remedios Trinidad, Pandi, Marilao, and Bocaue prepared by the Provincial Corn Coordinator, Mr. Benigno G. Cruz, Senior Agriculturist. The data set ranges from year 2014 to 2020 of maize accomplishment report. The data set consists of five year data with the following parameters such as Municipality, Year, Month, Type of corn, Area, Production, Yield, Status and Season. The researchers performed attribute evaluation using a data mining tool into the set of possible attributes whose attributes (Municipality, Year, Month, Type of corn, Area, Production, Yield, Status and Season.) were selected based on the defined attributes used by different researchers who conducted similar field of study, then the data was subjected to an attribute evaluator. There are eight independent variables and one dependent variable called Status. Year, Area, Production, are all nominal variables, meaning the order or the sequence of its represented data has no bearing or weight in the analysis. The remaining attributes are ordinal, which means that there is a clear order but no specific unit of measurement in the value it represent.

## 2.2 Processing

This study aims to predict future yield of maize, and determine the data mining technique to be implemented in order to select the best algorithm, which can predict yield status such as high and low depending on the yield.

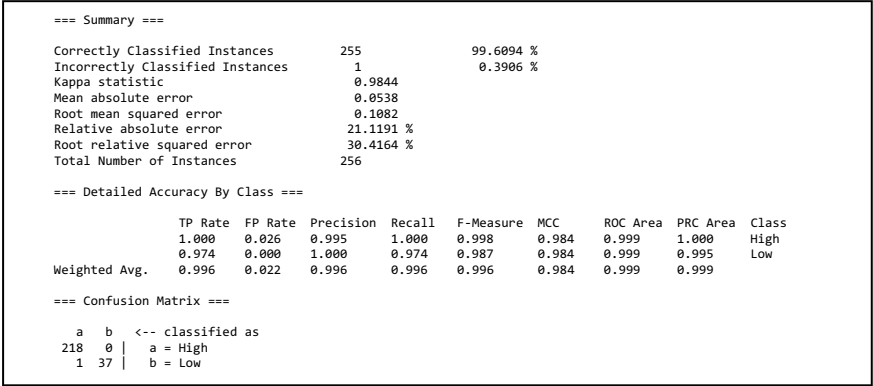
In order to identify the major predictors, the gathered historical data was filtered and analyzed using different techniques. For the given set of parameters, for instance, given the Municipality, Year, Month, Type of corn, Area, Production, Yield, Status and Season the kind of yield is expected like High or Low.

Applying the random forest algorithm to the data set, the algorithm determines the best split among the variables and creates a decision tree with the best split as its root node [2] as well as the leaf nodes of status such as high and low.

**Attribute Selection.** In the literature as indicated, J48, Bayes, Random Forest and JRip are among the most popular classification algorithm for predicting the yield of maize [3–5], These algorithms which are available in WEKA data mining tool were also tested under four test options, namely “Full Training Set”, “Supplied Test Set”, “Cross Validation” and “Percentage Split.” The last two options were carried out using the 10-Fold Cross Validation and 70–30 Split-sample validation.

In “Full Training Set”, the whole set of data containing the nine significant attributes were tested. The confidence factor was set to 0.25 with a minimum of ten (10) objects or instances per leaf and tree (3) folds of data to be used for error pruning. In “Supplied Test Set”, one hundred-fourteen (114) instances (approximately 30% of the fill data set) were randomly chosen. Its purpose is to test the accuracy of the model built by the algorithm. In GreedyStepWise search method selection through Cross-Validation, area planting appeared 8 out of 10 folds (80%) and month appeared 10 out of 10 folds (100%) which makes them significant predictors. As presented





**Fig. 1** WEKA data mining tool under cross validation

in Fig. 1, there are two hundred fifty-six instances tested in CfsSubset Evaluation. WEKA was used to carry out the CfsSubsetEvaluation in order to further confirm the importance of attribute month, area planting. CfsSubsetEvaluation gives merit to attributes which has a higher correlation with the class or the dependent variable and low correlation to other independent variables [6].

In BestFirst method, month and area planting were found to be the significant subset with 0.39 value of merit of best subset found. This study further shows that there are nine attributes included in the CfsSubsetEvaluation using the BestFirst method in the full training set implementing forward searching starting at an empty subset and will terminate the searching after five nodes. Though the literature is not saying any standard for the ideal value of merit of best subset found, it could be traced from its formula.

$$M_s = \frac{kr_{cf}}{\sqrt{k + k(k - 1)r_{ff}}}$$

(1)

where

- $M_s$  Merit of subset
- $K$  number of features
- $r_{cf}$  mean feature-class correlation
- $r_{ff}$  feature-feature intercorrelation.

A higher merit will be yield, if there is a great mean of feature-class correlation, meaning the independent variable is highly correlated to the dependent variable, which further explains that the higher the value of “merit” the better. A higher merit score represents a better subset [7].

*Training and Testing.* In order to select the most appropriate data mining technique or algorithm best suited for Yield Prediction of Maize, the researchers consider the following effects of weather/climate variability on rice production [8, 9] and rice

yield prediction using satellite imagery and neural networks [10]. The rice yield prediction using radar remote sensing and crop model [11] and crop yield prediction using Random Forest [12].

The researchers came up with a shortlist of various algorithms revealed from the review of related studies and literature. Selection was made after studying and testing the different approaches such as Bayes Classifiers (Naïve Bayes), Decision Tree (J48), Random Forest and Rule Classifier (JRip).

As indicated in the literature, J48, Bayes, Random Forest and JRip are also among the most popular classification algorithm for predicting the Yield of Maize. These algorithms which are available in Weka data mining tool were also tested under four options, namely, “Full Training Set”, “Supplied Test Set”, “Cross Validation”, and “Percentage Split”.

Random forest accuracies were very promising this algorithm considered significant parameters of classifiers such as accuracy percentage, standard error, kappa value and time to build the model [13] the researchers were able to used and compare the efficiency of performance of the selected classification algorithm implemented using Weka.

The results of Weka data mining tool under four options were presented, the first option is “Full Training Set”, the algorithm resulted in high accuracy of (100), standard error (0), kappa value (1), and time to build the model in (0) seconds. The outcome of the second test option, which is “Supplied Test Set”, also achieved the highest accuracy (100), standard error (0), kappa value (1), and time to build the model (0.02) seconds. The third option, “Percentage Split” reached accuracy (100), standard error (0), kappa value (1), and time to build the model (0) seconds and lastly fourth option, “Cross Validation”, the algorithm got the second highest accuracy rate of (99.6094), standard error (0.3906), kappa value (0.9844), and time to build the model (0.01) seconds.

Among the four test options cross-validation were utilized by the researchers in this study, the goal of this technique is to define a data set to test the model in the training phase in order to limit problems like overfitting and underfitting.

Overfitting happened when capturing the unseen data is poorly generalized by noise and capturing patterns. The model does a great job on the training data, but it does much worse on the test data. Underfitting is defined as a model that does not capture enough patterns in the data. Both the test set and the training set show poor model performance. The ideal model should be capable of performing well on both the train and test sets and provide insight into how the model will generalize to a different collection of data.

The dataset was divided into two subset, the Training Sample and the Testing Sample which comprises 70% and 30% of data respectively as suggested [14]. The training set is for building the model and the testing is to test the validity of the model as shown in Fig. 1. Using the training set, there are 256 instances predicted as True Positive (TP) and 0 as False Negative (FN) which means that there is a total of 256 (out of 370—which is approximately 70% of total data sets) instances classified correctly, thus yielding an overall percentage of 100% correct classification. On the other hand, when the model built was tested using the testing data set, it was found

out that 114 instance out of 370 were classified correctly, giving an overall percentage of 30% correct classification.

Approximately two hundred fifty-six (256) instances were selected from the data set, out of the given datasets, two hundred fifty five (255) were classified as True Positive and thirty-seven (37) instances as False Negative. Only one (1) instance were misclassified. To measure accuracy, the True Positive (TP) Rate, False Positive (FP) Rate, Precision, Recall, and *F*-Measure values are used.

Table 1 shows that random forest algorithm used to predict the yield of maize as “High” or “Low” got a computed kappa value of 0.9844, which is substantial enough to conclude that the algorithm used is reliable.

*Prediction and Validation.* Cohen’s Kappa will be used to assess the accuracy of the predictions made by the system on the yield prediction of maize in the Province of Bulacan as either “High or Low.”

The reliability or degree of similarity between two or more variables is assessed using Cohen’s Kappa (*K*) analysis [15]. There are no intermediate levels of disagreement between the two raters’ ratings; they are either in agreement or disagreement. The equation for *K* is:

$$K = \frac{P_{(A)} - P_{(E)}}{1 - P_{(E)}} \tag{2}$$

where

- $P_{(A)}$  number of agreements
- $P_{(E)}$  number of agreements expected by chance.

*K*’s value is calculated, and then the magnitude criteria in Table 2 are applied [16] can then be used to interpret its kappa value.

As shown in Table 2, in the literature [17], showed that  $K > 0.6$  or even  $K > 0.5$  was the threshold for acceptable interrater reliability.

*Yield Prediction of Maize with GIS System*

*Results.* In order to prevent overfitting and obtain a more accurate prediction, the training process also included the use of a Random Forest Decision Tree. This prediction included tenfold cross-validation.

The calculated degree of agreement (kappa value) between the generated prediction (yield of maize) by the developed system and the actual yield of maize from

**Table 1** True positive (TP) rate, false positive (FP) rate, precision, recall, and *F*-measure values

Class	TP rate	FP rate	Precision	Recall	<i>F</i> -measure
High	1.00	0.026	0.995	1.000	0.998
Low	0.974	0.000	1.000	0.974	0.987
Weighted average	0.996	0.022	0.996	0.996	0.993

**Table 2** Interpretation of the magnitude of Kappa values

Kappa statistic	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

**Table 3** Cross-validation weighted average for model performance comparison

Classifier	F-measure	Recall	Precision	TP rate	FP rate	Kappa value
Random forest	0.996	0.996	0.996	0.996	0.022	0.9844
Naïve Bayes	0.837	0.816	0.889	0.816	0.162	0.4738
J48	0.996	0.996	0.998	0.996	0.022	0.9844
JRip	0.996	0.996	0.996	0.996	0.022	0.9844

**Fig. 2** Dataset visualization report with GIS for high or low status of yield



the report of the Provincial Agriculture Office was compared by the researchers in order to assess the system’s accuracy and dependability in predicting the yield of maize. Those data were individually inputted into the system (testing) and the system generated the prediction. Out of two hundred fifty-six instances (256), two hundred eighteen (218) were classified correctly as True Positive and thirty-seven (37) instances as False Negative. Only one instance were misclassified. Using the Cohen’s Kappa formula and the kappa value of the developed system it is computed as 0.9844, interpreted as “Almost Perfect”, based on the Analysis of the Kappa Value’s Magnitude. The total number of correctly classified cases divided by the total number of instances is how the accuracy of right prediction can be calculated that is:

$$\text{Percentage Correct} = \frac{218 + 37}{256} (99\%) \quad (3)$$

The algorithm used to predict the yield of maize as “high” or “low” got a computed kappa value of 0.9844, which is substantial enough to conclude that the algorithm used is reliable.

The necessity to test the stability of the machine learning model and how well it will generalize to new data makes validation one of the most crucial tools a data scientist uses. It must ensure that the model accurately captures the majority of the patterns in the data and does not over-infer noise, or low bias and variance.

Table 4 shows that among the four test options random forest using “Cross Validation” has the second highest accuracy percentage similar to J48 and JRip, while Naïve Bayes has the lowest accuracy percentage.

### 3 Conclusions

The algorithm used to predict the yield of maize as “high” or “low” got a computed kappa value of 0.9844, which is substantial enough to conclude that the algorithm used is reliable.

If the accuracy deceives into thinking the classifier is really effective. Mathews Correlation Coefficient (MCC) aids in determining the classifier’s incapacity to accurately categorize samples, particularly those from the negative classes. MCC is between  $-1$  and  $1$ . According to the definition, “MCC is high only if your classifier is doing well on both the negative and the positive aspects.”

The comparison of the performance results of the four classifiers (J48, random forest, Naïve Bayes, and JRip) considering important parameters of accuracy, absolute error, kappa value and time to build the model evaluated that random forest is the best algorithm to be used in classifying the yield of maize in the context of this study. Random forest was considered best although it ranked equally with JRip and J48, in case classifiers perform nearly equal. The best algorithm has been determined to be the one with the highest accuracy and shortest execution time.

Using WEKA, a set of attributes was analyzed; during the pre-processing stage, attributes such as Municipality, Year, Type of Corn, Production, Yield, Status, and Season were automatically deleted because they were deemed to be highly irrelevant.

In GreedyStepWise search method selection through Cross Validation, area planting appeared 8 out of 10 folds (80%) and month appeared 10 out of 10 folds (100%) which makes them significant predictors. It was also observed that Municipality, Year, Type of Corn, Production, Yield, Status, and Season were disregarded as significant predictors in WEKA, while on that same method using 10-Fold Cross-validation, the variable area planting was found to be 80% only significant. To further verify the significance of attribute month, area planting, CfsSubsetEvaluation was performed in WEKA. CfsSubsetEvaluation gives merit to attributes which has a higher

**Table 4** Four test options random forest, J48, JRip and Naïve Bayes

Techniques	Algorithm	Test options	Accuracy	Std. error	Kappa	Time to build (sec.)
Bayes classifiers	Naïve Bayes	Full training set	83.5938	16.4063	0.5335	0
		Supplied test set	88.8889	11.1111	0.5787	0.02
		Cross validation	81.6404	18.3594	0.4738	0
		Percentage split	79.2208	20.7792	0.3383	0
WEKA tree classifiers	J48	Full training set	100	0	1	0
		Supplied test set	100	0	1	0.01
		Cross validation	99.6094	0.3906	0.9844	0.02
		Percentage split	100	0	1	0
Rule classifiers	JRip	Full training set	100	0	1	0
		Supplied test set	100	0	1	0.02
		Cross validation	99.6094	0.3906	0.9844	0.03
		Percentage split	100	0	1	0
WEKA tree classifiers	Random forest	Full training set	100	0	1	0
		Supplied test set	100	0	1	0.01
		Cross validation	99.6094	0.3906	0.9844	0.01
		Percentage split	100	0	1	0

correlation with the class or the dependent variable and low correlation to other independent variables. A higher merit will be yield if there is a great mean of feature-class correlation, meaning the independent variable is highly correlated to the dependent variable, which further explains that the higher the value of “merit” the better.

The study Yield Prediction of maize using random forest algorithm in the long run provides a framework that can be used to advance the implementation of data mining in farming. The Provincial Agriculture Office will certainly benefit in the output of

this research especially in data-driven decision-making, policies, plans, projects, and activities relative to the yield prediction of crop.

**Acknowledgements** The authors would like to acknowledge the assistance extended by the Department of Agriculture Regional Field Office III, Provincial Agriculture Office in Bulacan, National Disaster Risk Reduction Management Center (NDRRMC), Philippine Rice Research Institute (PhilRice), Philippines Atmospheric Geophysical and Astronomical Services Administration (PAG-ASA), Bureau of Soil and Water Management (BSWM), Regional Crop Protection Center (RCPC), and Bulacan State University.

## References

1. Tantalaki N, Souravlas S, Roumeliotis M (2019) Data-driven decision making in precision agriculture: the rise of big data in agricultural systems. *J Agric Food Inf* 20(4):344–380. <https://doi.org/10.1080/10496505.2019.1638264>
2. Fratello M, Tagliaferri R (2018) Decision trees and random forests. *Encycl Bioinform Comput Biol: ABC Bioinform* 1:3
3. Roux YL, Grobler J (2020) Investigating the use of machine learning for South African edible garnish yield prediction. In: *International conference on innovative techniques and applications of artificial intelligence*. Springer, Cham, pp 202–214
4. Ahmed GN, Kamalakkannan S (2022) Micronutrient classification in IoT based agriculture using machine learning (ML) Algorithm. In: *4th international conference on smart systems and inventive technology (ICSSIT)*, pp 1–9. <https://doi.org/10.1109/ICSSIT53264.2022.9716293>
5. Yange TS, Egbunu CO, Rufai MA, Onyekwere O, Abdulrahman AA, Abdulkadri I (2020) Using prescriptive analytics for the determination of optimal crop yield. *Int J Data Sci Anal (IJDSA)* 6(3):72–82
6. Mokarram M, Ghasemi MM, Zarei AR (2020) Evaluation of the soil fertility for corn production (*Zea Mays*) using the multiple-criteria decision analysis (MCDA). *Model Earth Syst Environ* 6:2251–2262. <https://doi.org/10.1007/s40808-020-00843-5>
7. Demisse GB, Tadesse T, Bayissa Y (2017) Data mining attribute selection approach for drought modeling: a case study for Greater Horn of Africa. [arXiv:1708.05072](https://arxiv.org/abs/1708.05072)
8. Lansigan FP, de los Santos WL, Coladilla JO (2000) Agronomic impacts of climate variability on rice production in the Philippines. *Agric Ecosyst Environ* 82(1–3):129–137. ISSN 0167-8809. [https://doi.org/10.1016/S0167-8809\(00\)00222-X](https://doi.org/10.1016/S0167-8809(00)00222-X). <https://www.sciencedirect.com/science/article/pii/S016788090000222X>
9. Stuecker MF, Tigchelaar M, Kantar MB (2018) Climate variability impacts on rice production in the Philippines. *PLoS ONE* 13(8):e0201426. <https://doi.org/10.1371/journal.pone.0201426>
10. Chen RC, Dewi C, Huang SW et al (2020) Selecting critical features for data classification based on machine learning methods. *J Big Data* 7:52. <https://doi.org/10.1186/s40537-020-00327-4>
11. Setiyono TD, Quicho ED, Holecz FH, Khan NI, Romuga G, Maunahan A, Garcia C, Rala A, Raviz J, Collivignarelli F, Gatti L, Barbieri M, Phuong DM, Minh VQ, Vo QT, Intrman A, Rakwatin P, Sothy M, Veasna T, Pazhanivelan S, Mabalay MRO (2018) Rice yield estimation using synthetic aperture radar (SAR) and the ORYZA crop growth model: development and application of the system in South and South-east Asian countries. *Int J Remote Sens*. <https://doi.org/10.1080/01431161.2018.1547457>
12. Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, Timlin DJ, Shim KM, Gerber JS, Reddy VR, Kim SH (2016) Random forests for global and regional crop yield predictions. *PloS ONE* 11(6):e0156571. <https://doi.org/10.1371/journal.pone.0156571>

13. Li H, Zhang C, Zhang S, Atkinson PM (2020) Crop classification from full-year fully-polarimetric L-band UAVSAR time-series using the random forest algorithm. *Int J Appl Earth Obs Geoinf* 87:102032
14. Liu H, Cocea M (2017) Semi-random partitioning of data into training and test sets in granular computing context. *Granular Comput* 2(4):357–386
15. Chen C, Quilang E, Alosnos E, Finnigan J (2011) Rice area mapping, yield, and production forecast for the province of Nueva Ecija using RADARSAT imagery. *Can J Remote Sens* 37:1–16. <https://doi.org/10.5589/m11-024>
16. Sabharwal CL (2021) Cohen's Kappa statistic and newKappaStatistic for measuring and interpreting inter-rater agreement
17. Livadiotis G (2017) Kappa distributions: theory and applications in plasmas. Elsevier



# Enhancement of Prototype Driving Simulator Using Available AI-Based Game Technology



Yun-Quan Cheng<sup>ID</sup>, Sarina Mansor<sup>ID</sup>, Ji-Jian Chin<sup>ID</sup>,  
Hezerul Abdul Karim<sup>ID</sup>, and Ban Kar-Weng<sup>ID</sup>

**Abstract** Simulators—games that simulate the real world in a virtual environment, such as racing simulators, have been widely studied and documented. Their uses could, however, be further expanded into the field of driving education. The motivations behind this study are to dive into the trend of immersive learning and exploit artificial intelligence (AI)-based game technology to benefit driving education. The preliminary work shows that the use of a driving simulator as a teaching tool for driving education has improved the passing rate of driving training. This paper proposes several enhancements to the prototype driving simulator, by utilizing rudimentary AIs to simulate basic traffic for certain scenarios. The driving simulator is developed in Unity and is paired with a driving rig consisting of a steering wheel and pedals. The project currently is a functioning driving simulator that can accept both controller and steering wheel inputs. For prototype enhancement, three scenario-based tracks and a mock-up town with a traffic light system and AI traffic were added. The prototype now includes five circuit tracks and two on-the-road tracks based on the Standard Examination syllabus, three scenario-based tracks, and a mock-up town with AI traffic, together with an automated test mode for the syllabus tracks. The free practice mode will be available for all tracks.

**Keywords** Driving education · Driving simulator · Game AI · Prototype enhancement

---

Y.-Q. Cheng · S. Mansor (✉) · H. A. Karim

Faculty of Engineering, Multimedia University, Cyberjaya, Selangor, Malaysia

e-mail: [sarina.mansor@mmu.edu.my](mailto:sarina.mansor@mmu.edu.my)

Y.-Q. Cheng

e-mail: [tylerchengyq@gmail.com](mailto:tylerchengyq@gmail.com)

H. A. Karim

e-mail: [hezerul@mmu.edu.my](mailto:hezerul@mmu.edu.my)

B. Kar-Weng

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia

e-mail: [kwban@mmu.edu.my](mailto:kwban@mmu.edu.my)

J.-J. Chin

Faculty of Science and Engineering, University of Plymouth, Plymouth, UK

e-mail: [ji-jian.chin@plymouth.ac.uk](mailto:ji-jian.chin@plymouth.ac.uk)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_47](https://doi.org/10.1007/978-981-99-3091-3_47)

## 1 Introduction

Simulators have been in existence for quite a while. They have been used extensively in specific skills training such as motor sports. Simulators have been studied and widely documented as mentioned by [1]. For home users, however, the increased affordability and accessibility of entry level racing simulation (sim) equipment have made these sim games incredibly popular recently [2]. A work done by Tiu et al. [3] proves that these sims can help minorly disabled people as well. These sims are usually racing sims, taking place on a simulated race track with the player racing with the AIs or other players through an online multiplayer setup. These sims show that the current generation of game technology is capable of simulating/mimicking real racing environments in games that are developed for consumer-level use. Virtual reality (VR) technologies have also become more accessible for consumers through heavy investments by major corporations such as HTC, as discussed by [4]. Mora et al. [5] posit that it is sound to take advantage of such events to harness the potential of VR technologies in education. We believe it is possible that we can harness the AIs in current game technologies and virtual reality to help improve and innovate driving education while pushing the boundaries further in terms of learning effectiveness and skill training efficiency.

Based on our review article [6], we found gaps within the studies relating to driving education. This motivated us to explore the possible use of combining three main areas of driving simulation: artificial intelligence (AI), computational intelligence (CI), and virtual reality (VR), to improve learning effectiveness in driving education. To the best of our knowledge, no such studies combining these aspects have been done in Malaysia, where the authors reside. Especially at the consumer-level, game development and VR have been increasingly more accessible for end-users, and we think it is time to maximize utilization of these technologies to spearhead its use in driving education for Malaysians.

The main contribution of this work is the enhancement of a prototype of a driving simulator [7] that consolidates the use of AI-based game technology. We present our development progress of the prototype that may benefit other studies in developing their own.

## 2 Driving Simulator Prototype

This section briefly describes the driving simulator that was developed in our preliminary work [7].

**Fig. 1** Driving simulator prototype with a steering wheel rig

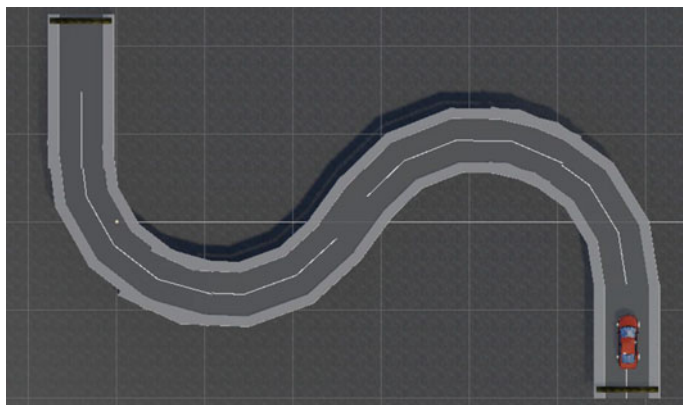


## 2.1 Background

The driving simulator prototype was developed with references from the Standardized License Exam of Malaysian Ministry of Transport [8, 9]. The platform that the prototype is developed on is Unity Engine, a popular free to use game engine by Unity Technologies [10]. The driving simulator at its current state can take input from a gamepad/controller and its main input system, the steering wheel rig as shown in Fig. 1. The preliminary prototype includes five circuit tracks and two on-the-road tracks from the syllabus. These tracks have two modes—a free practice mode and a test mode. The free practice mode is a mode where the user can practice without time limits. The test mode provides an automated test environment where a time limit is imposed. Based on our experiments that we conducted, it showed some potential in improving learning effectiveness [7].

## 2.2 Syllabus-Based Tracks

The syllabus-based driving tracks directly take cues from the real-life examination . These circuit tracks include S-circuit, Z-circuit, three-point turn circuit, hill climb circuit, and a parallel parking circuit. The remaining two theory circuits are the yield circuit and cross junction circuit. Example of S-circuit is as shown in Fig. 2. More details of syllabus-based tracks can be found in [7].



**Fig. 2** S-circuit

These circuits run on a basic core flow, in which the circuits have a trigger check to see if the required task has been done correctly when the user reaches the end of the circuit. These tracks have a common system where if the car falls out of the track, it would be deemed an automated failure and prompt an end menu.

### ***2.3 Practice and Test Modes***

The tracks have 2 different modes, which are practice and test mode. The difference between the two modes is the time limiter. In the test mode, there is a time limiter that will automatically fail the user if the time goes over the limit.

## **3 Prototype Advancements**

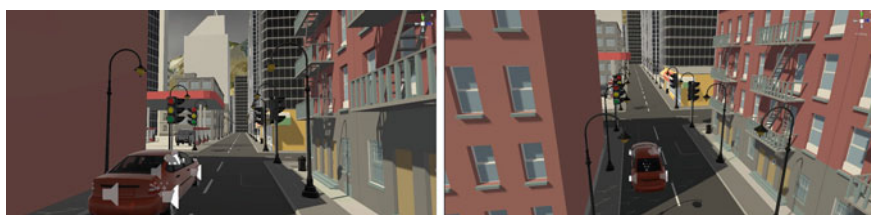
This section describes the advancements that we have extended onto the preliminary prototype, enhancing various aspects of the simulation experience. We added another three scenario-based tracks and a mock-up town with rudimentary AI traffic based on some tracks that are already available. These tracks are available in the free practice mode only.

### ***3.1 Scenario Tracks***

These tracks are designed to replicate possible scenarios that would happen in real traffic. For example, traffic lights breaking down and heavy traffic. These tracks do



**Fig. 3** Scenario 1 circuit



**Fig. 4** Scenario 2 circuit (angle 1 and 2)

not have test mode as time limiter is not needed, and it requires human judgement. Three scenario tracks were designed, with each track serving a different modular purpose. These tracks are (i) Scenario 1: a T-junction scenario track with traffic lights, shown in Fig. 3, (ii) Scenario 2: a cross junction scenario track with traffic lights, shown in Fig. 4, (iii) Scenario 3: a highway section, shown in Fig. 5. Due to the modular system design, these tracks can be modified accordingly to suit a specific purpose or scenario without redesigning the whole track layout. Scenario track 1 and Scenario track 2 are results of extension work done on the T junction circuit and the cross junction circuit from the syllabus tracks. Scenario track 3 is a new track designed from the ground up. Currently, the tracks are still empty tracks with layouts as appropriate scenarios are still being designed to be implemented.

### 3.2 *Mock-Up Town*

The mock-up town is designed and developed based on the Scenario 1 track. Enhancements that were put into the mock-up town include a complete loop road, extra



**Fig. 5** Scenario 3 circuit



**Fig. 6** Mock-up town (angle 1 and 2)



**Fig. 7** Mock-up town (angle 3 and aerial view)

environmental elements, traffic light system, and an AI car to act as traffic. The motivation behind this was to simulate a small town and some minor traffic to ease the user in driving in an almost true-to-life situation. It also acts as a testing ground for some of the driving simulator systems. Figures 6 and 7 show multiple angles of the mock-up town.

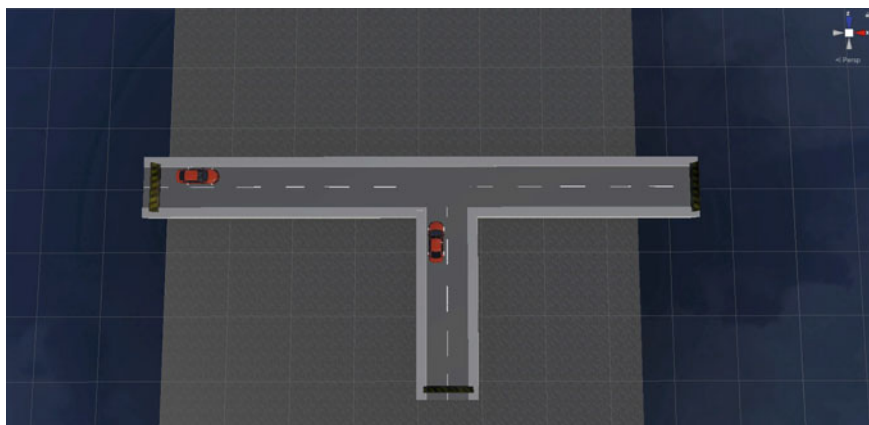
### 3.3 Summary

All tracks and systems were developed and built on to the Unity platform and are operational. The scenes are optimized visually to give a smooth user experience. As a result, the scenes run at least 60 frames per second (FPS) on 1080p resolution. We ran the simulator at 1440p resolution, and all scenes achieved 144 FPS performance. Loading times were almost instantaneous on all except Scenario 3 and the mock-up town, as both scenes have heavy environmental elements that might slow down the loading process.

## 4 Discussion

### 4.1 Advancement work

The current Scenario tracks and mock-up tracks are a direct result of the advancement work done to the preliminary prototype and are an improvement on the previous prototype. The advancement work consists not only of visual upgrades, it also introduces some performance optimization and extra systems such as the traffic light system. The results of all the visual improvements are shown in the Figs. 8 and 9, showing the evolutionary change of the prototype track that is based on the T-junction circuit.



**Fig. 8** T-junction track, first rendition





**Fig. 9** Left (Scenario 1, second rendition), right (mock-up town, third rendition)

## 4.2 Modularity in Track Designs

The prototype follows a fundamental design that is modular. We wanted the prototype to have the ability to be re-purposed and create new track layouts from existing tracks to ease the process and save on development time. As mentioned, the Scenario Tracks 1 and 2 and mock-up town are designed based on the syllabus tracks. This shortened the development process and showed an incremental change in enhancements done to the tracks. It also allowed us to trial and error new systems in the newer tracks more conveniently as we already established that the basic system is functional.

## 4.3 Layout of Mock-Up Town

The layout of the mock-up town is design as a loop is to:

- Allow the AI car/traffic to loop around
- Allow the user to drive around in a loop endlessly
- Ease the development process
- Visual and performance optimization

With regards to the first two points, the purpose was to let the AI(s) loop around freely to create a sense of occupancy and to allow the driver to practice their driving in an endless loop. This layout eased the development process by not requiring many junctions and separate gaps to fill environment elements. The loop was done to have an internal space that is filled with environmental elements just enough to cover the gaps and maintain performance while being esthetically pleasing.

## 4.4 Simple Traffic Lights System

In the mock-up town, a simple traffic light system is implemented. The core concept is where one of the traffic lights acts as a main timing resource to other traffic lights. The secondary traffic light will take timing cues when the lights go red on the main





**Fig. 10** Left (Traffic light array in mock-up town showing red), right (traffic light array in mock-up town showing green, and blinking yellow)

traffic light. This then starts the timing and process of the secondary traffic light. The timings are all set in the source code prior: The secondary traffic light only takes timing cues on when to turn green, and the rest is an internal loop. There is also a blinking yellow light system for "offline" traffic lights, which is a loop of blinking yellow lights with a 2 s gap in between; the traffic light arrays are shown in Fig. 10. There is also a system that automatically prompts a fail UI/Menu when the user runs the red light. The trigger is placed at half a car's length after the traffic light to allow for some tolerance against users missing the stop area.

## 4.5 The AI Car

We used a plug and play AI car with custom scripting in the mock-up town to test the implementation of a way-point-based traffic system. The implementation is achieved when the AI car runs in the loop of the town, successfully navigates its own way-points shown in Fig. 11 (left). A main challenge faced was due to the AI car only having the ability to follow one way-point, therefore we used a tracker method to overcome this issue. The tracker tracks which way-point the AI car has arrived at and tells the AI car to go to the next way-point, shown in Fig. 11 (right). As the way-points are set to go in the opposite direction from the user, shown in Fig. 12, the car did not cross over onto the user's lane. This AI is self-contained, and no other input is needed. There are two more sophisticated systems that have failed to be implemented. These systems both have unique features of their own, one of the features being able to spawn in AI traffic and allow the AI traffic to navigate in a major connected route randomly. The other one has the feature that allows the traffic to follow traffic light systems. The reason these systems were unable to be implemented was the difference in Unity Engine versions as there were no solutions to this issue. Hence, it is why we used the way-point system traffic AI albeit it being rudimentary.



**Fig. 11** Left (One of the AI car way-points), right (the AI car navigating to its way-point)



**Fig. 12** Left (the AI car on the right of the user's car), right (the AI car from inside the user's cockpit)

## ***4.6 Planned Experiment and Analysis***

The planned experiment will be mainly testing on the efficacy of conventional single screen and VR. As the experiments are still being designed and planned, different methods of experimentation might be used in actual experimentation. To get reliable results, the experiment will be using the simulator's automated test modes. The analysis will be done upon the results from the automated tests to ensure the validity of the results.

## ***4.7 Implications***

This simulator will allow learner students to have the confidence to drive a car even before having any real-world experiences. Among the possible implications that this will have is a reduction in the cost of driving education, as less fuel is used for initial practical training and practice. It also reduces the greenhouse gases emitted from using gasoline cars. This method could also produce more efficient results by simply allowing the students to have more practical time that was not possible with cars as space in the training grounds and time are limited. Hence, costs, emissions, and safety risks can be lowered even when rate of repeatability is increased.

## 5 Conclusions

In this paper, we have presented a prototype advancement of our driving simulator prototype, which includes scenario tracks, a mock-up town, and various systems that have been discussed. Currently, the prototype is operational with its discussed features. As the plug and play AI is still very rudimentary even with custom scripting, we wish to work on it in future to develop a self-contained AI car that roams around on its own instead of following a set route. Focused on driving education, future use could see application mainly in driving education, but we do not rule out the possibility of use in application such as training devices for the disabled and reinforcement training for the elderly or people with fear of driving.

**Acknowledgements** The authors would like to thank the Ministry of Education of Malaysia in providing financial support for this work through the Fundamental Research Grant Scheme (FRGS/1/2020/SS0/MMU/02/1).

## References

1. Burkhardt JM, Corneloup V, Garbay C, Bourrier Y, Jambon F, Luengo V, Job A, Cabon P, Benabbou A, Lourdeaux D (2016) Simulation and virtual reality-based learning of non-technical skills in driving: critical situations, diagnostic and adaptation. *IFAC-PapersOnLine* 49(32):66–71. <https://doi.org/10.1016/j.ifacol.2016.12.191>, cyber-Physical Human-Systems CPHS 2016
2. Why is sim racing so popular? <https://www.upshiftstore.co.uk/pages/why-is-sim-racing-popular>. Accessed 01 Jun 2021
3. Tiu J, Harmon A, Stowe J, Zwa A, Kinnear M, Dimitrov L, Nolte T, Carr D (2020) Feasibility and validity of a low-cost racing simulator in driving assessment after stroke. *Geriatrics* 5:35. <https://doi.org/10.3390/geriatrics5020035>
4. Cipresso P, Giglioli IAC, Raya MA, Riva G (2018) The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. *Front Psychol* 9:2086–2086. 10.3389/fpsyg.2018.02086, 30459681[pmid]
5. Mora CE, Martín-Gutiérrez J, Añorbe-Díaz B, González-Marrero A (2017) Virtual technologies trends in education. *EURASIA J Math Sci Technol Educ* 13(2):469–486. <https://doi.org/10.12973/eurasia.2017.00626a>
6. Cheng YQ, Mansor S, Chin JJ, Abdul Karim H (2021) Driving simulator for drivers education with artificial intelligence traffic and virtual reality: a review
7. Cheng YQ, Chin JJ, Mansor S (2020) Simulator prototype for driving education, final Year Project Thesis 2020. Multimedia University
8. JPI theory test preparation and practice. <https://driveinmalaysia.com/jpj-mock-test/>
9. Official portal of road transport department of Malaysia. <https://www.jpj.gov.my/en/web/main-site/utama>
10. Unity real-time development platform. <https://unity.com/>

# The QOM Toolbox: An Object-Oriented Python Framework for Cavity Optomechanical Systems



Sampreet Kalita and Amarendra K. Sarma

**Abstract** We present an open-source Python framework to obtain classical and quantum properties of cavity optomechanical systems. Using the `BaseSystem` object, users can interface such systems to invoke built-in methods that can analyze both stationary and dynamical properties. The toolbox also contains solvers for stability and nonlinear dynamics. It is also scalable and can be used to interface many-body dynamical systems. Other features include automated loopers, configurable plotters and a graphical user interface. Here, we present an overview of the the QOM toolbox and discuss its advantages and applications in physics.

**Keywords** Optomechanics · Numerical simulation library · Open-source toolbox · Nonlinear classical dynamics · Linearized quantum dynamics

## 1 Introduction

Quantum optomechanical (QOM) systems [1, 2] have developed into a prominent platform of study in the past decade, owing to their contributions to fundamental physics and quantum technologies. Such systems are now being used to generate non-classical states for quantum information processing [3, 4], perform ultra-sensitive measurements of force and displacement [5, 6], design quantum devices such as transducers and memories [7, 8] and explore quantum phenomena in the macroscopic domain [9]. Typically modelled by a laser-driven cavity with a moveable end-mirror, the QOM formalism has also been extended to study quantum electromechanical (QEM) systems [10], hybrid optoelectromechanical (OEM) systems [11] and very recently to quantum magnomechanics (QMM) systems [12] and Bose-Einstein Condensates (BEC) systems [13].

A traditional approach to theoretically study the dynamics of such open quantum systems is by simulating the Lindblad master equation [14], which describes

---

S. Kalita (✉) · A. K. Sarma  
Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India  
e-mail: [sampreet@iitg.ac.in](mailto:sampreet@iitg.ac.in)

**Table 1** List of available numerical packages

Package	Platform	Applicable to
Qutip [21]	Python	Open quantum systems
Strawberryfields [22]	Python	Quantum optical circuits
Atom [23]	C++	Atom-light systems
QuantumOptics [24]	Julia	Open quantum systems
Scqubits [25]	Python	Superconducting qubits
QOpenQuantumTools [26]	Julia	Open quantum systems
QuantumCumulants [27]	Julia	Open quantum systems
Qotoolbox [28]	MATLAB	Open quantum systems

the evolution of the density operator for an ensemble of identical systems. Another useful approach is using Monte-Carlo methods [15] to simulate the quantum trajectory of single systems stochastically. Numerical libraries that implement such approaches are also available in multiple programming platforms (see Table 1). However, most packages do not directly support linearized QOM systems, nor contain built-in functions to analyze the multi-stable behaviour of QOM systems, nor support the calculation of nonlinear dynamical properties for many-body QOM systems in the semi-classical limit. In this paper, we present the QOM toolbox to overcome such limitations by incorporating (i) solvers for stationary as well as dynamical behaviour with estimation of multistability [16] and dynamical stability [17], (ii) methods to detect nonlinear behaviour like fixed-point and limit-cycle oscillations [18], and (iii) standard measures for classical and quantum correlation like synchronization [19] and entanglement [20]. Together with these, our toolbox also features (i) automatable loopers that can sweep through select system variables and output user-defined properties, (ii) highly configurable plotters that are templated for figures in academic journals, and (iii) a graphical user interface (GUI) to loop the variables, solve for measures and plot the results with a simple set of clicks. Backed by numerical libraries like NumPy and SciPy and featuring the highly customizable visualizations offered by Matplotlib and Seaborn, the QOM toolbox is aimed as an alternative to writing explicit code and executing repetitive blocks.

The paper is organized as follows. In Sect. 2, we introduce the structure of the toolbox and the workflow of the numerical analysis. In Sect. 3, we demonstrate the usage of the toolbox by presenting the classical and quantum behaviour of a simple QOM system. Finally, we highlight the applications of our toolbox in Sect. 4 and summarize our work in Sect. 5.

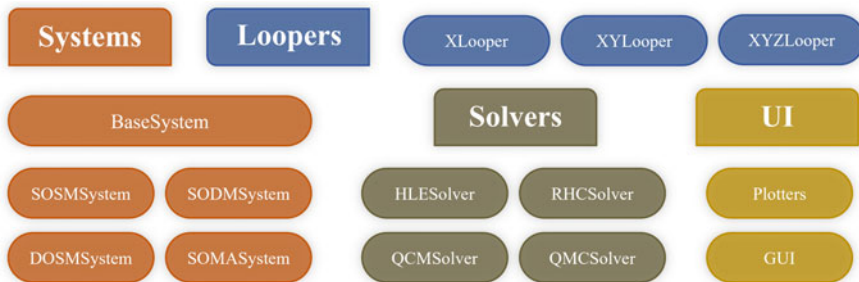
## 2 Structure and Workflow

The QOM toolbox is a wrapper-styled, scalable framework featuring multiple modules to calculate stationary and dynamical properties of linearized QOM systems. Although its modules are built to intuitively support multi-mode as well as many-body quantum optomechanical systems, the toolbox can be used to study any system following the optomechanical formalism (see Sect. 4). The key features of the toolbox can be summarized as

- Inheritable system classes with built-in numerical methods.
- Solver modules to calculate classical and quantum signatures.
- Automatically managed loops and parameter validation modules.
- Configurable visualizations without the requirement of explicit plotting.

It may be noted here that even though the QOM toolbox is built on the object-oriented approach, it supports functional and imperative approaches to obtain the same results. However, for faster implementation, the use of virtual classes is advised, the advantages of which we detail in later sections. Here, we briefly introduce the structure of the toolbox.

The QOM toolbox consists of four primary sub-packages (Fig. 1): the `loopers`, the `solvers`, the `systems` and the `ui`. A QOM system is interfaced to the toolbox through the modules in the `systems` sub-package, ideally by inheriting the `BaseSystem` class. Such an inherited class can be modelled for optomechanical systems containing any combination of optical and mechanical modes. To make things simpler, the sub-package also carries pre-defined classes for a single optical and a single mechanical mode (`SOSMSystem`), two optical modes with a single mechanical mode (`DOSMSystem`), a single optical with two mechanical modes



**Fig. 1** A representation of the QOM toolbox. The `systems` are a collection of classes templated for typical combination of modes and inherit the `BaseSystem` class. The `solvers` contain modules for the calculation of classical and quantum signatures and are called from the system classes. The `loopers` iterate over different parameters of the system and can consider upto three variables for looping using either serial or parallel computation. The `plotters` consists of the modules that display the results from a single run or a looped run. The `gui` module also contains a graphical user interface which integrates all the other modules into a single clickable window

**Table 2** Some of the methods built into the BaseSystem class and their requirements

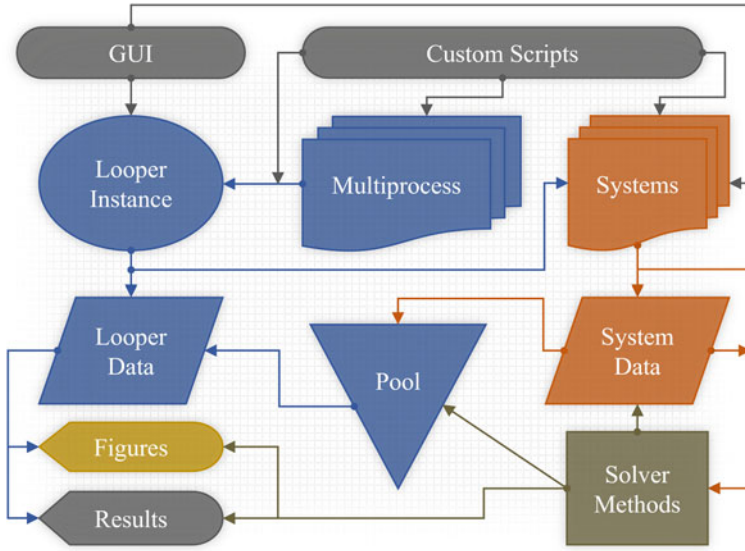
Method	Required methods/functions
get_averaged_eigenvalues	get_A, get_ivc, get_mode_rates
get_classical_phase_difference	get_ivc, get_mode_rates
get_lyapunov_exponents	get_A, get_ivc, get_mode_rates
get_mean_optical_occupancies	get_ivc, get_oss_args
get_measure_average	get_ivc, get_mode_rates
get_measure_dynamics	get_ivc, get_mode_rates
get_measure_stationary	get_ivc
get_modes_corrs_dynamics	get_ivc, get_mode_rates
get_modes_corrs_stationary	get_ivc
get_optical_stability_zone	get_ivc, get_oss_modes
get_pearson_correlation_coefficient	get_ivc, get_mode_rates

(SODMSystem) and an array of optomechanical systems with single optical modes and single mechanical modes (SOMASystem). Each of these modules support inherited methods that calculate stationary and dynamical properties of the modes and correlations, intracavity field intensities and properties that reflect its classical non-linear dynamics, by utilizing functions that return the initial values and constants (get\_ivc), the drift matrix (get\_A) and the rate of change of the classical mode amplitudes (get\_mode\_rates), etc. Table 2 lists some of BaseSystem’s built-in methods. Additionally, SOMASystem contains unique methods to calculate various properties for the collective dynamics of the classical modes.

The systems sub-package relies on the solvers sub-package to calculate dynamics as well as to determine signatures such as stability and correlations. Specifically, the HLESolver module outputs the classical amplitudes and quadrature correlations, the RHCSolver module determines stability, the QCMSolver module contains methods that calculate certain correlations like quantum entanglement or quantum synchronization and the QMCSolver module contains stochastic algorithms to obtain trajectories which mimic a single experimental run for the system.

The loopers sub-package contains wrappers that can iterate over the system parameters or a user-defined function. By default, the XLooper, the XYLooper and the XYZLooper modules nest one, two and three loops, respectively. To speed things up, they support running loops in multiple threads or parallel processes. Options for post-processing the output to calculate thresholds and gradients are also available. The ui sub-package consists of the plotters and the gui. The plotters sub-package features configurable visualizations for the results of systems and loopers. Notably, the MPLPlotter module supports a wide range of 1D, 2D and 3D plots that are templated for academic journals. It is integrated into the rest of the modules and does not require an explicit instance. However, it can also be used as a standalone module.





**Fig. 2** Workflow of the QOM toolbox. The GUI or the script calls one or more instances of the looper classes or directly calls a system class. Each system either solves for or loads the dynamics or calls other solver classes for method-specific requirements. Each looper object spawns a system instance for every set of parameters and collects their results. The results are then displayed in the console or through figures

Once the system is interfaced, the user can run scripts to invoke methods defined in the `loopers` or the `systems`. These methods can also be called through the `gui` sub-package, which simplifies everything to a set of drop-down lists and buttons. We illustrate this complete workflow of the toolbox in Fig. 2. It may be mentioned here that multiprocessing is currently supported only through custom scripts and is not integrated into the GUI.

### 3 Formalism and Simulation

The Hamiltonian of a simple QOM system can be written as [29]:

$$H = -\hbar\Delta_0 a^\dagger a + \hbar\omega_m b^\dagger b - \hbar g_0 a^\dagger a (b^\dagger + b) + i\hbar A_l (a^\dagger - a), \quad (1)$$

where  $a$  ( $b$ ) is the annihilation operator of the optical (mechanical) mode,  $\Delta_0$  is the laser detuning,  $\omega_m$  the mechanical resonance frequency,  $A_l$  the laser amplitude and  $g_0$  the optomechanical coupling constant. It can be seen here that the interaction term  $-\hbar g_0 a^\dagger a (b^\dagger + b)$  is inherently nonlinear. This nonlinear optomechanical coupling is at the root of all interesting phenomena observed in QOM systems. Although



different proposals involving hybrid setups have attempted to enhance the coupling strength, the strong coupling regime is difficult to achieve experimentally for a wide range of systems and therefore,  $g_0$  is inherently weak. To circumvent this, most studies utilize a semi-classical treatment by driving the systems with strong coherent fields, which leads to large classical amplitudes for the intra-cavity photon number. This results in two preferable outcomes. Firstly, the nonlinear classical dynamics ( $\alpha = \langle a \rangle$ ,  $\beta = \langle b \rangle$ ) containing well-defined trajectories can now be separated from the linearized quantum dynamics ( $\delta a = a - \alpha$ ,  $\delta b = b - \beta$ ) containing stochastic noises [14]. In the Markovian limit, the noises can be regarded as zero-mean and delta-correlated, and one can obtain quantum correlations, i.e., the second moments of the quantum fluctuations [1]. Secondly, the effective coupling strengths get enhanced by a factor proportional to the classical amplitude of the optical modes [2]. This leads to a stronger effective optomechanical coupling in the linearized regime.

Under this linearized approximation, the time-evolution of the complete system can be expressed by using the two sets of variables:

1. Classical complex-valued modes ( $\alpha$ ,  $\beta$ ) obeying coupled differential equations

$$\frac{d\alpha}{d\tau} = -\left(\frac{\kappa'}{2} - i\Delta'_0\right)\alpha + ig'_0\alpha(\beta^* + \beta) + A'_I, \quad (2)$$

$$\frac{d\beta}{d\tau} = -\left(\frac{\gamma'}{2} + i\right)\beta + ig'_0\alpha^*\alpha, \quad (3)$$

where  $\kappa$  ( $\gamma$ ) are the optical decay (mechanical damping) rates and the prime superscript denotes normalization with  $\omega_m$  with  $\tau = \omega_m t$ .

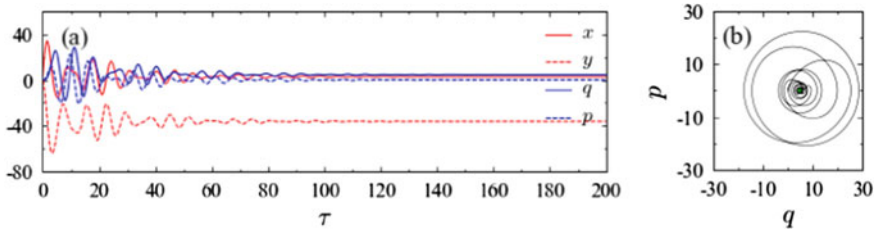
2. Quantum real-valued quadratures ( $X = (\delta a^\dagger + \delta a)/\sqrt{2}$ ,  $Y = i(\delta a^\dagger - \delta a)/\sqrt{2}$ ,  $Q = (\delta b^\dagger + \delta b)/\sqrt{2}$ ,  $P = i(\delta b^\dagger - \delta b)/\sqrt{2}$ ) whose correlations obey a simple matrix differential equation

$$\frac{d\mathbf{V}}{d\tau} = \mathbf{A}'\mathbf{V} + \mathbf{V}\mathbf{A}'^T + \mathbf{D}', \quad (4)$$

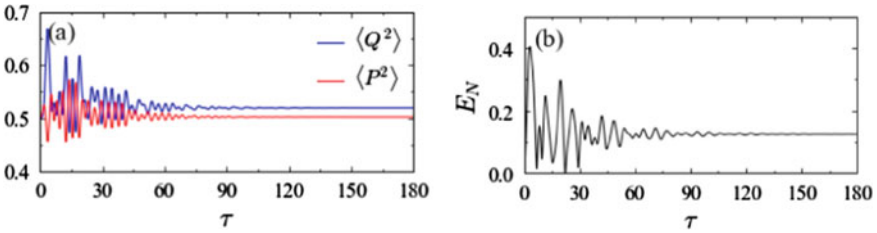
where  $\mathbf{V}_{jk} = \frac{1}{2}\langle u_j u_k + u_k u_j \rangle$ , with  $\mathbf{u} = (X, Y, Q, P)^T$ , the normalized noise matrix  $\mathbf{D}' = \text{Diag}[\kappa'/2, \kappa'/2, \gamma'(n_{th} + 1/2), \gamma'(n_{th} + 1/2)]$  with  $n_{th}$  being the thermal mechanical vibration quanta, and the normalized drift matrix  $\mathbf{A}'$  is

$$\begin{pmatrix} -\frac{\kappa'}{2} & -\Delta' & -2G'_{0I} & 0 \\ \Delta' & -\frac{\kappa'}{2} & 2G'_{0R} & 0 \\ 0 & 0 & -\frac{\gamma'}{2} & \omega'_m \\ 2G'_{0R} & 2G'_{0I} & -\omega'_m & -\frac{\gamma'}{2} \end{pmatrix}, \quad (5)$$

where  $\Delta' = \Delta'_0 + g'_0(\beta^* + \beta)$  and  $G'_{0R}$  ( $G'_{0I}$ ) are the real (imaginary) parts of  $G'_0 = g'_0\alpha$ .



**Fig. 3** Time-evolution of **a** classical amplitudes ( $x = (\alpha^* + \alpha)/\sqrt{2}$ ,  $y = i(\alpha^* - \alpha)/\sqrt{2}$ ,  $q = (\beta^* + \beta)/\sqrt{2}$ ,  $p = i(\beta^* - \beta)/\sqrt{2}$ ) and **b** phase-space trajectory of the mechanical motion given by Eqs. 2–3 for  $\tau_{\max} = 10^3$ . The last 10 cycles ( $2\pi \times 10 \approx 628$  points) in **b** are marked in green. Here, the modes settle down into stationary values for the parameters  $A'_l = 25.0$ ,  $\Delta'_0 = -1.0$ ,  $g'_0 = 0.005$ ,  $\gamma' = 0.005$ ,  $\kappa' = 0.15$  and  $n_{th} = 0.0$



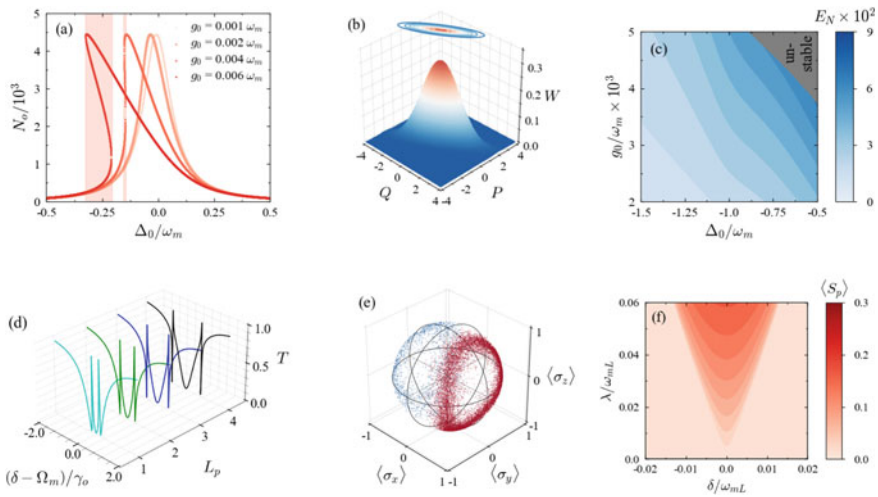
**Fig. 4** Time-evolution of **a** quadrature correlations ( $\langle Q^2 \rangle$ ,  $\langle P^2 \rangle$ ) of the mechanical mode governed by Eqs. 4–5 and **b** optomechanical entanglement computed using the logarithmic negativity measure [20]. Parameters used are the same as in Fig. 3

Equations 2 and 3 constitute the rate of change of the mode amplitudes returned by the `get_mode_rates` method, whereas the `get_A` method returns the matrix **A** in Eq. 5. Once interfaced, the `get_modes_corrs_dynamics` or the `get_measure_dynamics` methods can be used to obtain the dynamical values with "measure\_type" set to "mode\_amp" and "idx\_e" set to [0, 1] (0 and 1 are the indices for  $\alpha$  and  $\beta$ ). We plot the corresponding dynamics in Fig. 3. These methods call the `HLESolver` module to obtain the dynamics numerically.

To obtain the quantum correlations between the quadratures obtained from the elements of the correlation matrix **V**, the value of "measure\_type" can be set to "corr\_ele", with the corresponding indices of the element in "idx\_e". For example, to obtain  $V_{22} = \langle Q^2 \rangle$  and  $V_{33} = \langle P^2 \rangle$ , one can set it as  $[(2, 2), (3, 3)]$ . Figure 4a shows the corresponding output. In Fig. 4b, we plot the logarithmic negativity measure for the quantification of optomechanical entanglement by setting "measure\_type" as "entan\_ln" and "idx\_e" as (0, 1) (0 and 1 denote the optical and mechanical modes respectively).

## 4 Applications and Advantages

As discussed in Sect. 2, the usage of the QOM toolbox extends beyond the calculation of dynamics. It can be used to obtain specialized properties like bistability, wigner distributions, dynamical stability, collective dynamics, entanglement and synchronization measures, etc. (see Fig. 5). The methods pertaining to all these are well-documented and we have provided a few examples featuring simple to complex implementations and reproduced papers [30]. Our toolbox can also be used to study other hybrid system that can be described by a linearized Hamiltonian. The primary benefits of our toolbox are (i) reduction of runtimes due to linearization, (ii) generalized and scalable built-in methods for a diverse set of properties and (iii) simple dictionary-based parameter models without requirement of code blocks for looping and plotting.



**Fig. 5** Applications of the QOM toolbox. **a** Intracavity photon number ( $N_o$ ) in an end-mirror QOM system for different laser detunings ( $\Delta_0$ ) and optomechanical coupling strengths ( $g_0$ ) normalized by the mechanical frequency ( $\omega_m$ ) [29]. The numerical values are obtained using the `get_mean_optical_occupancies` method. The shaded areas are multi-stable. **b** Wigner distribution ( $W$ ) of the mechanical mode in a membrane-in-the-middle QOM system [31] obtained using the `get_wigner_single_mode` method. **c** Stationary quantum entanglement ( $E_N$ ) between the optical and mechanical modes of an end-mirror QOM system for different laser detunings ( $\Delta_0$ ) and optomechanical coupling strengths ( $g_0$ ) normalized by the mechanical frequency ( $\omega_m$ ) [32]. The entanglement is obtained using the `get_measure_stationary` method after stability analysis using the `get_optical_stability_zones` methods that invokes the `RHCSolver` module. **d** Transmission ( $T$ ) of a probe field (detuning  $\delta$  normalized by optical decay  $\gamma_o$ ) with variation in the winding number ( $L_p$ ) around mechanical frequency  $\Omega_m$  in a ring-BEC configuration [13]. **e** Bloch sphere depicting the qubit bistability in a hybrid QOM system [33] obtained using the `QMCSolver` module. **f** Averaged quantum phase synchronization ( $\langle S_p \rangle$ ) between the mechanical modes of two optically coupled QOM systems for different coupling strengths ( $\lambda$ ) and mechanical frequency differences ( $\delta$ ) normalized by the frequency of the first mechanical oscillator ( $\omega_{mL}$ ) [34]. Here, the `get_measure_average` method is used. All figures are obtained using the toolbox

## 5 Conclusion

We presented a Python framework containing modular methods to simulate linearized quantum dynamics alongside nonlinear classical dynamics. With a simple example, we demonstrated the calculation of dynamical properties and extended its usage to hybrid architectures. Although our toolbox is highly scalable, the time required to simulate  $N$ -body systems still scales as  $\mathcal{O}(N^2)$ . Further, the space complexity scales with the dimension of simulated time  $t$ . To overcome this, we are currently exploring predictive models that rely on neural ordinary differential equations [35]. Another bottleneck is CPU-based parallelization. Whilst current supercomputers support a good number of parallel threads, such computing facilities are not accessible to everyone and users are limited to the handful of CPU cores. Therefore, we are currently testing a GPU-based implementation of the toolbox in order to provide highly parallelizable loopers. Together with this, constant efforts are being made to further modularize and smoothen the user experience.

**Acknowledgements** S.K. would like to thank Subhadeep Chakraborty, Pratyusha Chowdhury, Ambareesh Sahoo, Roson Nongthombam and Shah Saumya Amit for useful discussions, and the PMRF scheme, MHRD, Govt. of India for financial support.

## References

1. Bowen WP, Milburn GJ (2015) Quantum Optomechanics, Taylor & Francis. <https://books.google.co.in/books?id=xqlcrgEACAAJ>
2. Aspelmeyer M, Kippenberg TJ, Marquardt F (2014) Rev Mod Phys 86:1391. <https://link.aps.org/doi/10.1103/RevModPhys.86.1391>
3. Braunstein SL, van Loock P (2005) Rev Mod Phys 77:513. <https://link.aps.org/doi/10.1103/RevModPhys.77.513>
4. Mari A, Eisert J (2009) Phys Rev Lett 103:213603. <https://link.aps.org/doi/10.1103/PhysRevLett.103.213603>
5. Tse M, et al (2019) Phys Rev Lett 123:231107. <https://link.aps.org/doi/10.1103/PhysRevLett.123.231107>
6. Bai CH, Wang DY, Zhang S, Liu S, Wang HF (2020) Phys Rev A 101:053836. <https://link.aps.org/doi/10.1103/PhysRevA.101.053836>
7. de Moraes Neto GD, Andrade FM, Montenegro V, Bose S (2016) Phys Rev A 93:062339. <https://link.aps.org/doi/10.1103/PhysRevA.93.062339>
8. Chan J, Alegre TPM, Safavi-Naeini AH, Hill JT, Jrause A, Groblacher S, Aspelmeyer M, Painter O (2011) Nature 478:89. <https://www.nature.com/articles/nature10461>
9. Westphal T, Hepach H, Pfaff J, Aspelmeyer M (2021) Nature 591:225. <https://www.nature.com/articles/s41586-021-03250-7>
10. Ekinci KL, Roukes ML (2005) Rev Sci Instrum 76:061101. <https://aip.scitation.org/doi/10.1063/1.1927327>
11. Midolo L, Schliesser A, Fiore A (2018) Nat Nanotechnol 13:11. <https://doi.org/10.1038/s41565-017-0039-1>. <https://www.nature.com/articles/s41565-017-0039-1>
12. Yuan H, Cao Y, Kamra A, Duine RA, Yan P (2022) Phys Rep 965:1. <https://doi.org/10.1016/j.physrep.2022.03.002>, <https://www.sciencedirect.com/science/article/pii/S0370157322000977>

13. Kumar P, Biswas T, Feliz K, Kanamoto R, Chang MS, Jha AK, Bhattacharya M (2021) *Phys Rev Lett* 127:113601. <https://link.aps.org/doi/10.1103/PhysRevLett.127.113601>
14. Gardiner CW, Collett MJ (1985) *Phys Rev A* 31:3761. <https://link.aps.org/doi/10.1103/PhysRevA.31.3761>
15. Brun TA (2002) *Am J Phys* 70(7):719. <https://aapt.scitation.org/doi/10.1119/1.1475328>
16. Dorsel A, McCullen JD, Meystre P, Vignes E, Walther H (1983) *Phys Rev Lett* 51:1550. <https://link.aps.org/doi/10.1103/PhysRevLett.51.1550>
17. DeJesus EX, Kaufman C (1987) *Phys Rev A* 35:5288. <https://link.aps.org/doi/10.1103/PhysRevA.35.5288>
18. Roque TF, Marquardt F, Yevtushenko OM (2020) *New J Phys* 22:013049. <https://iopscience.iop.org/article/10.1088/1367-2630/ab6522>
19. Mari A, Farace A, Didier N, Giovannetti V, Fazio R (2013) *Phys Rev Lett* 111:103605. <https://link.aps.org/doi/10.1103/PhysRevLett.111.103605>
20. Vitali D, Gigan S, Ferreira A, Boehm H, Tombesi P, Guerreiro A, Vedral V, Zeilinger A, Aspelmeyer M (2007) *Phys Rev Lett* 98(3):030405. <https://link.aps.org/doi/10.1103/PhysRevLett.98.030405>
21. Johansson JR, Nation PD, Nori F (2013) *Comput Phys Commun* 184:1234. <https://doi.org/10.1016/j.cpc.2012.11.019>, <https://www.sciencedirect.com/science/article/pii/S0010465512003955>
22. Killoran N, Izzac J, Quesada N, Bergholm V, Amy M, Weedbrook C (2019) *Quantum* 3:129. <https://doi.org/10.22331/q-2019-03-11-129>, <https://quantum-journal.org/papers/q-2019-03-11-129/>
23. Javanainen J (2017) *Comput Phys Commun* 212:1. <https://doi.org/10.1016/j.cpc.2016.09.017>, <https://www.sciencedirect.com/science/article/pii/S0010465516302880>
24. Krämer S, Plankensteiner D, Ostermann L, Ritsch H (2018) *Comput Phys Commun* 227:109. <https://doi.org/10.1016/j.cpc.2018.02.004>, <https://www.sciencedirect.com/science/article/pii/S0010465518300328>
25. Groszkowski P, Koch J (2021) *Quantum* 5:583. <https://doi.org/10.22331/q-2021-11-17-583>, <https://quantum-journal.org/papers/q-2021-11-17-583>
26. Chen H, Lidar DA (2022) *Commun Phys* 5:112. <https://doi.org/10.1038/s42005-022-00887-2>, <https://www.nature.com/articles/s42005-022-00887-2>
27. Plankensteiner D, Hotter C, Ritsch H (2022) *Quantum* 6:617. <https://doi.org/10.22331/q-2022-01-04-617>, <https://quantum-journal.org/papers/q-2022-01-04-617/>
28. Tan SM (1999) *J Opt B: Q Semiclass Opt* 1(4):424. <https://iopscience.iop.org/article/10.1088/1464-4266/1/4/312>
29. Sarma AK, Kalita S (2022) Tutorial: Cavity Q Optomech. <https://doi.org/10.48550/arXiv.2211.02596>, <https://arxiv.org/abs/2211.02596>
30. Kalita S (2018) The QOM toolbox documentation. <http://sampreet.github.io/qom/>. [Online; updated November 16, 2022]
31. Banerjee P, Kalita S, Sarma AK (2022) Mechanical squeezing in quadratically-coupled optomechanical systems. <https://doi.org/10.48550/arXiv.2210.00510>, <https://arxiv.org/abs/2210.00510>
32. Sarma AK, Chakraborty S, Kalita S (2021) *AVS Q Sci* 3:015901. <https://avs.scitation.org/doi/10.1116/5.0022349>
33. Nongthombam R, Kalita S, Sarma AK (2022) Synchronization of a superconducting qubit to an optical field mediated by a mechanical resonator. <https://doi.org/10.48550/arXiv.2205.12214>, <https://arxiv.org/abs/2205.12214>
34. Kalita S, Chakraborty S, Sarma AK (2021) *J Phys Commun* 5(11):115006. <https://iopscience.iop.org/article/10.1088/2399-6528/ac3204>
35. Jiahao TZ, Hsieh MA, Forgoston E (2021) *Chaos: An Interdiscip J Nonlinear Sci* 31(6):111101. <https://aip.scitation.org/doi/10.1063/5.0065617>

# Preserving Filipino Native Dishes Using Android-Based Application: A Heritage Cooking Tutorial



Aries M. Gelera, Alyssa Joi A. Gonzales, Bryan James V. Torres,  
and Marvin G. Sison

**Abstract** As a result of contemporary cooking methods, many of the local cuisines in the Philippines are starting to be forgotten. Today, some restaurants have begun to serve native dishes, but in small quantities. In some restaurants, the serving of native dishes is primarily centered on the most popular. With that, the vast majority of historical foods that shaped Philippine culture are now considered endangered and might be lost forever. Cooking tutorials using mobile devices are popular nowadays. Applications such as this help individuals easily learn how to cook certain dishes. In this study, the researchers developed a cooking tutorial mobile application that can preserve Filipino historical cuisines to help Filipino get used to different dishes that are part of their heritage. The goal of the study to provide instructions on how to properly prepare traditional dishes using actual cooking time and voice guide was achieved. To successfully develop the application tutorial, the researchers made use of the unified method, since this model focuses on the exchange of ideas and messages taking place between the user and the application. Moreover, statistical analysis of data revealed that the application was helpful and effective for people who wanted to learn how to cook native dishes. Users appreciated the application because it could explain the step-by-step process on cooking Filipino native dishes. Additionally, the application offers information on each recipe's historical significance. People who enjoy cooking will find the application highly helpful, user-friendly, and simple to use.

**Keywords** Cuisines · Heritage · Native dishes · Unified model · Tutorial · Mobile app · Cooking time · Voice guide

---

A. M. Gelera (✉) · A. J. A. Gonzales · B. J. V. Torres · M. G. Sison  
Department of Computer Studies, Cavite State University-CCAT Campus, Rosario, Cavite,  
Philippines  
e-mail: [aries.gelera@cvsu.edu.ph](mailto:aries.gelera@cvsu.edu.ph)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information  
and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_49](https://doi.org/10.1007/978-981-99-3091-3_49)

591

# 1 Introduction

There are more than 80 tribes in the Philippines, and each has its own unique cultural differences. Gender, religion, ethnicity, and food differences set the Filipinos apart from other races around the world [1]. Prior to the colonization of other countries, the Filipino way of life was modest. They get food simply by rowing out to sea in their little bancas, wading knee-deep in rice paddies, planting in their backyard, and hunting in the woods. Everything that nature had to offer, including fish, fresh produce, fruits, and many other foods, was incorporated in the Filipinos' conception of culture.

The Spanish, Chinese, and American civilizations, among others, have all had an impact on Philippine food. From that time, food in the Philippines evolved. The food in the Philippines is frequently spicy and has a lot of flavors. Rice, fish, poultry, hog, beef, vegetables, and fruits are the most typical items used in Filipino cuisine. Filipino food is often cooked with soy sauce, vinegar, and garlic [2].

Cuisines from the Philippines have been passed down from generation to generation. And as the latest generation, it is important to understand and learn the history of Filipino foods, as it is considered one that shapes the cultural identity of the Philippines. Filipino culture is shaped by their food. In addition to the food itself, it is made even more remarkable by the treasured memories that are shared with whomever they are with [3]. According to a PhilStar reporter named Mr. Juan, many traditional and cultural Filipino cuisines fear extinction as the Philippines enters a new period. Some of these delicacies have already been lost to time, and it would be terrible for today's generation to not know how to prepare them [4].

Today, Philippine cuisine continues to evolve as new techniques and styles of cooking find their way into one of the most active melting pots of Asia. Philippine dishes are one of the most popular dishes in the world and have been known for centuries [5]. Furthermore, Filipino foods have been offered and served not only in Asia but also on other continents. In the continuing development of modern cooking techniques, there are more changes in how we can learn to cook a dish. In today's new generation, technology helps us a lot in almost everything, and because of technology, cooking can be learned instantly.

The study's primary goal is to familiarize Filipinos with the variety of local cuisines throughout the Philippines. Cooking lessons for customary and ethnic foods will be made available through this study. By introducing the newest generation to authentic Filipino cuisine, this study will also benefit them. This paper serves as an example of the value of traditional and cultural foods in the Philippines.

## 2 Methodology

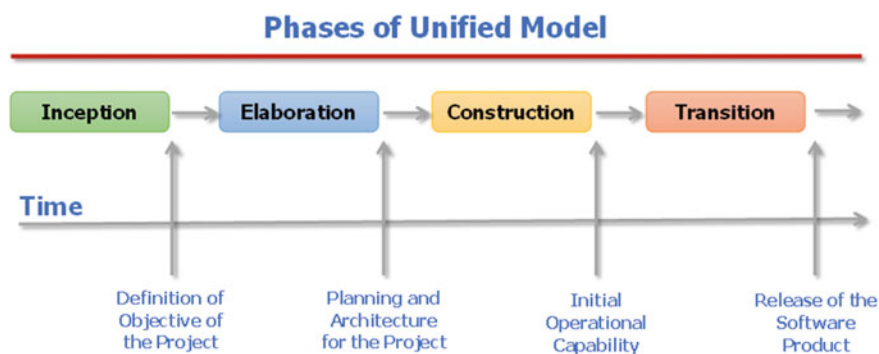
The researchers made use of the Unified Model to develop the heritage cooking due to the fact that this methodology specializes in handling the communication of ideas and communications that go both ways from sender to receiver and vice versa [6]. Because of its ability to specify, visualize, construct, and document software systems, the unified model is the best model for developing applications. The model proved to be extremely successful for object-oriented system development [7]. The four stages of this methodology are inception, elaboration, construction, and transition [8] (Fig. 1).

### 2.1 Inception Phase

The project scope and boundary constraints were determined during this phase, and developers also developed an approximation of the system's vision. The developers obtained information from books, journals, and online references to analyze and categorize the elements that were inferred when creating the application's general requirements. Interviews with subject-matter experts and brainstorming sessions regarding the potential outcomes of the application were conducted by the developers to gather data. Additionally, during this stage, the developers can also find problems that need to be solved.

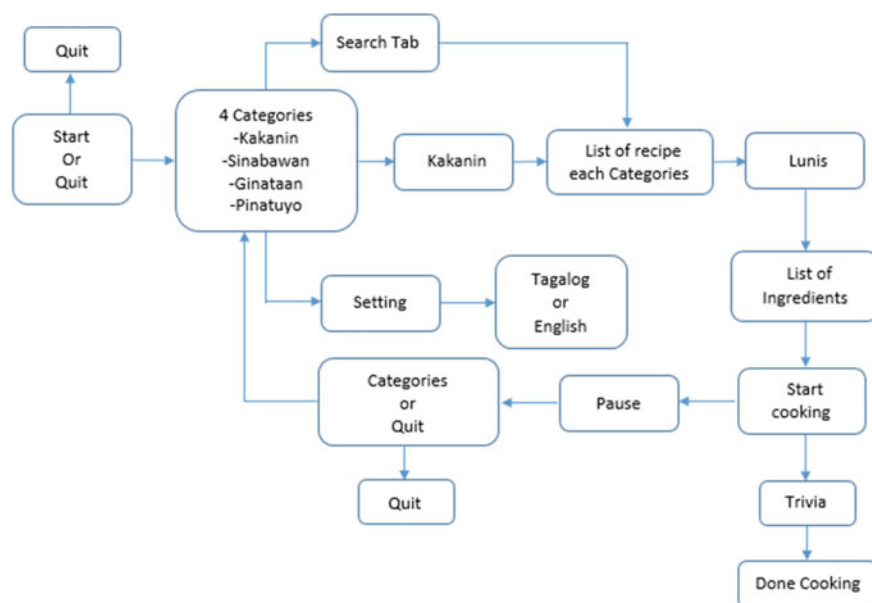
### 2.2 Elaboration Phase

In order to provide a solid foundation for the majority of the design and implementation efforts made in the construction phase, the baseline architecture of the system



**Fig. 1** Unified model





**Fig. 2** Process design

must be established during the elaboration phase. All of the information obtained during this phase was examined by the developers. They also used this step to develop the application in the best possible way. Moreover, developers simulated the specifications using different process flow diagrams that served as a pattern in developing the application. Part of it was the hardware requirements, software requirements, and development application that were used in order to build the Android application. Finally, this phase also allowed the developers to easily identify the possible risks that might occur during the development of the application. Figure 2 on the next page shows the process design of the application.

## 2.3 Construction Phase

The building phase's objective is to categorize the remaining needs and finish the application development using Unity 3D and Android Studio. During this phase, the developers used the Java programming language to develop the application and SQL Lite for its database. The coding and construction of the application represented the phase's turning point. In addition, test cases are run to make sure the program is operating correctly.

## 2.4 Transition Phase

Making sure that software is accessible to its end users is the main goal of the transition phase. In the transition phase, which is the last phase of the unified model, multiple iterations were applied. Testing the product in advance of release and making minor tweaks in response to user input are all part of the transition phase. Lastly, to accommodate its users' needs, minor improvements are made depending on user feedback.

## 3 Results and Discussions

### 3.1 Application Features

Figure 3 shows the home page of the application. This part of the application displays the “Help” button, the “Settings” button, the “Application logo,” and the “Magsimula” and “Umalis” buttons. Figure 4 shows the category page of the application. This displays the different categories in the application, such as Pinatuyo, Ginataan, Sinabawan, and Kakanin. Figure 5 shows a list of recipes. This feature of the application can display a list of recipes based on the selected category. Samples are Ginisang Bihon sa Bariles, Laswa, Kinulob, Ginarep, and Pamplina. Also, Fig. 6 shows the actual cooking of the recipe, including a time display for when to add ingredients and when to finish cooking.

**Fig. 3** Home page



**Fig. 4** Category page



**Fig. 5** List of recipes



**3.2 Testing**

The application underwent testing to ensure that the cooking tutorial was accurate and of high quality.

Table 1 shows the results of data in terms of accuracy that was gathered from the different trials conducted. The accuracy test of the application was based on

Fig. 6 Actual cooking



comparing the actual cooking time of the application with the actual time on the analog clock. This is to determine if the application’s cooking time is the same as the real-time cooking experience. The trials conducted prove that the application has a 100% accuracy rate from the expected output to the actual cooking time.

Table 1 Accuracy test based on actual cooking time

Trial	Expected output	Cooking of time	Remarks
1	8:00–8:03	3 min	Accurate
2	9:34–10:10	36 min	Accurate
3	10:20–11:20	1 h	Accurate
4	11:55–12:05	10 min	Accurate
5	12:16–12:23	7 min	Accurate
6	8:00–8:05	5 min	Accurate
7	7:15–7:35	20 min	Accurate
8	11:20–11:55	35 min	Accurate
9	12:05–12:16	11 min	Accurate
10	10:10–10:20	10 min	Accurate
11	8:00–8:03	3 min	Accurate
12	9:34–10:10	36 min	Accurate
13	10:20–11:20	1 h	Accurate
14	11:55–12:05	10 min	Accurate
15	12:16–12:23	7 min	Accurate

**Table 2** Descriptive interpretation of the mean

Score	Interpretation
4.51–5.00	Excellent
3.51–4.50	Very satisfactory
2.51–3.50	Satisfactory
1.51–2.50	Unsatisfactory
1.00–1.50	Needs improvement

### 3.3 Evaluation

A software evaluation tool compliant with ISO/IEC 25,010 was used to assess the application. Five (5) HRM instructors with expertise in the culinary arts and fifteen (15) HRM students each completed the evaluation forms. Additionally, it was assessed as an IT specialist in the field of developing mobile applications by six (6) IT faculty members. Table 2 was used to tabulate, compute, and interpret their results.

Table 3 shows the interpretation of data terms in terms of usability that was gathered from the HRM students and HRM instructors of the CvSU-CCAT campus participants of the evaluation. The developers computed that the software obtained a weighted mean of 4.65, which has an equivalent interpretation of excellent. This means that the HRM students and instructors believed that the software fully meets and far exceeds their expectations in terms of usability. Theoretically, the researchers can conclude that the HRM students and professors agreed that the software is easy to use.

Table 4 displays the responses provided by the CvSU-CCAT campus HRM professors and students who participated in the evaluation regarding their level of satisfaction. The program's weighted mean was calculated by the developers to be 4.8, which is equivalent to excellent. This indicates that according to the HRM students and instructors who participated in evaluating the application, it completely fulfills and greatly surpasses expectations in terms of satisfaction.

**Table 3** General rating of the applications' usability

Criterion	Mean	Interpretation
Appropriateness	4.65	Excellent
Learnability	4.65	Excellent
Operability	4.50	Very satisfactory
User interface aesthetics	4.50	Very satisfactory
Accessibility	4.60	Excellent
Mean	4.65	Excellent

**Table 4** General rating of users' satisfaction

Criterion	Mean	Interpretation
Usefulness	4.80	Excellent
Trust	4.45	Very satisfactory
Pleasure	4.45	Very satisfactory
Comfort	4.55	Excellent
Total	4.80	Excellent

## 4 Conclusions and Recommendations

Unlike other cooking tutorials, the developed heritage cooking uses actual cooking time instruction on how to cook certain native dishes. This actual cooking time instruction will instruct user on when to start cooking, add ingredients, and finish cooking. The developers concluded that the application was helpful and effective for people who wanted to learn how to cook native Filipino dishes. The users appreciated the application because it could explain the step-by-step process of how to cook the native dishes of each recipe and provide trivia about the recipe they picked. The application is user-friendly, easy to understand, and useful to people who love cooking, especially those aligned with the cooking industry. This also determined that the system met its objectives with the following features in the application:

- The application can give accurate cooking time along with voice directions for each recipe's preparation.
- The application allows the user to change the language from Tagalog to English and vice versa.
- The application can name the dishes, search any region in the Philippines for a recipe, and give background information on the history of certain local delicacies.

To further improve the awareness on the Filipino native dishes, it is recommended to conduct a technology acceptance study to confirm the level of users' awareness of the historical foods in the Philippines. It is also suggested that the cooking application should use advanced Alexa features for a voice user interface, enabling users to engage with the application's content using natural voice services.

## References

- Diversify (2013) Retrieved from diversifyoss.com: <https://diversifyoss.com/newsroom/undersanding-filipino>
- Nataly (2020) philippinecuisine.net. Retrieved from Philippine cuisine and culture: <https://philippinecuisine.net/filipinos-cuisine/>
- Juan CU (2021) Forgotten foods. Retrieved from philstar.com: <https://www.philstar.com/opinion/2021/04/06/2089182/forgotten-foods/>
- Peneva C (2020) Retrieved from discoverthephilippines.info: <https://discoverthephilippines.info/filipino-food-culture-and-traditions/>

5. [www.vigattintourism.com](http://www.vigattintourism.com) (2014) Retrieved from [www.vigattintourism.com](http://www.vigattintourism.com): <https://www.vigattintourism.com/tourism/articles/Philippine/>
6. Osis J, Donins U (2017) Software designing with unified modeling language driven approaches. Topological UML Model 53–82. <https://doi.org/10.1016/B978-0-12-805476-5.00002-2>
7. Fitsilis P, Gerogiannis VC, Anthopoulos L (2013) Role of unified modelling language in software development in Greece—results from an exploratory study. IET J 8(4):143–153
8. (2002) Methods and tools for software architecture. Comput Aided Chem Eng 11:229–266. [https://doi.org/10.1016/S1570-7946\(02\)80013-X](https://doi.org/10.1016/S1570-7946(02)80013-X)
9. Garrido-Merchán EC, Albarca-Molina A (2018) Suggesting cooking recipes through simulation and Bayesian optimization. In: Yin H, Camacho D, Novais P, Tallón-Ballesteros A (eds) Intelligent data engineering and automated learning—IDEAL 2018. IDEAL 2018. Lecture notes in computer science, vol 11314. Springer, Cham
10. Doman K, Kuai CY, Takahashi T, Ide I, Murase H (2011) Video CooKing: towards the synthesis of multimedia cooking recipes. In: Lee KT, Tsai WH, Liao HYM, Chen T, Hsieh JW, Tseng CC (eds) Advances in multimedia modeling. MMM 2011. Lecture notes in computer science, vol 6524. Springer, Berlin, Heidelberg
11. Schäfer U, Arnold F, Ostermann S, Reifers S (2013) Ingredients and recipe for a robust mobile speech-enabled cooking assistant for German. In: Timm IJ, Thimm M (eds) KI 2013: advances in artificial intelligence. KI 2013. Lecture notes in computer science, vol 8077. Springer, Berlin, Heidelberg
12. Kawano Y, Sato T, Maruyama T, Yanai K (2013) [Demo paper] mirurecipe: a mobile cooking recipe recommendation system with food ingredient recognition. IEEE Int Conf Multimedia Expo Workshops (ICMEW) 2013:1–2. <https://doi.org/10.1109/ICMEW.2013.6618222>
13. Garvin T, Chiappone A, Boyd L, Stern K, Panichelli J, Edwards Hall L, Yaroch A (2019) Cooking matters mobile application: a meal planning and preparation tool for low-income parents. Public Health Nutr 22(12):2220–2227. <https://doi.org/10.1017/S1368980019001101>

# Development of a Web-Based Graduate Tracer Information System with Data Analytics



Karlo Jose E. Nabablit and Edgardo S. Dajao

**Abstract** Graduate tracer study provides information that helps academic institutions in analyzing the performance of alumni employed in different organizations and industries. The conventional method struggles in evaluating the gathered data from the graduates. Thus, this research focused on the development of a web-based graduate tracer information system for a State University in the Philippines that ease the problems from the non-optimal collection and unsystematic compilation of information. The Rapid Application Development (RAD) model was used to develop the web-based information system. The Hypertext Markup Language (HTML), Bootstrap CSS framework, and JavaScript scripting language were used to develop the end-user interface, while Hypertext Preprocessor (PHP) and MySQL were used for the functions and modules. Alpha test was carried out for the system testing, and the result showed that it effectively manages alumni's information and capable of generating relevant reports for the end-users. The developed system provides descriptive and correlational data analytics, which can also be used to investigate the determinants of early employment.

**Keywords** Tracer · Data analytics · Graduate · Information system · Alumni tracking · Rapid application development

---

K. J. E. Nabablit (✉)

Department of Computer Studies, Cavite State University-CCAT Campus, Rosario, Cavite, Philippines

e-mail: [karlojose.nabablit@cvsu.edu.ph](mailto:karlojose.nabablit@cvsu.edu.ph)

E. S. Dajao

Graduate School of Engineering, Pamantasan Ng Lungsod Ng Maynila, City of Manila, Philippines



## 1 Introduction

Higher academic institutions play an essential role in nation-building by producing graduates with high levels of intellectual thinking and behavioral and technical competencies aligned with national and international educational standards. In the Philippines, the Commission on Higher Education mandates HEIs to keep track of their graduate's performance to determine the quality and the extent of functionality they deliver to the graduates. Tracer analysis is one type of empirical research that gives significant information for assessing the effects of education and training at a particular institution of higher education. In the framework of quality assurance, this information may be used for the institution's ongoing growth [1]. The employability of the graduates is one of the measures of success of higher educational institutions, making it an integral component in the light of quality education. Higher education accrediting organizations like the Accrediting Agency for Chartered Colleges and Universities in the Philippines (AACCUP) require graduate tracer studies.

Cavite State University–CCAT Campus is a higher education institution serving the diverse communities of Cavite and its nearby provinces. Since its integration into the Cavite State University system under CHED Memo No. 27, s. 2000, the campus has offered various courses that cater to the needs of its stakeholders [2]. In pursuing the university's mandate, to provide good-quality education, it is imperative to regularly conduct graduate tracer studies to assess its graduates' early labor market employment. The university's researchers utilize Google Forms to collect graduates' data. It does not have any mechanism to validate data input which permits duplicate data entry and a response from unknown respondents. Data processing involves data cleaning and filtering. Since it deals with a large amount of data and is being validated manually, the process is considered to be tedious and cumbersome. Despite carefully preparing the data, it is still prone to human error, which can compromise the accuracy of the result. The non-optimal collection and unsystematic compilation of alum data suggest a need for a systematic system for graduate tracing. The study focuses on developing an ICT solution to optimize and streamline graduate data acquisition and processing. It also aims to provide analytical reports commensurate to items needed for the accreditation process.

## 2 Literature Review

### 2.1 *Graduate Tracer Study*

Tracer studies are the surveys utilized by higher education institutions to track their alums. These questionnaires are typically designed and conducted online to a random sample of alumni one to two years following their graduation [3]. It is essential to monitor the activities and accomplishments of graduates as it is attributed to sustaining the quality of service to society as expressed by the university's vision,

mission, and goals. Frequently, the most important component of such institutional tracer studies is the input from curriculum development and other areas of enhancing study conditions and provisions [4]. Information about job search, employment conditions, and work is taken as signals of the graduates' labor market from different study programs. Moreover, the tracer study offers quantitative structural data on employment and career, the nature of work and related abilities, and the professional orientation and experiences of their graduates [5].

Graduate tracing is descriptive in nature. It uses descriptive design, a scientific method that involves observing and describing the subject's behavior without influencing it in any way [6]. Descriptive design is often used as a precursor to quantitative design, and these experiments are tedious [7]. Descriptive analysis, descriptive analytics, or descriptive statistics, is the process of using a statistical technique to describe or summarize data. It is famous for its ability to generate accessible insights from otherwise uninterpreted data. Descriptive analysis has four categories, the measure of frequency, central tendency, dispersion or variation, and position [8]. Tracer studies go beyond providing merely descriptive information concerning graduates' performance in the job market. Stakeholders and users are curious as to what aspects of the study circumstances and provisions have an impact on the employment outcomes since the results of the tracer studies offer insights and contributions to understanding the graduates' labor market status. To prevent misinterpretation of results, context variables such as economic circumstances, regional job market, and individual mobility and motivation must be taken into account [9].

A correlational approach is also being utilized in the conduct of the tracer study. Correlation analysis is a statistical technique that measures the strength of the link between two variables. A high correlation indicates that the association between two or more variables is strong, whereas a low correlation indicates that the variables are hardly related. In other words, it is the analysis of the relationship's strength using relevant statistical data [10]. A tracer study for PNU graduates used Chi-square goodness of fit to determine the relationship between the graduates' fields of specialization and their occupations after graduation [5]. Meanwhile, a graduate tracer study for the Industrial Engineering Program utilized Pearson's Correlation to determine the relationship between learning and program assessment in the employment status and current position [11]. Further, a tracer study for the Electronics Engineering Program also used Pearson's Correlation to determine the relationship between the program's assessment and the graduates' employment status [12].

## ***2.2 Graduate Tracer Information Systems***

Information systems are interdependent components that gather, analyze, store, and distribute data to support an organization's decision-making, coordination, control, analysis, and visualization processes [13]. Several studies have been conducted on the development of tracer study information systems. Each of these was conceptualized and developed to help their respective university track their alum activities

after graduation. The first study focused on developing a Tracer Study Information System using the Waterfall Model. The application program may gather alum data and generate alum data reports for university accreditation on QS Stars and BAN-PT [14]. Another study is about the development of the Codeigniter-Based Tracer Study Application using procedural model development research design from Brog and Gall. The study was conducted to aid the collection of data of alumni of Faculty Education’s Student and to fulfill various accreditation and administrative needs about data of alumni [15]. The third study used the Iterative model in the system development life cycle for developing the system. The system empowers the user to quickly evaluate alumni performance and access job posting career opportunities on the net. Hypertext Markup Language (HTML) and PHP: Hypertext Processor (PHP) was applied to code the machine instruction. The Android Studio was utilized for mobile development; the Eclipse application was used to tailor the system environment, and the Bluestacks application to connect other platforms [16].

### 3 Methodology

The study used a rapid application development software model. It consists of four phases: requirements planning, user design, construction, and cutover (Fig. 1). The RAD model is suitable for the study due to its key benefit of fast project turnaround. This approach enables project managers and stakeholders to monitor progress correctly and communicate in real-time about emerging difficulties or modifications. This leads to increased productivity, accelerated development, and efficient communication. [17].

#### 3.1 Requirement Planning

In this phase, a participatory interview with the client was conducted to determine the project’s goals and expectations. A review of related literature and studies was also undertaken. The system’s functional requirements were identified through the

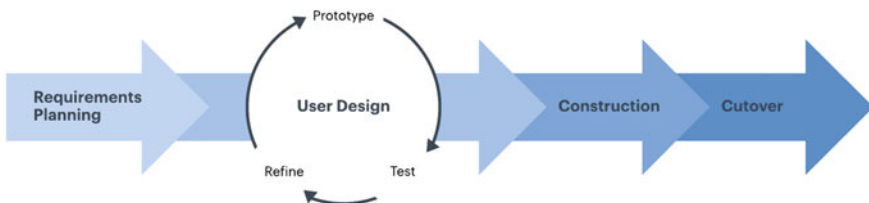


Fig. 1 Rapid application development software model

current problems encountered in conducting tracer studies and the concepts learned in internet and library research.

### **3.2 User Design**

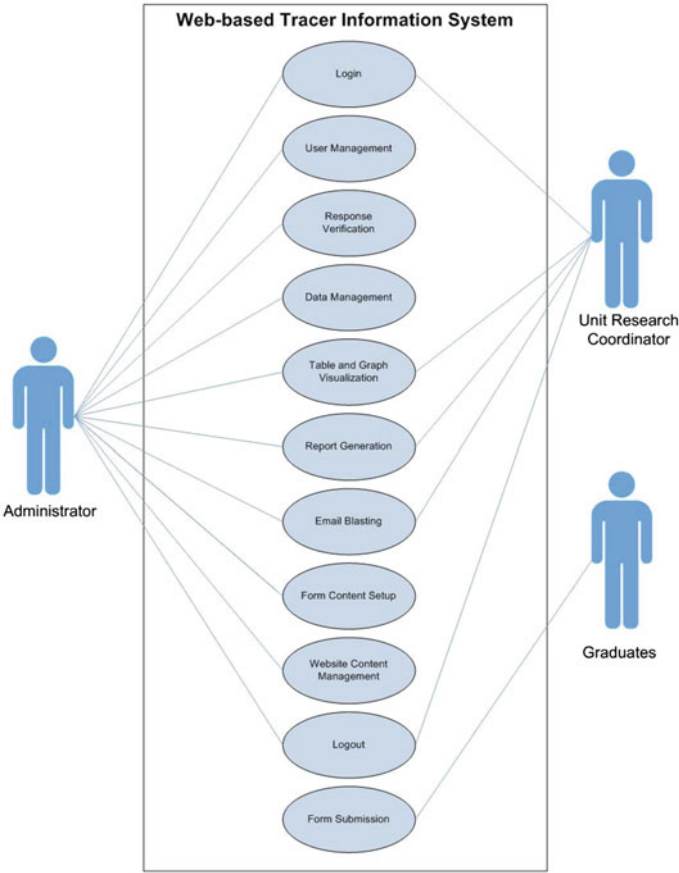
Upon identifying the project scope, each component of the web-based system capable of expediting the process of acquiring, processing, storing, and presenting graduate information was designed. This activity includes creating a Use Case Diagram, Entity Relationship Diagram, and wireframe of the system. The initial design was presented to the potential users, and feedback was gathered and used for further refinement. The design revisions were all worked out in an iterative process until reaching a satisfactory design. The use case diagrams describe a set of actions or use cases that can be performed by one or more external users of the systems. It is used in system analysis to identify, clarify, and organize system requirements [18]. Figure 2 shows the use case diagram of the system.

The web-based system can be accessed according to the account's level. There are three accounts, the administrator, research coordinators, and the graduates. The administrator has full access to the system features, including managing accounts, graduate data, and website content. The research coordinator's account is limited in viewing and generating the data of the graduates under each program. Graduates can only access the graduate tracer form. The Entity Relationship Diagram (ERD) of the system's database was also created. ERD depicts the system's database's logical structure and presents the relationship among its entity sets [19]. Figure 3 shows the Entity Relationship Diagram for the graduate's data.

The study used a relational database to avoid data anomalies common in other types of databases. The database design was realized through creating (8) tables namely: (1) *general\_information*; (2) *educational\_background*; (3) *professional\_examination*; (4) *advance\_trainings*; (5) *employment\_data*; (6) *reason\_for\_staying*; (7) *relevant\_skills* and (8) *cvsu\_feedback*. These tables are interlinked to ensure easy retrieval of the data needed.

### **3.3 Construction and Cutover**

In this phase, the researcher codes the system and ensures that the developed system adheres to the design formulated in the previous stage. HTML, CSS, and JavaScript scripting languages were used to create the website interface, while PHP and MySQL were used to develop their functions and modules. Once a module is done, it is subjected to design and validation checks. If passed, the development will proceed to the next module. Otherwise, further refinement will be undertaken until passed. This process continues until the completion of the system. The system was tested



**Fig. 2** Use case diagram of web-based graduate tracer information system

through alpha testing, where bugs and errors were fixed. This was done to ensure that the final system met the expected requirements without errors.

**4 Results and Discussion**

**4.1 Description of the System**

The system is a web-based application that collects alums’ data through its digital graduate tracer survey form (Fig. 4). The system has an analytic dashboard allowing users to view the total number of respondents, employment rate, top companies

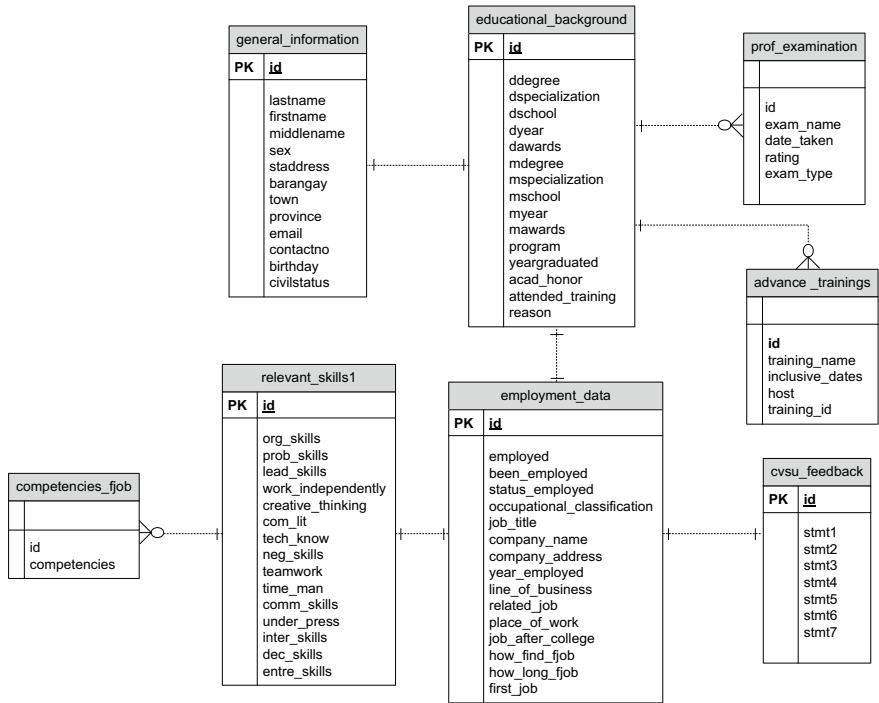


Fig. 3 Entity relationship diagram

hiring graduates, and skills orientation (Fig. 5). Furthermore, it provides descriptive analytics of verified graduates' data using frequency count, percentage, standard deviation, and coefficient of variation. Demographics of graduates, educational background, employment data, and first job information are illustrated using a frequency distribution table and graphs (Fig. 6). Additionally, the system presents correlation analytics that describe the relationship between the school-acquired skills of the graduates and (1) job search duration and (2) initial monthly gross income.

4.2 Test Results

A series of testing was done in each module to ensure that the system is free from defects and yields an accurate result. Table 1 presents the test results of the various test cases. The data acquisition function of the system proved to be accurate as it ensures that all required fields in the digital tracer form are filled out, the data being submitted is in the correct format, and they are being saved correctly in the database without data loss. The descriptive analysis page of the system proved to be 100% accurate in presenting and calculating the frequency count, mean, and standard deviation as

The screenshot displays the 'CvSU WEB - TRACER' survey interface. It features a list of skills to be rated on a scale of 1 to 5. The skills listed are: Problem solving skills, Leadership Skills, Ability to work independently, Creative Thinking, Computer Literacy, Technical Knowledge, Negotiation Skills, and Teamwork / Team Orientation. Each skill has a corresponding row with five radio buttons for rating. In the 'Creative Thinking' row, the rating '4' is selected. The interface is clean with a dark header and light body.

Fig. 4 Digital graduate tracer survey form

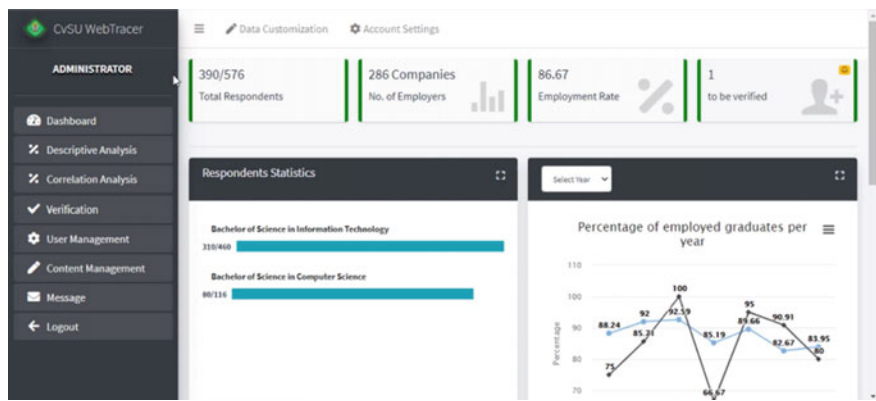


Fig. 5 Dashboard with analytics

it matched the expected value generated by Statistical Package for Social Sciences (SPSS). Furthermore, the correlation analysis page of the system also proved to be 100% accurate in presenting and calculating the rho coefficients, which describe the relationship between the school-acquired skills and (1) Job search duration and (2) Initial gross income. Moreover, the system generated an accurate verbal interpretation for descriptive and correlation analyzes.



Fig. 6 Frequency distribution table and graphs of graduate's data

Table 1 Test case results

Functionality	TC ID	Status	Detail
<i>Frontend</i>			
Digital graduate tracer survey form input validation	TC-001	Passed	Accomplished GTS form with complete and correctly formatted data entry
Submission of graduate tracer form	TC-002	Passed	Graduate information is saved correctly in the database without any data loss
<i>Backend</i>			
Adding of new program and year level	TC-003	Passed	New program and year levels were added to the database and digital graduate tracer form's option
Response verification	TC-004	Passed	Automatic verification of respondent's authenticity and checking of duplicate entries
A glance view of graduate information	TC-005	Passed	The graduate's information was displayed on the dashboard
Viewing of the descriptive and correlation analytics of the graduate's information	TC-006	Passed	The descriptive and correlation analytics generated by the system are accurate as they match the results generated by the SPSS
Report generation	TC-007	Passed	Various reports in the form of tables and charts/graphs are generated and accurate



## 5 Conclusions and Recommendations

After employing the principle of software engineering in software development and performing a series of tests, it can be concluded that the university and its alums can already use the web-based Graduate Tracer Information System. The web application can be utilized to streamline the process of data collection and processing. It can also display and manage alum data and generate reports that can be used for accreditation and other purposes. The developed system must be integrated into the existing web portal of Cavite State University–CCAT. The utilization of the system will prove its effectiveness in aiding the cumbersome process of graduate tracing in the university. Moreover, it is further recommended to make the digital tracer form customizable. With this, the graduate tracer instrument can be modified tailored-fit to a specific study program. It can also be customized with a focus on the 21st-century skills of the graduates, which are emerging considerations when hiring graduates. Furthermore, regression analysis for predicting the employment rates of graduates can be an added feature of the system.

## References

- Schomburg H (2003) Handbook for graduate tracer studies 2003. Incher Kassel, Germany
- Commission on Higher Education (CHED) (2010) Memorandum Memo No. 27 Series 2000 “Issuance of the Implementing Guidelines on the integration of CHED- Supervised Institutions (CSIs) to State Universities and Colleges (SUCs), Phase II” Author F.: Contribution title. In: 9th international proceedings on proceedings, pp 1–2. Publisher, Location
- Tumapon T (2016) Graduate tracer studies. <https://www.manilatimes.net/2016/09/09/opinion/columnists/graduate-tracer%20studies/284763/>
- Schomburg H (2016) Tracer studies at a glance. Carrying out tracer studies: guide to anticipating and matching skills and jobs (pp 25). European Union
- Gines AC (2014) Tracer study of PNU graduates. *Am Int J Contemp Res* 4(3):81–98
- Shuttleworth (2008) Descriptive research design. Retrieved from <https://explorable.com/descriptive-research-design>
- Aguila (2016) Employability of computer engineering graduates from 2013 to 2015 in one private higher institution in the Philippines. *Asia Pacific J Educ Arts Sci* 3:48–54
- Bush T (2020) Descriptive analysis: how-to, types, examples. [https://pestleanalysis.com/descriptiveanalysis/#What\\_Is\\_Descriptive\\_Analysis](https://pestleanalysis.com/descriptiveanalysis/#What_Is_Descriptive_Analysis)
- Schomburg H (2016) Tracer studies in a glance. Carrying out tracer studies: guide to anticipating and matching skills and jobs (pp 16). European Union
- Franzese M, IulianoA (2018) Correlation analysis. *encyclopedia of bioinformatics and computational biology: ABC of bioinformatics* 1-3:706–721
- Curbano RJP, Bustamante RV (2018) Graduate tracer study for industrial engineering program from batch 2013 to 2015. In: Tang S, Cheah S (eds) *Redesigning learning for greater social impact*. Springer, Singapore. [https://doi.org/10.1007/978-981-10-4223-2\\_28](https://doi.org/10.1007/978-981-10-4223-2_28)
- Balba N (2019) Graduate tracer study for electronics engineering program (2016–2018). *LPU—Laguna J Multi Res* 3:82–86
- Bourgeois DT, Smith JL, Wang S, Mortati J (2019) *Information Systems for Business and Beyond*. Open Textbooks. 1. <https://digitalcommons.biola.edu/open-textbooks/1>, pp 3
- Hafiz M et al (2020) Development of Tarumanagara university tracer study information system. *IOP Conf Series Mater Sci Eng* 1007:012117

15. Mohammad SH et al (2020) Development of Codeigniter-Based Tracer Study Application. *Adv Eng Res* vol 196, pp 304–309
16. Ponte AP, Dacalos KCG, Lina PR (2019) *Innovatus*, pp 125–130
17. Lucidchart (2022) Four phases of rapid application development methodology. Retrieved from <https://www.lucidchart.com/blog/rapid-application-development-methodology>
18. Fakhrouddinov (2014) UML use case diagrams. Retrieved from <https://www.uml-diagrams.org/use-case-diagrams.html>
19. Freeman (2021) What is an entity relationship diagram (ERD). Retrieved from <https://www.edrawsoft.com/what-is-entity-relationship-diagram-erd.htmls>

# Distance Education Opportunities for the Elderly in Thailand: Opportunity to Access Distance Education and Factors Affecting Such Opportunity



Phisit Nadprasert, Chanoknart Boonwatthanakul, Supanita Sudsaward, Duangbhorn Sapphayalak, and Likkhasit Putkhiao

**Abstract** The aim of this study was to investigate the present situation of lifelong learning of the elderly as well as factors impacting the opportunity and accessibility of distance education for the elderly. This study was divided into two stages. This article presents only the results of the first stage, which is a quantitative study that used a printed survey and an online survey to collect data from 791 participants. In this survey, the questionnaires were distributed to samples from all five geographical regions in Thailand. It was found that nearly all the elders were married and lived with a small family. Their education varied, but mostly they had completed elementary school education, and most did not work anymore. Their average income was most commonly between 1000 and 5000 Thai Baht that came from elderly welfare given by the government. They mostly did have electronic devices, that included televisions and smartphones, and mostly they learned for self-development and used their devices to get access to social media such as Line and YouTube, and tended to spend less than 3 h in the morning for their learning. They reported learning mostly for easing their loneliness and maintaining their ability, perception, and memory. The people who influenced their decisions to learn were mainly just themselves. Moreover, they indicated having little knowledge about the topics of law in daily life and elderly

---

P. Nadprasert (✉) · S. Sudsaward · D. Sapphayalak  
Office of Educational Technology of Sukhothai, Thammathirat Open University, Pakkret,  
Thailand

e-mail: [Phisit.nad@stou.ac.th](mailto:Phisit.nad@stou.ac.th)

S. Sudsaward

e-mail: [Supanita.sud@stou.ac.th](mailto:Supanita.sud@stou.ac.th)

D. Sapphayalak

e-mail: [Duangbhorn.sap@stou.ac.th](mailto:Duangbhorn.sap@stou.ac.th)

C. Boonwatthanakul · L. Putkhiao

School of Educational Studies of Sukhothai, Thammathirat Open University, Pakkret, Thailand

e-mail: [Chanoknart.boo@stou.ac.th](mailto:Chanoknart.boo@stou.ac.th)

L. Putkhiao

e-mail: [Likkhasit.put@stou.ac.th](mailto:Likkhasit.put@stou.ac.th)

people's rights. They were mostly interested in learning more about physical exercise, the use of Line application, music, and rights protection for the elderly. Most said they were likely to use the opportunity to participate in distance learning at only a medium level because of the lack of accessibility, internet network, health problems, vision problems, insufficiency of study time due to agriculture job, lack of advisors, and poor-quality devices.

**Keywords** Distance education · Elderly · Opportunities · Access to distance education

## 1 Introduction

Thailand is taking full advantage of the “aging society” demographic, with a growing population of the elderly. Educating the elderly can develop their existence as well as improving their quality of life. One of the most important attributes of a strong society is that everyone is supported so that they can maintain dignity, value, and quality of living throughout their life. Education is a regular flow of necessary relevant information and promotes lifelong learning. It can create benefits in the life of the elderly without discrimination by age. With the added capabilities of distance education technology, the elderly can now access learning resources anywhere and anytime. This research aims to study the present conditions of distance education for the elderly in Thailand. The sample population studied was 791 elderly people living in Thailand, obtained through stratified random sampling technique. The samples are selected from all five geographical regions of the country, specifically from provinces with a high number of elderly people and provinces with schools or networks for the elderly.

## 2 Methodology

The research process was divided into two phases, each of which used both quantitative and qualitative research methods. Phase 1 was intended to evaluate the present opportunities to access distant education among the elderly and to analyze the factors affecting such opportunities. Results from Phase 1 are presented in this paper.

### *Populations and Samples*

1. The study population is the elderly in Thailand. Based on data from the 2021 survey of The Older Persons in Thailand, National Statistical Office, Ministry of Digital Economy and Society. The total number of elderly persons aged 60 years or above in Thailand is 13,358,751 classifying into 5,974,022 men and 7,384,729 women.

2. The sample population was 791 elderly people, obtained through stratified random sampling technique. The samples were selected from all 5 regions of Thailand, specifically from provinces with a high number of elderly and provinces with schools or networks for the elderly: Bangkok and vicinity provinces (central region), Chiang-Mai Province (northern region), Khon-Kaen Province (northeast region), Songkhla Province (southern region) and Chonburi Province (eastern region).
3. The research tool was a questionnaire asking about elderly people's level of demand for accessing distant education. Data were collected from elderly people at least 60 years old. The questionnaire, vetted by three experts, had questions examining conditions, factors, and opportunities for the access to distance education by the elderly.

### 3 Data Collection and Findings

The printed and online questionnaires were sent to the selected samples. The online questionnaire was administered to the samples who could answer the questionnaire online and the printed questionnaire was sent to the samples who could not answer the questionnaire online, as follows.

Table 1 shows that most of samples were female ( $n = 542$  or 68.52%), the rest male ( $n = 249$  or 31.48%), whose age ranges are mainly 61–65 years ( $n = 241$  or 30.47%), followed by the age ranges of 66–69 years ( $n = 194$  or 24.52%) and 70–75 years ( $n = 156$  or 19.72%).

Data shows that most of the elderly samples ( $n = 383$  or 48.442%) were married, followed by divorced/widowed ( $n = 269$  or 34.01%), and single ( $n = 100$  or 12.64%), respectively.

The majority of the samples ( $n = 293$  or 37.04%) were living the central region of Thailand, followed by those living in the eastern region ( $n = 145$  or 18.33%) and those in the southern region ( $n = 144$  or 18.20%).

Most of the samples ( $n = 419$  or 52.97%) were living in Mueang District/Municipalities, which are more urban areas, whilst the rest ( $n = 372$  or 47.03%) lived outside Mueang District/Municipalities. As for the samples' families, most samples ( $n = 491$  or 50.72%) were living in a small family (with 2–4 people staying together), followed by the samples living in a big family with 5 or more people staying together ( $n = 195$  or 24.65%), and those staying in care centers or elderly housing ( $n = 84$  or 10.62%), respectively.

Table 2 shows that most of the samples ( $n = 293$  or 37.42%) have graduated from Grade 4 of primary school, followed by groups of samples not studying and having graduated from Grade 6 of primary school ( $n = 98$  or 12.39% per group) and the group of samples with bachelor's degrees ( $n = 96$  or 12.14%). Most of the samples reported they no longer worked ( $n = 454$  or 57.32%) whilst the other 329 samples (41.54%) who were still working were divided into 133 samples (39.23%) having farming/agricultural work,  $n = 111$ (32.74%) doing businesses/trading, and

**Table 1** Demographic particulars of the elders

Participants' personal particulars		Number	%age
1. Gender	Male	249	31.48
	Female	542	68.52
	Total	791	100.00
2. Age range	61–65 years	241	30.47
	66–69 years	194	24.52
	70–75 years	156	19.72
	76–79 years	101	12.77
	80–85 years	81	10.24
	86 years and over	18	2.28
	Total	791	100.00
3. Marital status	Single	100	12.64
	Married	383	48.42
	Divorced/widowed	269	34.01
	Separated	39	4.93
	Total	791	100.00
4. Region of residence	The north	68	8.60
	The center	293	37.04
	The northeast	145	18.33
	The south	144	18.20
	Total	791	100.00
5. Area of residence	Mueang district/municipality	419	52.97
	Outside Mueang district/municipality	372	47.03
	Total	791	100.00
6. Current living situation	Living alone	71	8.98
	Living in a small family (With 2–4 people)	401	50.70
	Living in a big family (With 5 people or more)	195	24.65
	Living in a big family (With 5 people or more)	38	4.80
	Other	84	10.62
	No specification	2	0.25
	Total	791	100.00

66 samples (19.47%) who were hired labors. From all the samples, most ( $n = 412$  or 52.09%) gained their income from government assistance for the elderly, followed by the group that received money from their offspring/other members of their families ( $n = 289$  or 36.54%) and the group that earned money from work ( $n = 179$  or 22.63%). Most samples (260 samples or 36.98%) earned only 1001–5000 baht per month, followed by the groups of those earning 5001–10,000 baht per month ( $n = 156$  or 22.19%) and those earning less than 1000 baht a month ( $n = 92$  or 13.09%). Most samples ( $n = 344$  or 43.49%) said they have enough income to cover expenses, but not to save, followed by the groups of those with enough income and to save ( $n = 234$  or 29.58%) and those with not enough income to cover expenses, but having no debt incurred ( $n = 80$  or 10.12%). Most samples ( $n = 644$  or 81.42%) reported that their major monthly expenses were their personal expenses whilst  $n = 344$  (42.23%) had to pay for telephone service and  $n = 314$  (39.70%) paid for water and electricity.

As for the technological devices that most of the samples used daily, the most common was normal TV sets, which were used by  $n = 409$  (or 51.71%), followed by smartphones, which were used by  $n = 367$  (or 46.40%) and smart TV sets, which were used by  $n = 204$  (or 25.79%). Most of the samples (290 or 36.66%) said they were regularly participating members of a group, club or elder's school (8–12 times a year), whilst  $n = 288$  (or 36.41%) were not members of a group, club or elder's school and the others  $n = 288$  (or 36.41%) were members of a group, club or elder's school but only participated less than 8 times a year.

Table 3 shows that most samples ( $n = 607$  or 76.74%) are interested in finding knowledge by themselves or accessing distance education, whilst the other  $n = 168$  (21.24%) are not interested. Most samples ( $n = 427$  or 59.98%) have devices to access education in their daily life whilst the other 240 samples (30.34%) do not. The most used devices are smartphones (used by  $n = 415$  or 68.37%), followed by tablets (used by  $n = 81$  or 13.34%) and smart TV sets (used by  $n = 78$  or 12.85%).

The most used form of online social media is Line application (used by  $n = 338$  or 55.68%), followed by YouTube (used by  $n = 253$  or 41.68%) and Facebook (used by 241 samples or 39.70%). Most of the samples spend late morning (09:00–12:00 h.) on online social media ( $n = 150$  or 24.71%), followed by those spending morning (06:01–09:00 h.) ( $n = 73$  or 12.03%) and midday (12:01–15:00 h.) ( $n = 61$  or 10.05%), respectively. Most samples ( $n = 222$  or 36.57%) spend less than 3 h a day on online social media, followed by groups of samples spending 3 h a day ( $n = 96$  or 15.82%) and 3–6 h a day ( $n = 82$  or 13.51%). From all samples, 233 (38.39%) used home internet, followed by groups of samples using 4G mobile internet ( $n = 139$  or 22.90%) and internet from care centers ( $n = 51$  or 26.69%).

The reason for most samples to learn is to spend free time ( $n = 361$  or 59.47%), followed by to maintain the ability to think and remember ( $n = 167$  or 27.51%) and to keep up their health and strength ( $n = 155$  or 25.54%). People with influence on the decision to learn cited by most samples is themselves ( $n = 383$  or 48.42%), followed by children, grandchildren and relatives ( $n = 81$  or 10.24%) and officials at care centers ( $n = 53$  or 6.70%).

Table 4 presents the online sources that most samples used to search for educational information and it was found that YouTube is the largest commonly used tool

**Table 2** Demographic particulars of the elderly (continued)

Participants personal particulars		Number	%age
7. Educational background	No formal education	98	12.39
	Primary school grade 4	296	37.42
	Primary school grade 6	98	12.39
	Secondary school	52	6.57
	High school	67	8.47
	Bachelor's degree	96	12.14
	Master's degree	25	3.16
	Doctoral degree	9	1.14
	Vocational diploma	24	3.03
	Higher vocational diploma	14	1.77
	Others such as grade 2 and grade 7	6	0.76
	No specification	6	0.76
	Total	791	100.00
8. Occupation	Working/employed	329	41.54
	Not working/ not employed	454	57.32
	No specification	9	1.14
	Total	791	100.00
9. Sources of income* (can choose more than 1 choice)	Offspring/family members	289	36.54
	Retirement fund	112	14.16
	Work	179	22.63
	Government assistance for the elderly	412	52.09
	No Income	77	9.73
	Others such as donations and remunerations from assistance to centers	38	4.80
	Offspring/family members	289	36.54
	Total	1107	100.00
10. Monthly income	No more than 1000 Baht	92	13.09
	1001–5000 Baht	260	36.98
	5001–10,000 Baht	156	22.19
	10,001–15,000 Baht	61	8.68
	15,001–20,000 Baht	46	6.54
	20,001–25,000 Baht	21	2.99
	25,001–30,000 Baht	20	2.84
	Higher than 30,000 Baht	47	6.69
	Total	703	100.00
11. Sufficiency of income	Enough to cover expenses, but none left over to save	344	43.49

(continued)



**Table 2** (continued)

Participants personal particulars		Number	%age
	Enough to cover expenses and some to save	234	29.58
	Not enough to cover expenses, but no debt incurred	80	10.12
	Not enough to cover expenses, with debts incurred	52	6.57
	Others such as unstable income	10	1.26
	No specification	71	8.98
	Total	791	100.00
12. Monthly expenses (can choose more than 1 choice)	Personal expense	644	81.42
	Water-electricity	314	39.70
	Telephone	334	42.23
	Traveling	155	19.60
	Loan payment	81	10.24
	Medical fee	232	29.33
	Savings	68	8.60
	Insurance premiums	80	10.11
	Donation/merit making	158	19.97
	Hobbies	129	16.31
	Leisure tours	92	11.63
	Educational fees/courses	47	5.94
	Others such as house repairs and decoration and social expenses	32	4.05
	Total	791	100.00
13. Technological devices for daily use (can choose more than 1 choice)	Normal TV	409	51.71
	Smart TV	204	25.79
	Radio	181	22.88
	Tablet	38	4.80
	Smartphone	367	46.40
	Desktop/laptop computer	55	6.95
	Electronic reading gadget (Kindle, Boox)	3	0.38
	Others such as normal telephone and iPad	33	4.17
	Total	791	100.00
14. Group activities	Regularly participating member of a group, club or elder's school (8–12 times a year)	290	36.66
	Participating member of a group, club or elder's school (less than 8 times a year)	181	22.88

(continued)

**Table 2** (continued)

Participants personal particulars		Number	%age
	Not a member of a group, club or elder's school	288	36.41
	No specification	32	4.05
	Total	791	100.00

( $n = 236$  or 53.88%), followed by Google ( $n = 206$  or 47.03%) and Facebook ( $n = 166$  or 37.90%); while the most convenient place for online learning for most samples ( $n = 227$  or 51.83%) was home, followed by the elder's school ( $n = 89$  or 20.32%) and the elder's club ( $n = 76$  or 17.35%).

Figure 1 shows that the elderly samples surveyed already had self-assessed levels of knowledge about the topics of aging, health and disease prevention on average at level 6.90 on a scale of 0–10; they had knowledge about the topics of how to live a good life, targets of life and quality of life, and exercise for the elderly on average at level 7.00; nutrition for the elderly, guidelines for taking medications, and benefits of herbs and vegetables on average at level 7.10; mental development for the elderly on average at level 6.90; changes in modern society and self-development skills for the elderly on average at level 6.70; information about Thai elderly people on average at level 6.30; and everyday laws and elders' rights on average at level 6.40.

The health-related issues that the elderly were most interested in learning about were found to be: exercise and physical therapy ( $n = 733$  or 92.67%), followed by health for the elderly ( $n = 311$  or 39.32%) and brain and memory training ( $n = 236$  or 29.84%).

The technology topics that most samples were interested in learning more about were Line application ( $n = 301$  or 38.05%), followed by YouTube ( $n = 279$  or 35.27%), and Facebook ( $n = 255$  or 32.24%).

The leisure activities that most elderly were interested in learning more about were: music ( $n = 303$  or 38.31%), followed by cooking ( $n = 273$  or 34.51%), and handicrafts ( $n = 160$  or 20.23%).

The topics of economics and law study that most elderly samples were interested in learning about were the protection for the elder's rights ( $n = 464$  or 58.66%), followed by health insurance and life insurance ( $n = 225$  or 28.45%) and laws involving the elderly ( $n = 215$  or 27.18%).

Data shows that the elderly samples surveyed; Opportunity to access distance education Most samples ( $n = 262$  or 33.12%) said they had moderate opportunities to access distant learning, followed by groups of samples with no opportunity ( $n = 121$  or 15.30%) and few opportunities ( $n = 120$  and 15.17%).

**Table 3** Results from the study on the access to distance education of the elderly

Access to distance education of the elderly		Number	%age
1. Interestedness in self learning	Yes	607	76.74
	No because of old age, lack of time, inconvenience etc	168	21.24
	No specification	16	2.02
	Total	791	100.00
2. Equipment daily used to access learning	No equipment	240	30.34
	With equipment	427	59.98
	No specification	124	15.68
	Total	791	100.00
3. Equipment (from item) used (more than 1 choice can be chosen.) ( $n = 607$ )	Smart TV	78	12.85
	Smartphone	415	68.37
	Tablet	81	13.34
	Desktop/laptop computer	45	7.41
	Electronic reading device (Kindle, Boox)	3	0.49
	Total	622	100.00
4. Social media used (more than 1 choice can be chosen.) ( $n = 607$ )	Line	338	55.68
	Facebook	241	39.70
	Twitter	22	3.62
	YouTube	253	41.68
	Podcast	6	0.99
	Instagram	32	5.27
	TikTok	78	12.85
	Others such as Blockdit and computer	24	3.95
	Total	622	100.00
5. Time of day for social media use ( $n = 607$ )	Early morning (03:00–06:00 h)	58	9.56
	Morning (06:01–09:00 h)	73	12.03
	Late Morning (09:01–12:00 h)	150	24.71
	Midday (12:01–15:00 h)	61	10.05
	Afternoon (15:01–18:00 h)	40	6.59
	Evening (18:01–21:00 h)	53	8.73
	Nighttime (21:01–24:00 h)	27	4.45
	No specification	145	23.88
	Total	607	100.00
6. Average time spent on social media each day ( $n = 607$ )	Less than 3 h	222	36.57
	3 h	96	15.82

(continued)

**Table 3** (continued)

Access to distance education of the elderly		Number	%age
	3–6 h	82	13.51
	More than 6 h	20	3.29
	No specification	187	30.81
	Total	607	100.00
7. Mean of access to internet ( $n = 607$ )	Home internet	233	38.39
	Phone internet—4G	139	22.90
	Wi-Fi in local community	4	0.66
	Wi-Fi at the elder's school	18	2.97
	Others such as care center	51	8.40
	No specification	162	26.69
	Total	607	100.00
8. Reasons for learning (More than 1 choice can be chosen)	To spend free time	361	59.47
	To spend time with child or family member	106	17.46
	To build network	70	11.53
	To learn a way to earn money	58	9.56
	To fulfill curiosity/desire to learn	132	21.75
	To maintain ability to think and remember	167	27.51
	To keep up health and strength	155	25.54
	Others such as to have self-development and to listen to news	14	2.31
	Total	1063	
9. The most influential person on decision to access learning	Myself	383	48.42
	Spouse	25	3.16
	Children, grandchildren, relative		81
	Friends	26	3.29
	Community leader	13	1.64
	Others such as officials at care center	53	6.70
	No specification	210	26.55
	Total	791	100.00

## 4 Discussion

Elderly individuals differ in terms of their interests, health and living background. Most of them have some desire to learn sparked from their own interest and are willing to learn about something they are interested in, or they want to learn topics

**Table 4** Results from the study on the access to distance education of the elderly (continued)

Access to distance education of the elderly		Number	%age
10. Knowledge of online learning sites for the elder	No, but interested in	298	37.67
	No and not interested in knowing	80	10.12
	Yes	140	17.70
	No specification	273	34.51
	Total	791	100.00
11. Sources of information about educational opportunities (more than 1 choice can be chosen)	Google	206	47.03
	YouTube	236	53.88
	Facebook	166	37.90
	TikTok	47	10.73
	Podcast	7	1.60
	Others such as line and TV	38	8.68
	Total	655	
12. The most convenient place to study online (anytime, anywhere) (more than 1 choice can be chosen) ( $n = 438$ )	Home	227	51.83
	Community center/ temple/religious center	25	5.71
	The elder's club	76	17.35
	The elder's school	89	20.32
	Educational institutes (all levels)	12	2.74
	Others such as care centers	72	16.44
	Total	417	

that can complement their daily life, or about something they have missed before [1]. The main problems facing the elderly that can cause limitations in their educational opportunities are health concerns and their physical decay. They have ability but their capabilities vary. Educators must develop the ability of elderly learners from what they already possess and what is relevant to their age, lifespan, occupation, income, skills and knowledge, education, prior knowledge, society, culture, attitude, behavior, and motivation, in order to assist them to learn properly and effectively. In previous studies on interest in design and choice of learning materials, there should be analysis, media design, media development, evaluation of the implementation to suit the target group [2], according to the elderly, they tend to spend their free time on recreation [3], self-development, culture-religion, self-value enforcement, self-finance management and additional occupation, for the sake of their personal interest. Moreover, they want to learn primarily for adjustment and for daily life, thus

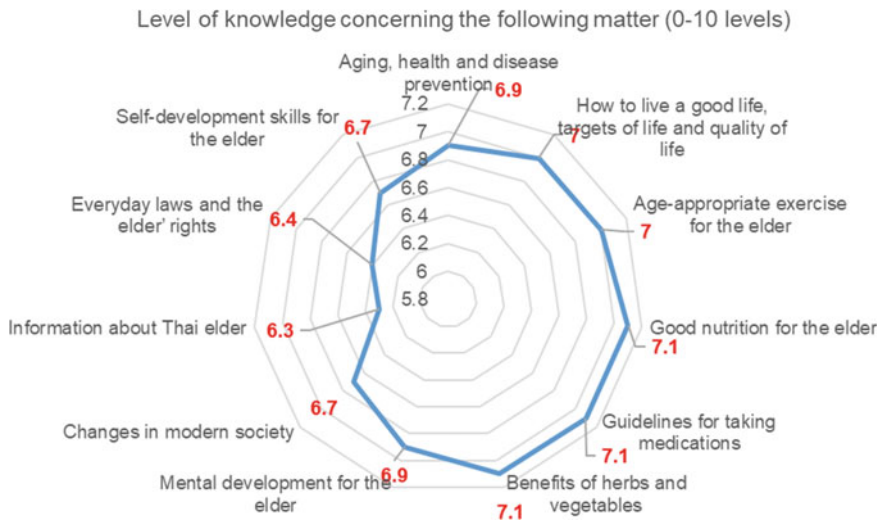
**Table 5** Opportunity to access distance education ( $n = 791$ )

Access to distant education of the elderly		Number	%age
Opportunity to access distance education	None because of the lack of free time, old age, no technological device for communication, headache	121	15.30
	Few because of the disability to use technologies, lack of money to spend on technologies and no chance to use	120	15.17
	Rather few because of old age, bad eyesight and forgetfulness after being taught by offspring	106	13.40
	Moderate because of the disability to access to networks, problems with eyesight and health, lack of time for learning because of agricultural work, the care-takers, and outdated devices	262	33.12
	Rather many because of the promptness of communication devices, the necessity, the availability of free time, chances to have online learning, easy accessibility and ability to learn anytime	84	10.62
	Many because of the availability of time and devices to access, the promptness to use technologies, the ability to learn anytime, easy accessibility and easier life made by technologies	49	6.19
	Plenty because of the preference to learn about unknown things, fundamental promptness, installation of fiber optics at home, availability of 5G phones and home internet, and the contemporary fun time for the elder	27	3.42
	No specification	22	2.78
Total		791	100.00

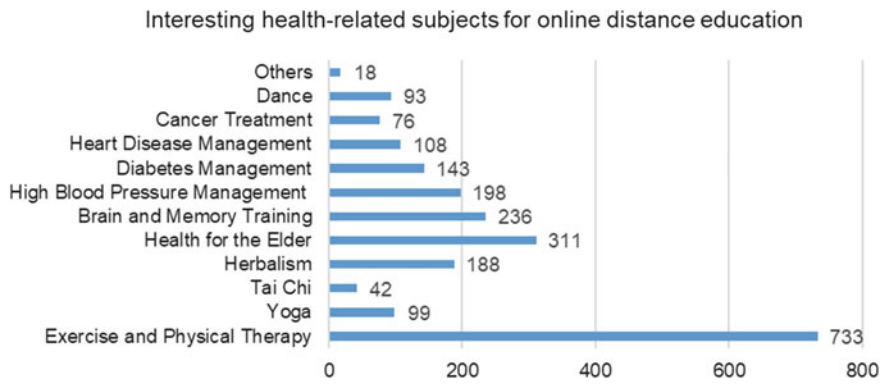
this is the essential learning of the elderly. That should be supported by education, and especially distance education.

The learning of the elderly is different from other people from different age groups. Considering the characteristics of the elderly, there are some physical and mental changes in various aspects as people age. Educators have studied how successful aging is perceived. Hence, the attributes of successful aging, according to the literature, should be looked at in the perspective of the elderly towards “successful aging” [4] in terms of physical-, mental-, social-, and economic points, including necessary abilities that elderly people need to be able to learn and enjoy living day to day. Moreover, the important things to teach the elderly include the perception of rights and duties, laws and legal issues in daily life, and media literacy for elders, as we have found that more elders in Thailand have been cheated.

A limitation of this study involved the survey data collection. In the situation of the spread of the virus COVID-19, data collection had to be done very carefully, especially because the informant group comprised elderly persons who are especially



**Fig. 1** Data level of knowledge from the self-assessment of the elderly on knowledge 0–10 level



**Fig. 2** Interest of the elderly on learning various health-related subjects ( $n = 791$ )

at risk for respiratory system viruses. The information was only obtained from those who were willing to take part in the survey.

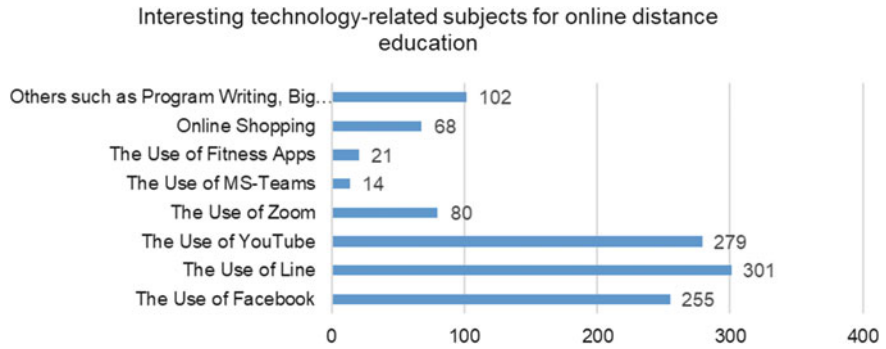


Fig. 3 Interest of the elderly in technology-related subjects ( $n = 791$ )

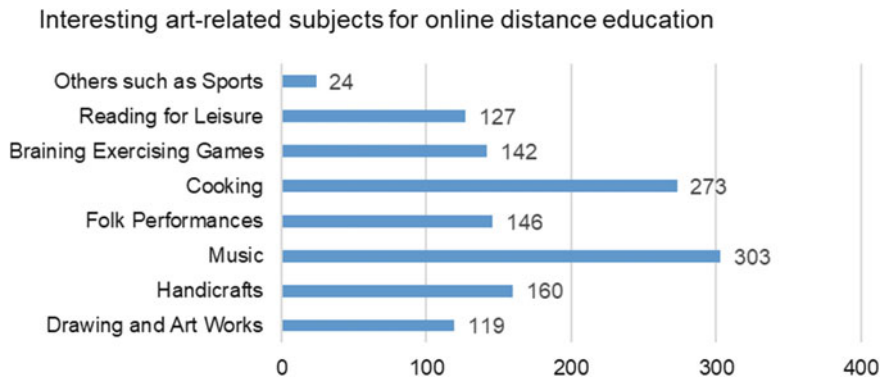


Fig. 4 Interest of the elderly in leisure activities subjects ( $n = 791$ )

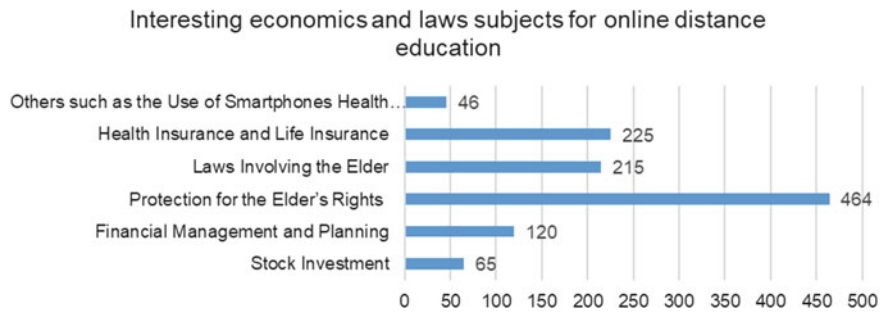


Fig. 5 Interest of the elderly in economics and laws subjects ( $n = 791$ )



## References

1. Delors J et al (1998) *Learning: the treasure within*. UNESCO
2. Branch RM (2009) *Instructional design: the ADDIE approach* (vol 722). Springer Science & Business Media
3. Novak M (2018) *Issues in aging leisure, recreation, and education*. Chapter 11, p 485
4. Phelan EA, Anderson LA, Lacroix AZ, Larson EB (2004) Older adults' views of "successful aging" how do they compare with researchers' definitions? *J American Geriatrics Soc* 52(2):211–216

# Motivation Prediction for Persuasive Intervention at Appropriate Timing to Promote Exercises



Tomoya Yuasa, Fumiko Harada, and Hiromitsu Shimakawa

**Abstract** This study aims to encourage users to take exercises based on Fogg Behavior Model (FBM), taking into account the user's activity status and motivation. FBM insists that it is effective to provide appropriate triggers that match motivation and ability of users for persuasion of them to engage in specific actions. Since motivation is a psychological factor, questionnaires have mainly been used to investigate current motivation. However, repeated questionnaires are heavy burdensome for respondents. This study challenges to construct a machine learning model that predicts motivation in a coming period from the relationship between physical activity data and motivation in the past. If a machine learning model can make appropriate predictions, it is possible to promote actions without questionnaires. To predict motivation, physical activity data obtained from a smartwatch-type action meter are used as an explanatory variable, while a questionnaire on motivation to exercise is used as an objective variable. As a result, the study has succeeded in obtaining more than 80% of correct answers.

**Keywords** Persuasive engineering · Motivation prediction · Exercise promotion

## 1 Introduction

More and more people get lack in exercise due to their working style and immerse into sedentary lifestyles year by year. It is one of the big issues in the world to keep them healthy, improving their unhealthy behavior. Numerous studies have already shown that regular physical activity has a positive impact not only on physical health

---

T. Yuasa (✉) · H. Shimakawa

Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

e-mail: [tomoya@de.is.ritsumei.ac.jp](mailto:tomoya@de.is.ritsumei.ac.jp)

H. Shimakawa

e-mail: [simakawa@cs.ritsumei.ac.jp](mailto:simakawa@cs.ritsumei.ac.jp)

F. Harada

Research Organization of Science and Technology, Ritsumeikan University, Shiga, Japan

e-mail: [harada@de.is.ritsumei.ac.jp](mailto:harada@de.is.ritsumei.ac.jp)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_52](https://doi.org/10.1007/978-981-99-3091-3_52)

629

but also on mental one [1–3]. Many people have heard at least once about the benefits of physical activity. The WHO reports, however, a quarter of the world’s adults do not meet its recommended level of physical activity [4].

In recent years, there have been many IT-based exercise promotion methods [5, 6] such as smartphone applications and smartwatch-type activity meters to promote physical activity. The tools allow users to monitor themselves by visualizing their own physical conditions. They make the users notice the lack of exercise, which is expected to promote them to take exercise. However, the improvement with the IT methods is far from a solution for the lack of exercise and the sedentary lifestyles. Apps on smartphones and smartwatches are so novel that they make users filled up with motivation, at first. However, as time goes by, the motivation wears off. The users reduce exercising even if they realize they are not getting enough exercise, and eventually they stop using the apps on smartphones and smartwatches. Simply introducing smartphones and smartwatches is not enough to raise user awareness. It is necessary to provide continuous intervention to persuade users to engage in exercise, utilizing the information obtained from them.

The study provides users with persuasive interventions, based on Fogg Behavior Model (FBM) that considers the user’s activity status and motivation. The FBM explains that triggers should be provided when motivation and ability are high [7]. Based on FBM, the motivation and ability of a particular user must be grasped on the spot to persuade the user to engage in exercise. This study especially try to know the motivation of users. A machine learning model is trained to predict their cumming motivation based on physical activity data obtained from the smartwatch-type action meter over a past period and a questionnaire about their motivation to exercise that the users answered periodically over the period. The paper discusses the effectiveness of the model.

This paper introduces the method of promoting people’s behavior in Sect. 2. Section 3 explains how to improve exercise based on FBM, taking the user’s activity status and motivation into account. Section 4, a machine learning model is constructed to predict motivation. Section 5 discusses the results of a study to demonstrate the usefulness of the proposed model. Section 6 summarizes remarks and presents future works.

## 2 Methods to Promote Action to People

### 2.1 *Behavior Promotion Using Smartwatch-Type Action Meters*

In recent years, smartwatch-type action meters have begun to spread and are a popular trend [8]. Consumer action meters, such as Fitbit and Garmin, measure various data related to user activity, such as steps, sleep, and heart rate. The data can be viewed on the action meter or in a smartphone application. The users can check their activity

anytime on these devices. The devices allow the users to monitor their physical activity to support them become more active. Some works have reported an increase in physical activity with the use of smartwatch-type action meters [8, 9].

One of the major problems with smartwatch-type action meters, however, is the difficulty of wearing them on a sustained basis. According to a survey, about one-third of U.S. adults stop using smartwatch-type action meters after six months [10]. As it implies, to achieve sustained monitoring with smartwatch-type action meters, it is necessary not only to wear smartwatch-type action meters for monitoring but also to provide users with measures to promote exercise. The measures can be realized by combining data from smartwatch-type action meters with a specific function.

## **2.2 Behavior Promotion Using FBM**

The Fogg Behavior Model (FBM) [7] has been proposed as a method to persuade users and promote user behavior. The FBM states that motivation, ability, and triggers must occur simultaneously for people to take action. At that time, motivation and ability must exceed a certain threshold. If motivation and ability do not exceed the threshold, FBM supports the user, showing the users images to arouse hope, lightening the user's burden, and so on, so that they exceed the threshold.

The idea of FBM has already been adopted in various fields [11–14]. However, in works using FBM, questionnaires are often used to estimate users' motivation and ability [12, 13]. In the case of periodic monitoring, the burden on users gets high, because it is tiresome work for users to answer questionnaires.

# **3 Exercise Improvement Considering the User's Activity Status and Motivation**

## **3.1 Methodology**

This paper proposes an exercise improvement method based on FBM that takes into account the user's activity status and motivation.

A smartwatch-type action meter acquires the user's physical activity data such as the number of steps and sleeping time. The average value of the number of steps for each day of the week is calculated from the obtained time series data of the number of steps. Smoothing with a Kalman filter is applied to the average value of the number of steps obtained for each day of the week. It allows us to obtain the user's regular activity status for each day of the week. From this activity status, we can determine the time of day when the user's ability is high.

To understand the user's motivation, the user answers a questionnaire about his/her motivation to exercise for a certain period. According to the results of the question-

naire, whether the user's motivation is high or low is labeled for each time point in the period. A machine learning model is trained to predict the user's motivation. The state of high or low motivation is used as the objective variable. As an explanatory variable, we use physical activity data obtained from a smartwatch-type action meter worn by the user.

To predict the user's motivation for a specific day, the physical activity data of the previous day is used as the explanatory variable. That is, to predict the results of the questionnaire answered on November 25 as the objective variable, the physical activity data of November 24 are used as the explanatory variable.

The study adopts a random forest to construct a machine learning model to identify the user's motivational state, using the above explanatory and objective variables. Feeding the physical activity information obtained from the user's smartwatch-type action meter as an explanatory variable to the model, the study predicts whether his/her motivation is high or low without questionnaires.

Let us assume the current ability of the user is high. If the predicted motivation is high, we should not urge the user to engage in exercise. We should just prompt the user with a reminder notification. When the motivation is low, the user should be urged to take exercise with an encouraging text message. It would be expected to improve the user's exercise.

### ***3.2 Smoothing the Number of Steps per Day of the Week***

In this study, to understand the user's activity status for each day of the week, an average value is calculated from the number of steps. In general, people's activities are often similar on the same day of the week or at the same time of the day. For example, on a specific day of the week, students attend the same class, while working people have regular meetings on the same day of the week. Therefore, people are likely to behave similarly on each day of the week. If only one specific day is extracted, behavior on the day may happen to be irregular (e.g., people catch a cold and stay in bed all day). The influence of such irregular days should be minimized as much as possible. To solve this problem, multiple identical days of the week were prepared to reduce the influence of the irregularities.

Furthermore, the smoothing by the Kalman filter can further reduce noise in the step counts averaged for each day of the week. For the time series data of the number of steps, a Kalman filter can obtain the best-estimated value of the true state at time  $t$  using data both before and after time  $t$ .

In the transition of the smoothed step count, the proposed method calculates the minimum and maximum values. The number of steps steadily increases during the period. In other words, the ability to acquire steps is higher than that in other periods. The proposed method intervenes according to the user's motivation during the period when the number of steps moves from the minimum to the maximum, especially at the time of the minimum.

### **3.3 Intervention Using Text Messages**

Text messages are an inexpensive and highly effective means of encouraging people to take action. Many previous studies have already shown the potential effectiveness of text messages in various fields [15, 16].

Messages created with the user's motivation neglected are considered to be ineffective. This study chooses text messages reflecting behavioral economics for people with low motivation. The message presents positive images of exercise to them. On the other hand, according to FBM, it is sufficient to simply send reminders to those with high motivation. The proposed method sends messages reminding them of exercise.

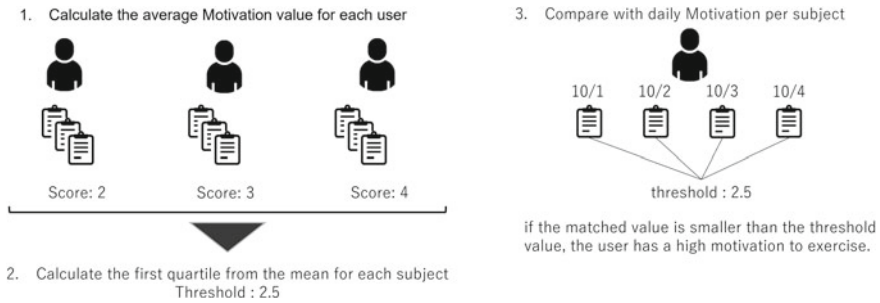
## **4 Building a Machine Learning Model to Predict Motivation**

### **4.1 Outline of the Survey**

The effectiveness is investigated for a machine learning model to predict motivation. The subjects are 11 working adults (5 males and 6 females) in their 30s to 50s. The duration of the experiment is five weeks. Each subject wore a smartwatch-type action meter during the period. The study used Fitbit's Fitbit Charge5 [17]. Fitbit devices are widely used as a measurement tool in physical activity and health promotion research [18]. A Fitbit Charge5 is hereafter referred to as a Fitbit. Subjects were instructed to wear a Fitbit all the time except when they take a bath. The time is used to charge a Fitbit.

During the experiment, subjects were asked a question about their motivation to exercise, "On a scale of 0–10, how motivated are you to exercise?" In the same way, the subjects were asked "How motivated are you to exercise on a scale of 0 (willing to exercise) to 10 (not willing to exercise)" for the 5 weeks. Instead of two choices of whether or not the user has the desire to exercise, the questionnaire was subdivided into 11 choices on a scale of 0–10. The reason for this is to unify users' feelings toward exercise. For example, if user A is rather unwilling to exercise, he/she may answer "I don't want to exercise". On the other hand, if user B wants to take a little exercise, he may answer "I would like to exercise". Though their feelings are similar to each other, it appears as a big difference, in the case of the two choices. When the answer options are narrowed in this way, there is a large possibility that the answer will depend on the scale used by the individual, resulting in a subjective evaluation. Therefore, we unified the scale of the user's feelings toward exercise by dividing the scale into more detailed steps, such as 0–10.

Of the five weeks of data obtained, the first four weeks were used as training data for model building, while the last week was used as test data.



**Fig. 1** Calculation method of the threshold value

## 4.2 Setting the Threshold

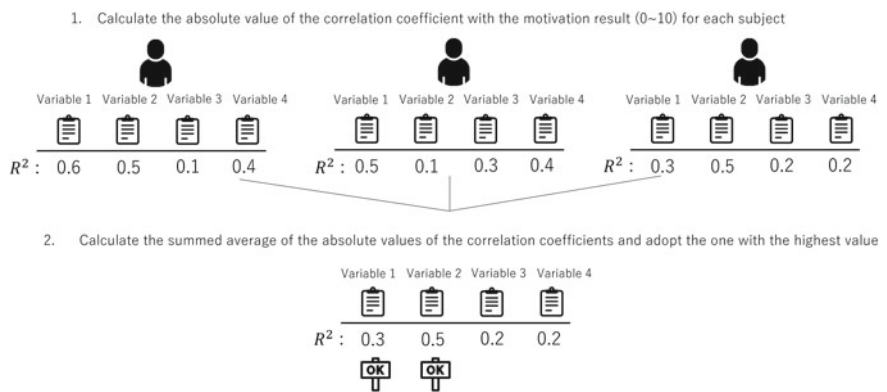
Since this study uses a rating scale from 0–10, it is necessary to know whether or not each user was relatively motivated to perform the exercise to construct the model. To this end, threshold values were set to ascertain whether users were in a state of high or low motivation, as shown in Fig. 1.

First, the average is calculated for the results of each subject’s responses to the question about their desire to exercise daily for four weeks. In this study, subjects were asked to respond to a daily motivation questionnaire for four weeks, but there were cases in which the results of all four weeks’ worth of responses were not available, or in which subjects responded multiple times on the same day. In such cases, the mean value for each subject was calculated using only the days on which responses were received. In the case of multiple responses on the same day, the one with the latest timestamp was used to adjust to one response per day.

Then, the first quartile of the total number of subjects’ answers is calculated based on the mean value of each subject’s answer. This first quartile value is set as the threshold value. This threshold value is compared again with the value of each subject’s daily intention to exercise. If the value is larger than the threshold value, the subject’s motivation to exercise is regarded as low. This state is defined as a state of low motivation to exercise. On the other hand, if the matched value is smaller than the threshold value, the subject has a high motivation to exercise. This state is defined as high motivation to exercise. The threshold was set to the first quartile because of the study by Agha and colleagues [13]. For the days with questions unanswered, the median value for each subject was interpolated. The value is compared to the threshold to label whether the subject was highly or lowly motivated.

**Table 1** List of explanatory variables obtained from wearable devices

No.	Variable name	No.	Variable name
1	Stress level	8	Exercise calories burned
2	Calorie consumption	9	Sleeping time (minutes)
3	Number of steps	10	Time awake and awake now (minutes)
4	Distance	11	Time in bed (minutes)
5	Light active time (minutes)	12	REM sleep (minutes)
6	Fairly active time (minutes)	13	Shallow sleep (minutes)
7	Very active time (minutes)	14	Deep sleep (minutes)



**Fig. 2** Selection of explanatory variables

**4.3 Preparation of Explanatory Variables**

The following physical activity data were collected from Fitbit. Table 1 shows the collected 14 items of physical activity data.

The difference from the previous day is calculated for the stress level, the number of steps, the time of being somewhat active (minutes), the time of being fairly active (minutes), and the time of being very active (minutes). These 5 values are added as candidates of explanatory variables in addition to the above 14 items. For the number of steps, a three-day moving average is also added to the candidates. In total, the candidates for model construction are 20 items , including 14 items of physical activity data that can be obtained from Fitbit and 6 items obtained by processing physical activity data.

Among the candidates, the study selects explanatory variables for the model construction, according to the correlation coefficients with the results of the exercise motivation question. The selection method is shown in Fig. 2.

First, the absolute values of the correlation coefficients between the 20 items and the results of the exercise motivation question were calculated for each subject. Next,



**Table 2** List of explanatory variables used in model building

No.	Variable name
1	Moving average
2	Very active time (minutes)
3	Number of steps
4	Distance
5	Deep sleep (minutes)
6	Stress level difference from previous day
7	Slightly active time difference from previous day

a sum mean was calculated for each item for each subject to determine which items were most likely to influence the results of the exercise motivation question for all subjects. Out of the calculated ones, the seven items with the highest values were used as explanatory variables for the model construction. The explanatory variables used are listed in Table 2.

The very active time (minutes) and somewhat active time (minutes) are not provided with precise definitions by the Fitbit manual, but it specifies the fairly active time (minutes) as representing very active time exceeding 6.0 METS [19]. 6 METS are achieved by, for example, 10 min of light jogging. Therefore, very active time (minutes) is considered to be the cumulative amount of time spent doing more intense exercise than fairly active time (minutes). For example, 8 METs can be obtained by running for 7 min. On the other hand, the somewhat active time (minutes) is considered to be the cumulative amount of time spent exercising more moderately than the fairly active time (minutes). For example, 15 min of brisk walking can result in the acquisition of 4 METs.

Although Fitbit has not provided an exact definition for deep sleep either, several previous studies have compared sleep data from Fitbit smartwatch-type action meters with polysomnography (PSG), where deep sleep has been compared and classified as N3 [20, 21].

A machine learning model was created to predict the user’s motivation to exercise using a random forest with the seven items in Table 2 as explanatory variables and the high or low motivation states described in Sect. 4.2 as objective variables. The random forest was run on Python 3.7.7 with RandomForestClassifier in the ensemble package of scikit-learn 1.0.2.

## 5 Survey Results and Discussion

### 5.1 Survey Results

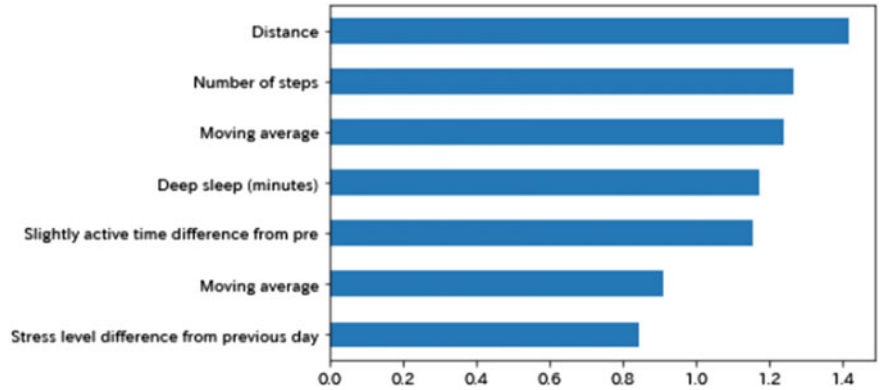
The accuracy of the model constructed in Sect. 4 was tested using the last week of the five-week study period described in Sect. 4.1 as the test data. The accuracy, precision, recall, and F1 scores for all subjects are shown in Tables 3 and 4, respectively.

**Table 3** Predicted results for a state of high motivation

Accuracy	0.81
Precision	0.28
Recall	0.36
F1	0.31

**Table 4** Predicted results for a state of low motivation

Accuracy	0.81
Precision	0.70
Recall	0.72
F1	0.70



**Fig. 3** Visualization results of feature importance

The results in Tables 3 and 4 show that the accuracy is high. Table 3 shows that it is difficult to predict a state of high motivation. On the other hand, Table 4 indicate that the prediction of low motivation can be made appropriately to some extent.

One reason for this result is the bias of the training data. As described in Sect. 4.3, the first quartile value was set as the threshold value for whether motivation is high or not, which is the objective variable in this study. Therefore, it is difficult to be highly motivated when the value is lower than the threshold value. On the other hand, it is easy to be in a low-motivation state, which is a state with a value higher than the threshold. The above results suggest that the number of low and high-motivation states is biased. Due to it, the accuracy of the prediction results for high-motivation states with a small number of states is not favorable, while that for low-motivation states with a large number of states is adequate to some extent.

## 5.2 *Validity of Variables in Predicting Motivation*

To investigate what variables had an impact on the prediction of users' motivation to exercise, the degree of important variables is visualized for the model constructed in Sect. 4. The visualization results are shown in Fig. 3.

The figure shows that distance is the most important variable for creating this model. It is very reasonable that the criterion of whether a person is gaining a lot of distance is necessary to predict motivation. It is also interesting to note that deep sleep (in minutes) is as important for predicting motivation as the number of steps taken or the slightly more active time difference from the previous day, both of which are directly related to exercise.

## 6 Conclusion

This paper proposes an exercise improvement method based on FBM that takes into account the user's activity status and motivation. In the proposed method, the minimum and maximum values are obtained from the smoothed step count for each day of the week. The period when the number of steps moves from the minimum to the maximum is considered to be a state of high ability because the number of steps steadily increases for a longer period than during other periods.

Based on the physical activity data obtained from the smartwatch-type action meter and the results of the user's answers to a questionnaire about his/her motivation to exercise, a machine learning model is built to predict the user's motivation, to provide interventions for users according to their motivation.

In this study, a survey was conducted to confirm the effectiveness of the machine learning model in predicting user motivation. The results of the survey showed that the percentage of correct answers exceeded 80%. In addition, the accuracy of the model was not good for identifying the state of high motivation. One possible reason for this is the lack of data on the number of highly motivated individuals. On the other hand, the accuracy is fairly high in a low-motivation state, which is a state with a value higher than the threshold.

In the future, to increase the number of data in the highly motivated state, it is necessary to add people who exercise regularly at the time of model training to prevent bias in the data. To achieve it, the study needs to expand the number of subjects, as well as conduct research over a longer period to obtain more general results.

The machine learning model that predicts the user's motivation has obtained a certain level of correct answers. To investigate whether the proposed method can improve the user's exercise, the study will incorporate the model into an exercise improvement method that takes the user's activity status and motivation into account.

## References

1. Hupin David, Roche Frédéric, Gremeaux Vincent, Chatard Jean-Claude, Oriol Mathieu, Gaspoz Jean-Michel, Barthélémy Jean-Claude, Edouard Pascal (2015) Even a low-dose of moderate-to-vigorous physical activity reduces mortality by 22% in adults aged  $\geq 60$  years: a systematic review and meta-analysis. *British J Sports Med* 49(19):1262–1267
2. Kraus WE, Powell KE, Haskell WL, Janz KF, Campbell WW, Jakicic JM, Troiano RP, Sprow K, Torres A, Piercy KL, et al (2019) Physical activity, all-cause and cardiovascular mortality, and cardiovascular disease. *Med Sci Sports Exercise* 51(6):1270
3. Carter T, Pascoe M, Bastounis A, Morres ID, Callaghan P, Parker AG (2021) The effect of physical activity on anxiety in children and young people: a systematic review and meta-analysis. *J Affect Disorders* 285:10–21
4. World Health Organization (2019) Global action plan on physical activity 2018–2030: more active people for a healthier world. World Health Organization
5. Romeo Amelia, Edney Sarah, Plotnikoff Ronald, Curtis Rachel, Ryan Jillian, Sanders Ilea, Crozier Alyson, Maher Carol et al (2019) Can smartphone apps increase physical activity? systematic review and meta-analysis. *J Med Internet Res* 21(3):e12053
6. Coughlin SS, Stewart J (2016) Use of consumer wearable devices to promote physical activity: a review of health intervention studies. *J Environ Health Sci* 2(6)
7. Fogg BJ (2009) A behavior model for persuasive design. In: *Proceedings of the 4th international conference on persuasive technology*, pp 1–7
8. Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, Dumuid D, Kasai D, O'Connor E, Virgara R, et al (2022) Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. *Lancet Digital Health* 4(8):e615–e626
9. Cadmus-Bertram LA, Marcus BH, Patterson RE, Parker BA, Morey BL (2015) Randomized trial of a fitbit-based physical activity intervention for women. *Am J Prevent Med* 49(3):414–418
10. Clawson J, Pater JA, Miller AD, Mynatt ED, Mamykina L (2015) No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist. In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pp 647–658
11. Rabbi Mashfiqui, Pfammatter Angela, Zhang Mi, Spring Bonnie, Choudhury Tanzeem et al (2015) Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. *JMIR mHealth uHealth* 3(2):e4160
12. Tran DT, Martinez I, Cross CL, Earley YF (2022) Mobile intervention pilot study in college students with elevated blood pressure. *J Cardiovascular Nurs* 10–1097
13. Agha S, Tollefson D, Paul S, Green D, Babigumira JB (2019) Use of the fogg behavior model to assess the impact of a social marketing campaign on condom use in Pakistan. *J Health Commun* 24(3):284–292
14. Akmal M, Niwanputri GS (2021) Spoonful: mobile application for reducing household food waste using fogg behavior model (FBM). In: *2021 international conference on data and software engineering (ICoDSE)*, pp 1–6
15. Patrick Kevin, Raab Fred, Adams Marc, Dillon Lindsay, Zabinski Marion, Rock Cheryl, Griswold William, Norman Gregory et al (2009) A text message-based intervention for weight loss: randomized controlled trial. *J Med Internet Res* 11(1):e1100
16. Figueroa CA, Deliu N, Chakraborty B, Modiri A, Xu J, Aggarwal A, Williams JJ, Lyles C, Aguilera A (2022) Daily motivational text messages to promote physical activity in university students: results from a microrandomized trial. *Ann Behav Med* 56(2):212–218
17. Advanced fitness + health tracker | shop fitbit charge 5. <https://www.fitbit.com/global/us/products/trackers/charge5>. Accessed on 20 Nov 2022
18. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, Hamilton CB, Li LC (2018) Accuracy of fitbit devices: systematic review and narrative syntheses of quantitative data. *JMIR mHealth uHealth* 6(8):e10527

19. Huberty JL, Buman MP, Leiferman JA, Bushar J, Hekler EB, Adams MA (2017) Dose and timing of text messages for increasing physical activity among pregnant women: a randomized controlled trial. *Trans Behav Med* 7(2):212–223
20. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC (2018) A validation study of fitbit charge 2™ compared with polysomnography in adults. *Chronobio Int* 35(4):465–476
21. Dong Xiaofang, Yang Sen, Guo Yuanli, Lv Peihua, Wang Min, Li Yusheng (2022) Validation of fitbit charge 4 for assessing sleep in Chinese patients with chronic insomnia: a comparison against polysomnography and actigraphy. *Plos one* 17(10):e0275287

# Implications of 3D Printing on Physical Distribution in Logistics and Supply Chain Management



Patrick Brandtner , Robert Zimmermann , and Jessika Allmendinger

**Abstract** 3D printing is a technology expected to have a huge impact on Logistics and Supply Chain Management (LSCM) and especially on physical distribution. The aim of the paper is to elaborate on the implications and benefits of 3D printing for (i) physical distribution strategies and processes, (ii) the role of the logistics service provider (LSP), and (iii) the connected processes of warehousing, picking and transport. By means of expert interviews, we analyze these implications and derive a set of impacts and potential future implications of 3D printing in LSCM. Our results show that experts expect huge potential from this type of technology. Experts agree that global transport can significantly be reduced in the future. However, it currently is limited in terms of handling large amounts of batch sizes and volumes as it is not designed for mass production. Furthermore, 3D printing will most likely take place in centralized and decentralized sites managed by companies or LSPs. The placement of 3D printers in private homes is currently not seen as a realistic option on a larger scale. As in current distribution approaches, LSPs will also play an important role in 3D printing-based distribution in the future. Experts expect them to offer corresponding business models (as e.g., print on demand) in the future. In conclusion, the importance of 3D printing as an alternative product method impacting LSCM will continue to rise.

**Keywords** 3D printing · Additive manufacturing · Physical distribution · Supply chain management · Logistics

## 1 Introduction

3D printing is a technology that is predicted to have a huge impact on logistics and supply chain management (LSCM). Recently, concrete application scenarios and use case for this technology started to emerge in, e.g., the context of smart manufacturing

---

P. Brandtner (✉) · R. Zimmermann · J. Allmendinger  
University of Applied Sciences Upper Austria, Steyr, Austria  
e-mail: [patrick.brandtner@fh-steyr.at](mailto:patrick.brandtner@fh-steyr.at)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_53](https://doi.org/10.1007/978-981-99-3091-3_53)

641

processes [1], spare parts management [2, 3] or in last mile distribution [4]. 3D printing hence will have a significant impact in future logistics and supply chain management. By reducing the need for physical transport, it has the potential to reshape traditional supply chains and logistics networks.

Predictions about the precise impact on SCM have been made for the last three decades. Still, actual, scientifically grounded research on its benefits and implications in practice are scarce [5]. Although already being used in a variety of industrial and mainly production-oriented settings, the real implications in LSCM still must be elaborated from a practical point-of-view. One of area that is thought to be heavily impacted by 3D printing is the physical distribution of goods in LSCM networks [6].

Thus, the aim of the current paper is to elaborate recent use cases and effects of 3D printing on exactly this area. By means of expert interviews, we analyze the implication of 3D printing on distribution logistics, the effects on the role of the logistics service provider (LSP), and the implications in warehousing, picking and transport. The paper therefore provides an expert-based view of 3D printing in distribution logistics grounded on actual and implemented use cases.

For separation of concern, we define the following three research question:

- RQ1: How does 3D printing effect the processes of and the strategic approach to distribution logistics?
- RQ2: How does 3D printing influence the role of logistic service providers involved in distribution processes?
- RQ3: How does 3D printing effect the processes of warehousing, picking, and transport?

The remainder of the paper is structured as follows: in Sect. 2, the background of the study, i.e., 3D printing, and distribution logistics is presented. Section 3 elaborates on the applied research method of the paper. Subsequently, Sect. 4 presents the detailed results, i.e., the impact of 3D printing on (i) distribution logistics in general, (ii) the role of the logistic service providers, and (iii) on warehousing, picking, and transport processes. Section 5 provides a discussion of results, and Sect. 6 concludes the paper, states limitations, and offers an outlook for future research opportunities.

## 2 Background

The background of the paper consists of the theoretical constructs of 3D printing and distribution logistics. Subsequently, these areas are described, and the theoretical background of the paper is presented.

## 2.1 3D Printing

3D printing is an umbrella term for a variety of additive manufacturing techniques. 3D objects, such as spare parts, are printed layer by layer based on physical or chemical melting or hardening processes. The main element of 3D printing is the layer-wise structure of the produced objects, i.e., the object is built by adding one layer after the other till the final form of the object is reached. 3D printing requires 3D models of the object to be built, e.g., in the form of Computer-Aided Design Models (CAD-models) [7]. Based on these models, a so-called 3D printer then produces or prints the object. Such printers may use a variety of input materials such as polymers, metals, ceramics, or other base materials.

For the current paper, 3D printing is defined as follows: 3D printing is the “*process of joining materials to make parts from 3D model data, usually layer upon layer, as opposed to subtractive manufacturing and formative manufacturing methodologies*” [8].

In contrast to the physical distribution of goods, 3D printing starts with (i) the developments of a 3D model. Subsequently, it covers the following steps: (ii) conversion of 3D model into STL format (standard triangulation language), (iii) consolidation and sending of the printing file, (iv) the installation and parameterization of the 3D printer, (v) the printing itself, (vi) the removal of the object from the printer, the (vii) postprocessing of the object, and finally (viii) the shipping to the usage destination or—if printed on site—the final usage [9, 10].

## 2.2 Distribution Logistics

LSCM has to fulfill a variety of tasks, ranging from procurement and transport, to inventory management, production and demand planning and to distribution and return-related activities [11]. In this paper, a focus is put on distribution logistics, comprising the physical distribution of goods from one location to another. 3D print might have several implications on this process, and as well on other processes as e.g., production planning and warehousing. Hence, our third research question also focusses on the processes of warehousing and picking.

In the current paper, the following definition of physical distribution is used: “*Physical distribution involves a group of interdependent parties involved in the process of distributing products or services so that they are ready for use by users with indicators of product availability, delivery time and place*” [12].

Several parties might be involved in physical distribution of goods, logistics service providers (LSP) being a central element of it. Our second research question addresses this important partner and elaborates specifically on implications on this interface.



### 3 Method

The base of our explorative research is a semi-structured expert interview. This was chosen as the nature of our topic is subjective, complex, and quantitative data are barely available. As such, expert interviews have shown to be well suited to collect insights regarding specific domain knowledge, which is difficult to uncover with survey-based methods only [13, 14]. A publicly available service and company profile platform was used to identify and contact possible experts in the domain of 3D printing. This platform gives an overview of European companies by providing data on employee numbers, contact details, and the product range of the respective companies. To identify suitable companies, we used the following criteria: Experience in the field of 3D printing and use of 3D printing processes. As such, we could identify 45 companies fitting to our criteria, which we contacted via e-mail. Out of these 45 companies, nine companies were interested in an interview and were subsequently interviewed. Information about the companies and the respective experts can be found in Table 1. Due to the huge physical distances between the authors and the experts, the interviews were conducted using phones. The interview data was later transcribed, and cross-interview comparison was used to identify commonalities and differences between the expert opinions [15, 16].

### 4 Results

In this chapter, we present the findings of our expert interviews regarding the impact of 3D printing on distribution strategies, the role of logistic service providers, and the processes of warehousing, picking, and transport.

#### *4.1 The Impact of 3D Printing on Distribution Logistics*

The impact of 3D printing on distribution logistics processes and strategy could be divided in centralized and decentralized distribution scenarios.

**Centralized distribution scenarios** During the expert interviews, only company 2, as producer of 3D printing systems, states that in the future “central locations” will be established, which will be used to develop and produce 3D printed parts. As a reason, the expert points out that in decentralized locations a consistent 3D print quality cannot be achieved. As such the experts suggest establishing central, or regional, 3D printing facilities, to bundle machinery and know-how. These facilities should be placed in junction areas, at approved service providers, or in subsidiaries. According to the expert, this is particularly interesting for large-scale companies with global and regional locations. Nevertheless, the expert states that even if the 3D printing systems are established in junction areas, local delivery is still a necessity.

**Table 1** Overview of companies and experts

#	Company orientation	Knowledge in 3D printing	Expert role	Employees	Revenue (2017)
1	Cleaning equipment for home and garden	Production and development of 3D prototypes	Regional Logistics SE EU	< 12.500	< 2.5bn
2	Manufacturing of machine tools and production facilities	Manufacturing 3D printers	Product manager 3D	< 7.500	< 2.5bn
3	Manufacturing of prototypes, products, and tools	Research and development of 3D printing products	CEO	< 300	< 50 m
4	Manufacturing of 3D print systems	Manufacturing 3D printers and on-demand 3D prints	Head of marketing	< 300	< 25 m
5	Mechanical engineering projects	3D printing metal parts using SLS process	Business organization Manager	< 50	< 10 m
6	Classical manufacturing processes	Production of models, prototypes, and small series by means of 3D printing	Head of Sales	< 50	< 10 m
7	Component manufacturing, refinement processes	3D metal printing	CEO	< 10	< 2 m
8	Prototype production, requirement development, research	Additive laser beam melting	CEO	< 10	< 2 m
9	3D printing of prototypes and components	Specialized in SLSL Print	CEO	< 10	< 2 m

Thus, 3D printing can only reduce global transports. A similar opinion is shared by the expert from company 1. The expert states that with centralized 3D printing locations, especially their spare part (more than 70,000 items) would no longer have to be shipped globally, thus avoiding large logistical efforts.

**Decentralized distribution scenarios** Particularly in the area of decentralized distribution scenarios for 3D printing, the experts expressed differing opinions.

*Local 3D printing facilities.* The first scenario mentioned was the placement of the 3D printing facilities locally. Only three companies can imagine that the local placement of 3D printing facilities, using several independent and smaller companies, is feasible. In this context, the expert from the comparatively small 3D printing company 9 currently do not advises buying 3D printing equipment. They argue that the investments for such equipment currently is very high, and thus 3D printing production is not profitable for most parts. In addition, they point out that know-how and high skilled workers are needed to meet the requirements in virtual construction. Although the number of people being able to operate 3D printing facilities is rising, the overall shortage of skilled workers is still very present. Also, the high costs of setting up such facility (up to 800.000€ for one additive manufacturing system) would be a high deterrence for local 3D printing facilities. As such, the expert regards a nationwide, Europewide or even worldwide expansion of local 3D print manufacturers as difficult.

The experts from company 2 points out that the CAD data used in the 3D printing process, once shared, is always reusable. In a local perspective this becomes problematic as local companies usually do not have partnership or other contractual relationships with each other. Thus, this could lead to the unrestricted printing of parts reducing the attractiveness of placing local 3D printing facilities.

However, the easy shareability of CAD is also seen as an opportunity. As such, the expert from company 1, 3, and 8 point out that especially spare parts could easily be made available in regions for which transportation is very time and cost intensive. Using local 3D printing facilities would, however, not only save shipping time and reduce monetary costs, it could possibly also reduce administrative costs, as sending files is seen as a much easier process compared to sending entire products.

Nevertheless, the expert from company 1 points out that it would also be fully feasible to use local service providers as a kind of extended workbench in order to save on delivery distances. In contrast, the experts from the smaller companies tend to reject this possibility. In this regard, they do not consider such cooperation as profitable as they fear increased competition and argue that their unique company performance would no longer be acknowledged.

*3D printing equipment in private households.* The printing of parts using commercially available 3D printers in private households is another decentralization scenario. According to company 6, the private 3D printing market is growing, with millions of printers already sold in the private sector. However, they point out that it will not be possible to produce complex 3D prints in the private sector in the future. Especially upstream and downstream processing steps, such as the correct calibration of the machine as well as turning and milling work are not feasible in a private setting. According to the expert of company 7, the vision that 3D printing can become an all-purpose solution, especially for private use, is a “fallacy” propagated by the media. Similarly, the expert from company 2 is also very skeptical in this regard. He states that one must understand how a printer works and which configurations must be set to produce a product with appropriate quality. In this respect, the expert questions to

what extent the quality of 3D prints can be kept consistently high for every household. The expert states that in a private setting, important parts of 3D printing (e.g., raw material quality, climate conditions) cannot be guaranteed, thus reducing the possible quality of private 3D prints.

*Mobile 3D Printing.* The scenario of 3D printing on the way between company and end user is considered unrealistic by most companies. Regarding metal printing, the experts from company 5, 7, and 9 state that it would be almost impossible to produce high-quality prints. This is because 3D metal printing systems are very sensitive, especially to vibrations. Also, they question if 3D printing to taking place inside vehicles or ships can be done economically as SLS systems weigh several tons and take several hours. The experts of company 5 and 7 see this scenario as more realistic in terms of plastic and smaller part printing. Nevertheless, they emphasize that still this will not be implementable on a large scale.

## **4.2 The Role of Logistic Service Providers**

The experts see great potential for large freight forwarders and other logistics providers in the use of 3D printing technology.

The expert from company 9 mentions that particularly in larger forwarding companies can use this potential to set up their own 3D printing centers and supply other locations. In this regard, the expert from company 2 cites an example from the company UPS, which plans to use 3D printers for plastic parts at their logistics nodes so that goods in demand can be produced and distributed closer to the customers. The expert states that the reason for this plan is that UPS regards 3D printing as a potential threat to their global business model. This is because 3D printing has the potential to reduce the need for products being shipped and thus could potentially make customers shift to predominantly hire local 3D printing service providers to do the manufacturing right next to their home address.

The expert from company 9 also mentions that processes can be eliminated because fewer parts need to be shipped globally in the future. In this respect, the expert regards it as quite realistic that shipping companies will also set up “printing centers” to produce parts and deliver them locally. The expert also mentions that setting up 3D printing systems in large logistics service companies is additionally attractive, since the transport system and the entire value chain can remain in use. The expert states that the product range of 3D printed objects will expand to meet global needs, while distribution will, however, remain local. Similarly, the expert from company 3 states that the logistics company “Dachser Group SE & Co. KG” is already developing scenarios in which the production of goods moves back to home countries thus reducing the global flow of goods.

The expert from company 4 suggests the introduction of “print centers”, which can take on the role of distribution center and 3D printing facility simultaneously. However, in comparison to traditional distribution centers, these “print centers”

would require much less space for warehousing as they could produce required items in an on-demand basis. A similar view is shared by the expert from company 8, which does not regard 3D printing as a “stand-alone” solution but as part of a process combination. In addition, the expert from company 8 imagines that logistics service provider can increase their importance using 3D printing technology if they open 3D printing service pools. These pools would group all 3D printing facilities in a certain area such as production orders would directly be allocated to the nearest or best-suited 3D printing facility. From this facility, the order would then be shipped to the target destination. Another possibility is to place 3D printing equipment in local competence centers or print stores of the logistics service provider. The experts from company 3 and 4 imagine, that there could be “printer shops” similar to the currently available “copy shops”. However, the expert from company 6 states that in this context, print centers would not be profitable as the demand for printing inexpensive small parts currently is very limited.

#### ***4.3 Impact on Warehousing, Picking, and Transport Processes***

In general, all companies state that little will change in respect to picking. Only in the areas of warehousing and transport significant changes are expected or observable.

**Warehousing** The expert of company 9 states that currently, warehousing remains very important. However, in the future, when shipping can be replaced by printing parts onside, warehousing becomes significantly less necessary. In this regard, the expert considers the implementation of a 3D printing department as vital in order to reduce the need for warehousing even further. The expert of the second company sees similar possibilities but points out that this transition will not be possible in the next few years and instead will take significantly longer. Taking a customer perspective, the expert of company 6 notes that 3D printing can reduce the time a customer needs to wait for a certain spare part. In this regard, for parts that were previously held in stock for months and thus had to be ordered in large quantities can now be produced on demand and no longer require immense stockpiling.

The experts from company 3, 5, and 8 note that the technology of 3D printing is not designed to keep large quantities of stock available but always the stock required for specific orders. This is already demonstrated by the companies 4 and 7. Instead of physical storage, company 4 specifies in the storage of digital CAD data. As such, they can send CAD data to companies with 3D printers and order a specific on demand production. In a similar fashion, company 7 was able to reduce the amount of stocked spare parts for one customer by 60%. Also, the expert of company 1 states that they see potential in 3D printing in order to reduce the number of stocked spare parts, which are not frequently needed.

**Transport** In terms of transportation, most experts see a change in a way that transport routes will shift from global delivery in the direction of local delivery. Due to the possibility of on-site 3D printing, the expert of company 5 estimates great time savings in the distribution process. Similarly, the expert of company 1 highlights that the production of parts can take place domestically in the future. Accordingly, the entire transport structure will increasingly run within regional distribution centers. Also, the expert from company 7 states that the only change in transport traffic will be seen in the decrease of volumes and weights of individual shipments.

**Picking** Concerning picking, the experts do not see any major changes happening. In this regard, the expert from the company 3 states that the logistics sector will only change when 3D printing production is installed locally and on a large scale. Company 1 also sees no change in the area of picking, as downstream processes after production continue to exist even with 3D printing in place. As such, goods still have to be sorted, packaged, and sent to customers. In addition, the expert of company 7 states that distribution hubs will not change or be eliminated as their importance remains the same even with 3D printing in place.

## 5 Discussion

Regarding the impact of 3D printing on distribution logistics, in the literature, the placement of 3D printing equipment at centralized and/or decentralized locations is seen as the future [17, 18]. However, reflecting on the results of our expert interview, different conclusions might be drawn. As such, our experts stated that the local and large-scale placement of 3D printing systems in many small locations, e.g., in small businesses, is possible. However, many experts view this approach critically, since on the one hand acquisition costs, qualified personnel, and quality standards can represent implementation barriers. Printing in private households is also considered as less feasible by the experts due to the lack of expertise and the possible fluctuations in quality. As such, only a few products could be produced in private printers. In particular, metal printing is not suitable for private use. As such, systems for metal printing also cost much more than small home printers for small products. A mobile solution in transport vehicles is also viewed critically by the experts. Such solution is seen as simply not feasible in terms of metal printing but could be feasible for the printing of small plastic parts. It can therefore be stated that many scenarios from literature are seen as less feasible in practice. However, the decentralized distribution of 3D printing equipment within small companies, but also in the hubs of larger companies with a widespread network of central warehouses, may find broader application in the future.

Regarding the impact of 3D printing on the role of logistic service providers, the literature shows that logistics service providers are already integrated into the processes of the focal companies and play an important role in the distribution of goods to the customer [19]. Through 3D printing, these logistics companies can

continue to take on a larger role and, for example, replace storage capacity in warehouses through printing on-demand [20]. Similarly, the interviews show that the role of logistics service providers can gain in importance through the use of 3D printing technology. Especially as they can use the already existing infrastructures to improve its efficiency by using 3D printing.

Regarding the impact of 3D printing on warehousing, our analysis shows the following. Generally speaking, 3D printing will not be able to replace classic manufacturing processes. Although a wide range of products can be produced by the 3D printer, it is unlikely that especially goods in regular demand will be replaced in large quantities by 3D printing. As such, these goods, will continue to require warehousing. However, the warehousing for irregularly demanded goods might be reduced by 3D printing. Overall, the premise that 3D printing will completely replace physical distribution centers and warehouses is thus unlikely. At most, experts mentioned a reduction in warehousing at the customer's premises, since at least a small amount of warehousing will always be required. In literature, similar findings can be found [17–19].

Taking the scenario of 3D printing in private households as an example, we can see that all sub-processes can be omitted. However, for every other scenario, there is no change or elimination of these activities. This was also stated by the interviewed experts. Even if printing is carried out locally, goods still have to be transported to the end customer. Even parcel service providers who carry out a last mile delivery, must at least transport, deliver and pick the goods, correctly in the process. Similarly, the analysis from literature and our research therefore shows that the processes of distribution logistics will not become obsolete unless 3D printing is done entirely at the customers home address. However, the transportation process might speed up because of 3D printing as locally printed products generally need to travel shorter distances.

## 6 Conclusion, Limitations, and Future Work

Addressing RQ1, it can be concluded that 3D printing effects distribution logistics in multiple ways. Generally speaking, 3D printing has the potential to bring the production line closer to the customer. As such, transports can also be moved to the local area. However, since 3D printing reaches its limits, especially for large-volume production, few changes are seen in distribution logistics overall. Nevertheless, it can be assumed that at least some global transportation can be reduced, with distribution logistics subsequently also focusing on local distribution. In contrast, scenarios that place 3D printing systems in mobile transports or private households are only possible for a few products and thus cannot find widespread distribution, while the central or local placement of such systems in specialized companies is seen as particularly relevant. Still, it must be noted that at least the distribution logistics from producer

to end customer must continue to be carried out and therefore, the associated sub-processes, such as storage in a distribution warehouse, transport and, if necessary, handling and picking processes, will not be eliminated.

Regarding RQ2, it can be concluded that logistic service providers gain a greater role through 3D printing technology. Although logistic service providers are already experiencing a great deal of integration in value-creating activities of focal companies, they also have great potential in terms of possible future business models, which make use of 3D printing technology (e.g., printing centers, print shops).

Regarding RQ3, it can be stated that 3D printing technology is not a process designed for mass production. Whereas in conventional manufacturing processes mainly economies of scale depress the price of a product, 3D printing cannot make use of this mechanism, as it focuses on the production of individualized products. Therefore, it is unlikely that conventional mass production will be completely replaced with 3D printing. For this reason, the demand for warehouses will not decline. However, 3D printing might optimize the use of warehouses as not every product needs to be kept on stock all the time but can be printed in an on-demand basis.

Regarding the processes of picking and transport, these physical processes of distribution logistics will not undergo great changes. However, it can be argued that the amount of transport and picking necessary for 3D printed products depends on the location of the printing machine. As such, printing directly at the customers home, implies that there is no need for delivery, storage or picking activities. In cases where printing takes place in special companies, little will change compared to today's distribution.

As every study, our study also has its limitations. Mainly because of the qualitative nature of our study, our results are subjective and only represent the views of the interviewed experts. Also, as we only interviewed nine experts from the German speaking area, the overall comparability of our results is thus limited. However, by applying the sophisticated expert selection procedure and the structured analysis approach, we ensured a reliable and proper gathering and analysis of results.

In summary, it can be said that 3D printing technology will continue to enter the corporate world in the coming years, with more and more companies using 3D printing as an alternative to classical manufacturing. As such, many fields of research require further consideration. In particular, in the field of supply chain management, the topic of 3D printing needs to be filled with more quantitative studies. In this regard, an interesting question would be to analyze the extent to which 3D printing technology affects global transportation. Also, it might be an interesting endeavor for future research to identify opportunities for future business models, which can arise for logistics service providers through the use of 3D printing technology.



## References

1. Ali S, Shin WS, Song H (2022) Blockchain-enabled open quality system for smart manufacturing: applications and challenges. *Sustainability* 14:11677. <https://doi.org/10.3390/su141811677>
2. Jakšič M, Trkman P (2018) 3D printing as an alternative supply option in spare parts inventory management. In: Fink A, Fügenschuh A, Geiger MJ (eds) *Operations research proceedings 2016. Operations research proceedings*, pp 617–622. Springer International Publishing, Cham. [https://doi.org/10.1007/978-3-319-55702-1\\_81](https://doi.org/10.1007/978-3-319-55702-1_81)
3. Rupp M, Buck M, Klink R, Merkel M, Harrison DK (2022) Additive manufacturing of steel for digital spare parts—a perspective on carbon emissions for decentral production. *Cleaner Environ Syst* 4:100069. <https://doi.org/10.1016/j.cesys.2021.100069>
4. Wang Y, Ropke S, Wen M, Bergh S (2022) The mobile production vehicle routing problem: using 3D printing in last mile distribution. *arXiv*
5. Beltagui A, Gold S, Kunz N, Reiner G (2023) Special issue: rethinking operations and supply chain management in light of the 3D printing revolution. *Int J Prod Econ* 255:108677. <https://doi.org/10.1016/j.ijpe.2022.108677>
6. Xiong Y, Lu H, Li G-D, Xia S-M, Wang Z-X, Xu Y-F (2022) Game changer or threat: the impact of 3D printing on the logistics supplier circular supply chain. *Ind Mark Manage* 106:461–475. <https://doi.org/10.1016/j.indmarman.2022.03.002>
7. Feldmann C, Schulz C, Fernströning S (2019) *Digitale Geschäftsmodell-Innovationen mit 3D-Druck*. Springer Fachmedien Wiesbaden, Wiesbaden
8. Lee J-Y, An J, Chua CK (2017) Fundamentals and applications of 3D printing for novel materials. *Appl Mater Today* 7:120–133. <https://doi.org/10.1016/j.apmt.2017.02.004>
9. Gibson I, Rosen D, Stucker B (2015) *Additive manufacturing technologies*. Springer, New York, New York, NY
10. Feldmann C, Gorj A (2017) *3D-druck und lean production*. Springer Fachmedien Wiesbaden, Wiesbaden
11. Brandtner P, Udokwu C, Darbanian F, Falatouri T (2021) Dimensions of data analytics in supply chain management: objectives, indicators and data questions. In: *2021 the 4th international conference on computers in management and business*. ACM, New York, NY, USA. <https://doi.org/10.1145/3450588.3450599>
12. Wilson T, Nguni W, Rwehumbiza DA (2022) The mediating influence of procurement strategy on the relationship between physical distribution and availability of contraceptives in public health facilities. *Bus Manage Rev* 25:42–59
13. Brandtner P, Udokwu C, Darbanian F, Falatouri T (2021) Applications of big data analytics in supply chain management: findings from expert interviews. In: *2021 the 4th international conference on computers in management and business*, pp 77–82. ACM, New York, NY, USA
14. Döringer S (2020) The problem-centred expert interview. Combining qualitative interviewing approaches for investigating implicit expert knowledge. *Int J Soc Res Methodol* 1:1–14. <https://doi.org/10.1080/13645579.2020.1766777>
15. Brandtner P, Mates M (2021) Artificial intelligence in strategic foresight – current practices and future application potentials. In: *Proceedings of the 2021 12th international conference on e-business, management and economics (ICEME 2021)*, pp 75–81
16. Albers S, Klapper D, Konradt U, Walter A, Wolf J (2009) *Methodik der empirischen Forschung. 3., überarbeitete und erweiterte Auflage*. Wiesbaden and sl: Gabler Verlag
17. Rogers H, Baricz N, Pawar KS (2016) 3D printing services: classification, supply chain implications and research agenda. *Int J Phys Dist Log Manage* 46:886–907. <https://doi.org/10.1108/IJPDLM-07-2016-0210>
18. Ryan MJ, Evers DR, Potter AT, Purvis L, Gosling J (2017) 3D printing the future: scenarios for supply chains reviewed. *Int Jnl Phys Dist Log Manage* 47:992–1014. <https://doi.org/10.1108/IJPDLM-12-2016-0359>

19. Bayraktar AN (2022) 3D Printing and Logistics. In: İyigün İ, Görçün ÖF (eds) Logistics 4.0 and future of supply chains. Accounting, finance, sustainability, governance & fraud: theory and application, pp 63–82. Springer Singapore, Singapore. [https://doi.org/10.1007/978-981-16-5644-6\\_5](https://doi.org/10.1007/978-981-16-5644-6_5)
20. Wu Q, Xie N, Zheng S, Bernard A (2022) Online order scheduling of multi 3D printing tasks based on the additive manufacturing cloud platform. J Manuf Syst 63:23–34. <https://doi.org/10.1016/j.jmsy.2022.02.007>

# Towards Prototyping Single-modal and Multimodal Interactions in Mixed Reality Games



Logan LaMont, Ged Fuller, Pratheep Kumar Paranthaman, Thomas Poteat, Dhvani Toprani, Qian Xu, and Nikesh Bajaj

**Abstract** Mixed reality (MR) blends the physical and digital environments using natural interactions. In MR, users experience computer-generated content in the physical world by wearing a head-mounted display unit. MR domain is still in its infancy with several open questions on immersion, comfort, user interactions, and user experience. Specifically, game user experience in MR is new to the research in game design and human-computer interaction. To understand the effective interaction modes with new technology such as MR, user experience analysis and interaction patterns need to be deeply explored. In this paper, we aim to explore novel ways in which we can effectively play games in an MR environment by analyzing different interaction modalities and user experience factors. This paper explains the prototypical process involved in developing 5 games in MR that include single-modal and multimodal interactions.

**Keywords** Mixed reality · Interaction modalities · Prototype development · Game design · User experience · Interaction design

---

L. LaMont (✉) · G. Fuller · P. K. Paranthaman · T. Poteat · D. Toprani · Q. Xu  
Elon University, Elon, NC, 27244, USA  
e-mail: [llamont@elon.edu](mailto:llamont@elon.edu)

G. Fuller  
e-mail: [gfuller3@elon.edu](mailto:gfuller3@elon.edu)

P. K. Paranthaman  
e-mail: [pparanthaman@elon.edu](mailto:pparanthaman@elon.edu)

T. Poteat  
e-mail: [tpoteat@elon.edu](mailto:tpoteat@elon.edu)

D. Toprani  
e-mail: [dtoprani@elon.edu](mailto:dtoprani@elon.edu)

Q. Xu  
e-mail: [qxu@elon.edu](mailto:qxu@elon.edu)

N. Bajaj  
Imperial College London, London, UK  
e-mail: [nbajaj@imperial.ac.uk](mailto:nbajaj@imperial.ac.uk)

# 1 Introduction

Mixed reality (MR), the blending of the physical and digital world, is the next step in spatial computing, and it is already starting to be utilized for many different applications. This includes keeping students engaged in learning [1], helping physical therapists in the rehabilitation process for their patients [6], improving collaboration in the workplace, or promoting a new form of immersive environment [5, 15]. When it comes to MR headsets there are a few options available and one of the newest options now is the Magic Leap 2. There have already been examples of this headset such as planning different surgeries, improving the onboarding process for manufacturers, and reducing the amount of mistakes workers are making as well as increasing their efficiency [9]. Another headset, Microsoft HoloLens 2, which was released in 2019 is the headset we used for our investigation in this project. The HoloLens 2 was designed for allowing users to better collaborate with each other in the virtual space that would be created within an application. The HoloLens 2 has many different functions, such as the ability to take in voice commands, recognize gestures to allow the user to interact with the environment, and a depth camera that can spatially map the surroundings with a 3D scan. [12]. There are a few potential challenges in making games for MR headsets such as the HoloLens 2 because there are some inherent limitations. One limitation is the field of view where the user can see holograms. The HoloLens 2 has a limited field of view which is approximately 54° (roughly half of what VR can offer) [14]. Another limitation is a headset such as the HoloLens 2 needs certain lighting conditions where it cannot be too bright or dark. Sometimes there can be physical discomfort associated with MR devices as well, with the weight of the headset or the user may not be comfortable with seeing holograms [4]. At the moment MR devices have limited hardware capabilities which makes it difficult to render complex 3D video game scenes so games have to be relatively simple.

Considering these limitations involved in MR game development process, in this project, we aim to explore the interaction modalities in HoloLens 2 and to prototype multiple games that are playable in the MR environment using different forms of interactions. Microsoft HoloLens 2 offers input and interactivity options like gaze, gestures, and speech commands, so our first step in this process is to identify: (1) what games can be played using different inputs and (2) which inputs contribute toward better user experience in the MR environment. For this, we started exploring single-modal (using only one form of interaction to work with the system like speech/gaze/gesture) and multimodal (combining more than one form of interaction like speech and gaze or gesture and gaze) interactions in the MR environment. In this paper, we present our prototyping process involved in developing 5 MR games using single-modal and multimodal interaction systems.

## 2 Related Work

The research conducted by [8] examined how physical and cognitive challenges contributed to immersion in an MR game. Sixty-eight participants played 34 games of the MR version of pong on a volleyball court. They were able to compare whether more physical challenges or cognitive challenges would lead to higher levels of immersion. They found that when the participants were faced with physical and cognitive challenges combined it led to the most immersion that participants are “doing while thinking.” This is a contrast to desktop games where physical movement is secondary to all else. In [11], the authors investigated players’ emotional factors in virtual reality (VR) games and for this study, they used an electroencephalogram (EEG) device to investigate the brain activity of players in three VR games. Also, in this study, the authors observed a significant difference between the performance metrics values (six emotional states of players) from the EEG system and players’ experience from the gameplay (self-reported questionnaire). This paper by [3] shows a conceptual surgery tool that uses the HoloLens headset and an Azure Kinect camera to simultaneously use body tracking and render a display of a model of the human body in real-time. The proposed use of a tool such as this would allow surgeons to plan out how they want to complete the surgery and it can also be used for learning, practicing those concepts in a simulated environment, and improving their skills. Research carried out by [2] investigates how the HoloLens headset could be used for patients with Alzheimer’s. With the HoloLens they developed an application using pre-existing strategies to combat memory loss and reduce the risk of strokes. Some of the MR-specific problems they ran into while developing included the movement of certain objects during the activity and this made the instructions to be unreadable. To mitigate this issue, they had to change their prototype so the instructions were always in view of the users regardless of where they looked. They proposed that MR will allow for the combination of all the different strategies used to try and combat Alzheimer’s.

An investigation by [13] explores a way the HoloLens 2 could be used to help alleviate phantom limb pain. Currently to alleviate phantom limb pain patients will sit in front of a mirror that shows them with a healthy limb which has been shown to reduce the pain felt by the patients. Using HoloLens and myoelectric sensors (EMG) technology they created a virtual limb that the patients can use to interact with the virtual environment. Once they begin testing they expect this approach to allow for a more immersive experience versus the mirror therapy and with these extra interactions help alleviate the pain that patients may feel. A study conducted by [7] used head-mounted displays (HMD) to assist assembly line workers’ interface so they can interact with the machines without having to manually interact and do everything virtually through the MR interactions with the headset. This study had participants to test a user interface that they interacted with using the HMD to more efficiently gather and process the needed information from the machines in the production line. The result of the study showed that this concept would be a sufficient enhancement to assembly line manufacturing, with the biggest limitation being hardware.

From the analysis of related work, we learned that MR technology is expanding in several domains like healthcare, education, and engineering. However, there is less research on applying MR in the field of games and this is due to the fact of inherent issues like limited field of view in the headset, lack of system resources/computational power to render complex 3D worlds, and physical discomforts associated with MR technology. Hence our main goal in this project is to explore the application of MR technology in the field of games and also to illustrate the prototyping process involved in developing games for the MR environment.

### 3 System Design

The MR games in this project were developed using the Unity game engine and mixed reality toolkit (MRTK) and the games were deployed on Microsoft HoloLens 2. For the design and development process, we selected three interaction modalities (gesture, gaze, and speech) in HoloLens 2 and developed games around them. In addition to the interaction modalities, we considered the limitations like the field of view, graphics, and physical discomfort. For the graphics, we used low-poly assets for seamless rendering and also to reduce the application size on HoloLens 2. To ensure comfort, we decided to keep all the games under 6 min, as anything beyond this can be overwhelming for new users in the MR environment. This assumption was based on our initial prototyping process using bodystorming approach (as explained in the next section). Also, we reduced all fast-paced movements within the games. The following sections describe the prototypical process we followed in developing the 5 MR games in this project.

### 4 Bodystorming Process

The early prototyping process in the development helps understand design aspects, playability issues, and feasibility. As a first stage in the process, we decided to apply bodystorming process to help us in gleaning insights into playability factors involved in the MR games. Bodystorming is the act of using physical objects to simulate what the product would look like if it actually existed [10]. This process takes the players' point of view into account by simulating what the product will do in the real world. To simulate this experience, we can use cheap arts and crafts to rapidly prototype elements needed for the experience. For this process, we built a mockup of MR headset out of straws so we can have an idea of what the HoloLens viewport will be like (see Fig. 1). We then used real-world measurements in order to decide how much space we had for all of our games. 3 m in every direction were the measurements we came up with for the area we had to work with for a playable experience. In this bodystorming process, we also roleplayed certain game mechanics and interactions of each game in order to fully understand how players will be moving around the



**Fig. 1** (left) a user wearing a mockup of MR headset for the bodystorming process in this project. This mockup MR headset was created using straws and masking tapes. (right) a user wearing Microsoft HoloLens 2 mixed reality headset

space and interacting with elements in our MR games. Specifically, we roleplayed picking up objects, gazing at objects, and walking around the space while wearing the mockup headset and playing within the confines of our measurements.

## 5 Prototype Development

After bodystorming process, we translated the real-world measurements and distance of objects from player in Unity game engine. This experience helped us smoothly transition into the development phase because we were able to set up certain goals for what we wanted the mechanics to act like and how we wanted the games designed. Following are the 5 MR games we developed in this project. Prototypes 1–4 involve single-modal interactions (gaze/gesture/speech) and prototype 5 involves multimodal interaction.

### 5.1 *Prototype 1—Gaze-Based Game*

In the first prototype game, we used gaze-based interaction and developed a tower defense game. In this game, the player uses gaze in order to defend their home base from aliens. The game consists of 3 towers and the goal of the player is to defend the towers from aliens. Aliens will constantly spawn in space until time runs out. There are 2 levels, each with an 80-second timer. Level 1 consists of aliens constantly spawning all around the player and going after a random tower. Level 2 increases the challenge by throwing in some allies that the player must not destroy (see Fig. 2).



**Fig. 2** (left) First prototype of gaze-based tower defense game deployed on Microsoft HoloLens 2. We used primitive shapes and basic UI in Unity to model this prototype. (right) the final version of the gaze-based game. After iterating over the first prototype, we arrived at this final version and the figure shows the viewport captured from HoloLens 2

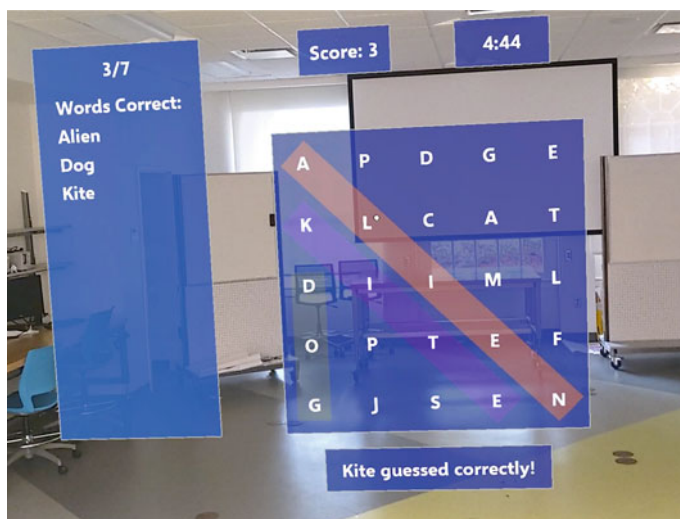
In this game, players will use gaze as a primary mode of interaction with the game system. Mechanics of this game include looking at an alien in order to destroy them in which the player will gain 1 point, and looking at an ally for a period of time will result in the player losing 5 points. In terms of the aliens' purpose, once they travel to and reach a tower, they will hit that tower once every 3 s. The player needs to make sure they are only destroying the aliens and not the allies in order to be successful in the game.

The player needs to keep at least 1 tower alive in order to go onto the next level, and at the end of level 2 the player must have at least 1 tower standing and a score of at least 25 to win the game. Each tower has a health bar to help the player know how close the aliens are to destroying one. A reticle will appear if the player looks at an object that is eligible to be destroyed (alien or ally). In terms of the UI, there is a scoreboard, timer, and level display that shows all information that is relevant to the player. During the bodystorming process, we learned that making users to frequently move their head for gazing mechanics in this game can cause neck strain. So to mitigate this, we reduced the speed of moving objects and also placed all the 3 towers to almost eye level, so that the users would not require to tilt head up and down when targeting flying aliens.

## 5.2 *Prototype 2—Speech-Based Game 1*

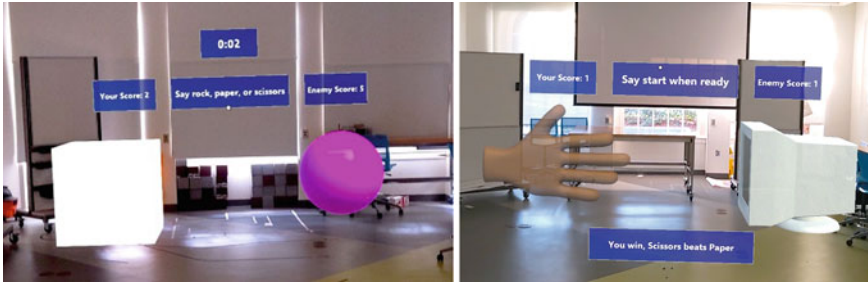
In this prototype, we developed a speech-controlled word search game where the player tries to find as many words as possible in a  $4 \times 4$  or  $5 \times 5$  grid of letters. 4 total levels exist in this game, 2 of them being on a  $4 \times 4$  grid and the other 2 being on a  $5 \times 5$  grid. A timer will continuously run for 5 min and the player has to try to go through every level before the time expires. The score will be tracked throughout the





**Fig. 3** prototype of our word search game deployed on HoloLens 2. The figure shows the game interface in HoloLens 2 and the interface contains the list of words correctly guessed,  $5 \times 5$  grid of letters, timer, current score, and feedback from system indicating the response to players' voice commands

game so if the player is unable to complete all word searches in time, they will still know how well they did. Speech/voice commands are the only way to interact with the game so this game uses a single-modal interaction. Players will say whatever word they find in the grid and the game will determine if that word is valid in the word search. Words in this game are at least 3 letters long, each level contains a certain number of words and once the player guesses all of those words, the next level will appear. If the player guesses all words in all levels within the time limit then they are declared as winner, if not, they lose and their final score will be presented. Once the player guesses a word correctly, a colorful outline will surround the word they guessed and that word will be added to a list of correctly guessed words in order to help the user know what they have already guessed. A total word count for the particular level the player is on will also be shown so they know how many words are present in the current level (see Fig. 3). From the bodystorming process, we learned that having the word search grids ( $4 \times 4$  or  $5 \times 5$ ) 4m away from the player will be an ideal position to scan through the grids without having to frequently move the head up and down or left and right.



**Fig. 4** (left) first prototype of speech-based rock, paper, scissors game developed in Unity and deployed on HoloLens 2. For this, we used basic elements like cubes, spheres, and simple UI. (right) final version of game with 3D models reflecting the player/computer choices

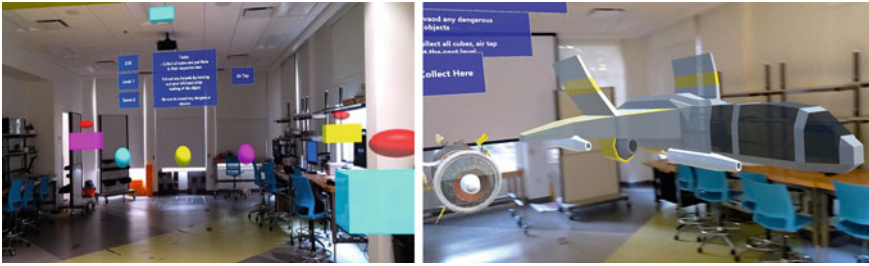
### 5.3 *Prototype 3—Speech-Based Game 2*

In addition to the word search game, we also created another speech-based game prototype and the reason for this prototype is to include more spontaneity within the speech-based game mechanics. During our trial runs of the previous prototype (word search game), we realized that the gameplay sessions ran longer/exceeded the specified time limit. In this game, we turned the rock, paper, scissors game into a speech-controlled one, where the player goes against the computer (see Fig. 4).

The player will provide a voice command as an input option for what they want to show (rock, paper, or scissors), then the computer will randomly pick an option to show as well (rock, paper, or scissors), and the outcome will be determined once both hands are shown. Whoever reaches the score of 7 first is declared as winner. When it comes to the UI and interface, the player will be able to see the score, when they are allowed to say their move, and what the outcome of the move was (win, lose, or draw).

### 5.4 *Prototype 4—Gesture-Based Game*

The fourth prototype in this project involved a gesture-controlled game where the goal of the player is to shut down hazards and collect objects in order to repair their spaceship. In this game, players will be trapped in space with a broken spaceship and they need to avoid the dangers of outer space in order to get back home. Players will face 3 levels with increasing difficulty. Each stage lasts for 210s and if the time runs out, the player loses. In the first level, the player will collect runes and put them into a sword and once all the runes are assembled, the sword will be used to slay enemies in future levels. In the second level, metal scraps will be collected into a broken engine to repair that engine. In the third level, barrels of oil will be collected



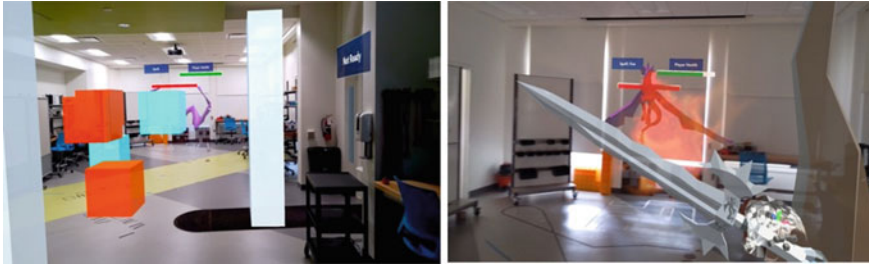
**Fig. 5** (left) first prototype of gesture-based game in unity and in this prototype, we tested the functionality of each gesture involved in this game. (right) the final version of this game deployed on HoloLens 2. The image shows the final level in the game with spaceship and engine models. This version was achieved after multiple iterations of the base prototype

into the repaired engine to get the engine properly running, which the player will then put into the ship to finish/win the game (see Fig. 5).

In terms of mechanics, players will show their left hand like a stop sign when close to a hazard (fire) on a collectible object (rune, metal scrap, and oil barrel) to get rid of that hazard. Players will then be able to pick up those hazardless objects and bring them to the proper spot to be assembled. Collectible objects will also be high up in the air and this is when players can use long-range grabbing in HoloLens 2 to collect those distant objects. Once the player has assembled all objects, they will air tap to go to the next level. In levels 2 and 3, there will be enemies moving around the space and the player must evade those enemies by walking around them. The sword collected from level 1 can be used by the player to kill the enemies. Finally, once the player collects all oil barrels into the repaired engine, their spaceship will appear and the player will place the fully repaired and running engine into the spaceship and escape. In this prototype, we incorporated gestures in HoloLens 2 like air tap, pinch, distance grab, and stop sign gestures (for turning off the fire).

### 5.5 *Prototype 5—Multimodal Interactions Game*

In this prototype, we developed a multimodal game that combines all 3 types of interactions used in previous games (speech, gesture, and gaze). In this game, players take the role of a wizard who is trying to defeat his arch-rival dragon. This dragon spawns minions that the player has to defend themselves against. The player needs to eliminate these minions and avoid the dreaded dragon's breath of fire in order to survive. While dodging all of these minions, players must charge up a sword that will eventually be used to slay the arch-rival dragon. Killing minions sent from the dragon is what charges up the sword until it is eventually fully charged which the player can pick up and attack the dragon. Three fully charged blows from the sword will slay the dragon and the player will win the battle.



**Fig. 6** (left) Initial prototype of multimodal game in unity. (right) the final version of the game with visual effects, sounds, and 3D models. In this viewport, player is holding the sword using right hand (pinch gesture) to attack the dragon

For the game mechanics, players have two spells (fire and ice) that they can use to kill the respective type of minion (fire and ice). The players execute the spell by saying the name of the spell as a voice command and then air tapping with their right hand while aiming (gazing) at their target. Once their target (minions) dies, a health pack has a chance to spawn that will slightly heal the player. Players also have the ability to freeze enemies by holding out their left hand and showing a stop sign while looking at an enemy. Once the sword is fully charged up, the player can pick up the sword, walk over to the dragon, and hit the dragon dealing a third of its health in damage. The dragon will spit fire throughout the game and the player has to use the generated cover in order to avoid getting eviscerated by the fire breath (see Fig. 6).

In this multimodal game, for selecting the attack spell (fire or ice) player uses voice commands, for aiming at the enemies player uses gaze, and for attacking the enemies, the player uses air tap and stop sign gestures. Finally for attacking the dragon player uses pinch gesture on the sword for picking the sword. Minions will spawn in 3 waves and after eliminating the entire wave of enemies, the sword will be fully charged. Players must be careful to avoid minions and the fire-breathing dragon as it will affect their health component in the game. Playtime depends completely on the player and how fast they are able to clear each wave, but this game should only take roughly 4–5 minutes to play. When it comes to UI, the player can see their health, the dragon's health, and if the sword is fully charged or not.

## 6 Conclusion and Future Work

The development of mixed reality applications involves spatial computing and modeling real-world interactions in the MR environment. So applying effective strategies for prototyping MR applications can streamline the development process. During the development process in this project, we noticed: (1) deploying games onto HoloLens 2 can be a time-consuming process, (2) using low-poly assets can improve game optimization and rendering, and (3) having play space between 3 and 4 m can be ideal

for spatial mapping in the MR environment. Lastly, using prototyping techniques like bodystorming can help in identifying project requirements, gathering insights on interactions, spatial constraints, and player comfort aspects.

In this paper, we presented the process of prototyping and developing 5 MR games and specifically, our prototypes were focused on using single-modal and multimodal interaction modalities in MR environment. We believe that the findings from this paper will contribute to ongoing research in experience design in immersive technologies and games research in mixed reality. For the future work, we plan to conduct user study with these 5 MR games and for the study, we will incorporate an electroencephalogram (EEG) system to capture the brain activity of the users while they play these MR games. From the EEG data, we will analyze the user emotions and experiences associated with the gameplay. In addition to user emotions, we will also conduct a comparative analysis between single-modal and multimodal MR games and investigate, the impact of each interaction modality on the gameplay.

## References

1. Akçayır M, Akçayır G, Pektaş HM, Ocak MA (2016) Augmented reality in science laboratories: the effects of augmented reality on university students' laboratory skills and attitudes toward science laboratories. *Comput Human Behav* 57:334–342 <https://www.sciencedirect.com/science/article/pii/S0747563215303253>
2. Aruanno B, Garzotto F, Rodriguez MC (2017) Hololens-based mixed reality experiences for subjects with alzheimer's disease. In: *Proceedings of the 12th Eiannual conference on Italian SIGCHI chapter. CHIItaly '17*, association for computing machinery. New York, NY, USA. <https://doi.org/10.1145/3125571.3125589>
3. Castelan E, Vinnikov M, Alex Zhou X (2021) Augmented reality anatomy visualization for surgery assistance with hololens: Ar surgery assistance with hololens. In: *ACM international conference on interactive media experiences. IMX '21*. Association for Computing Machinery, New York, NY, USA, pp 329–331. <https://doi.org/10.1145/3452918.3468005>
4. Fagan K (2022) Here's what happens to your body when you've been in virtual reality for too long. *Business Insider*. <https://www.businessinsider.com/virtual-reality-vr-side-effects-2018-3>
5. Feick M, Tang A, Bateman S (2018) Mixed-reality for object-focused remote collaboration. In: *The 31st annual ACM symposium on user interface software and technology adjunct proceedings. UIST '18 Adjunct*, Association for Computing Machinery, New York, NY, USA, pp 63–65. <https://doi.org/10.1145/3266037.3266102>
6. Fu Y, Hu Y, Sundstedt V (2022) A systematic literature review of virtual, augmented, and mixed reality game applications in healthcare. *ACM Trans Comput Healthc* 3(2). <https://doi.org/10.1145/3472303>
7. Hahn J, Ludwig B, Wolff C (2018) Mixed reality-based process control of automatic printed circuit board assembly lines. In: *Extended abstracts of the 2018 CHI conference on human factors in computing systems. CHI EA '18*, Association for Computing Machinery, New York, NY, USA, pp 1–6. <https://doi.org/10.1145/3170427.3188652>
8. Hu G, Bin Hannan N, Tearo K, Bastos A, Reilly D (2016) Doing while thinking: physical and cognitive engagement and immersion in mixed reality games. In: *Proceedings of the 2016 ACM conference on designing interactive systems. DIS '16*, Association for Computing Machinery, New York, NY, USA, pp 947–958. <https://doi.org/10.1145/2901790.2901864>
9. Leap M (2022) Magic leap 2. <https://www.magicleap.com/magic-leap-2>

10. Microsoft: thinking differently for mixed reality. Microsoft Research Blog (2020). <https://learn.microsoft.com/en-us/windows/mixed-reality/discover/case-study-expanding-the-design-process-for-mixed-reality>
11. Paranthaman PK, Bajaj N, Solovey N, Jennings D (2021) Comparative evaluation of the EEQ performance metrics and player ratings on the virtual reality games. In: 2021 IEEE conference on games (CoG), pp 1–8 (2021)
12. Pollefeys M (2020) Microsoft hololens 2: improved research mode to facilitate computer vision research. In: Microsoft research blog. <https://www.microsoft.com/en-us/research/blog/>
13. Prahm C, Bressler M, Eckstein K, Kuzuoka H, Daigeler A, Kolbensschlag J (2022) Developing a wearable augmented reality for treating phantom limb pain using the microsoft hololens 2. In: Augmented humans 2022. AHs 2022, Association for Computing Machinery, New York, NY, USA , pp. 309–312. <https://doi.org/10.1145/3519391.3524031>
14. Skalski P, Tamborini R, Shelton A, Buncher M, Lindmark P (2011) Mapping the road to fun: natural video game controllers, presence, and game enjoyment. *New Media Soc* 13(2):224–242. <https://doi.org/10.1177/1461444810370949>
15. Sylaiou S, Kasapakis V, Dzardanova E, Gavalas D (2018) Leveraging mixed reality technologies to enhance museum visitor experiences. In: 2018 international conference on intelligent systems (IS). pp 595–601

# Possibility of Utilising Information Technology to Promote Local Production for Local Consumption of Agricultural Products and Future Challenges



Tomoko Kashima, Takashi Hasuike, and Shimpei Matsumoto

**Abstract** Currently, many agricultural sites do not manage information that can be used for analysis of production, distribution, and consumption of local produce. Even if data are available, it is not integrated and used effectively. In particular, the balance of supply and demand of production is not visualised at the point of production, and production is conducted based on experience and intuition, which results in over-supply or shortages. Furthermore, the link between the consumer and producer is tenuous, and the requirements of both sides are not shared. Therefore, accurate and interactive sharing of information is needed between the producer and consumer, the producer and distributor, and others. Therefore, in this study, we visualised the supply situation faced by the producer and the supply situation of local agricultural products for business operators. We also surveyed consumer behaviour, collected information to expand the supply of locally grown agricultural products, digitised the current status, and visualised the balance of supply and demand to serve as reference for methods to share information in future, and investigated the future possibility of utilising digital technology for agricultural DX.

**Keywords** Local production for local consumption · Agricultural information · Agricultural DX

---

T. Kashima (✉)  
Kindai University, Hiroshima, Japan  
e-mail: [kashima@hiro.kindai.ac.jp](mailto:kashima@hiro.kindai.ac.jp)

T. Hasuike  
Waseda University, Shinjuku City, Japan

S. Matsumoto  
Hiroshima Institute of Technology, Hiroshima, Japan

## 1 Introduction

In recent years, there has been a demand for promoting innovation by utilising digital technology in agriculture; this includes the plan compiled by the Ministry of Agriculture, Forestry and Fisheries for agricultural DX [1]. However, data for analysis are lacking in many regions, which makes it difficult to ascertain the current situation. Therefore, it is necessary to collect data using information technology and understand the current situation, placing as little burden on producers as possible. By replacing areas that are currently implemented based on experience and intuition with data, and organising the currently available data into useable information, it becomes possible to understand the current situation, determine targets, predict sales and profit, conduct planned production, increase the ratio of local production for local consumption, and examine means to improve consumer satisfaction.

In this study, we visualised the status of the supply of agricultural products for local direct sales stores in city A, visualised the status of the supply of local agricultural products for business operators, and surveyed consumer behaviour unique to the region. This enabled the examination of ways to expand the supply of local agricultural products in the future.

## 2 Visualising the Supply Situation of Agricultural Products for Local Direct Sales Stores

Inventory in direct sales stores is not managed. The volume of sales can be checked using the POS system, but the producer may display products in the store, so it is not possible to confirm the inventory unless the producer himself/herself determines the quantity shipped to the store and calculates the difference with the volume of sales. This situation makes it difficult to manage the inventory for the entire store. Therefore, the producer and store staff manage the inventory through visual checks only. When the inventory is low, a member of the store staff telephones the producer to request product shipment. Therefore, in addition to visualising the current supply status of agricultural products at the store, we confirmed whether supply balance could be achieved by transmitting the visualised information to the producer in real time. We also confirmed that enabling the producer to check the store's supply status in real time makes it possible to check whether the producer's behaviour is changing.

### 2.1 Implementation Details

The content of the four items listed below was implemented.

- Four web cameras were installed in a store (direct sales store B) to determine the inventory status using web cameras, and the data were analysed.



- The producers were provided with information on the experiment and information on the transmission of inventory status through the web cameras (approximately 1000 mails were posted, and briefings and POP attachments were sent), and the logs were analysed to confirm the usage status of web camera transmission.
- Shipment was managed in the direct sales stores (targeting approximately 1000 people), an entry control system was introduced, human detection sensors were installed (system construction), and entry data were collected to analyse producer behaviour.
- The producers were surveyed after completion of the experiment.

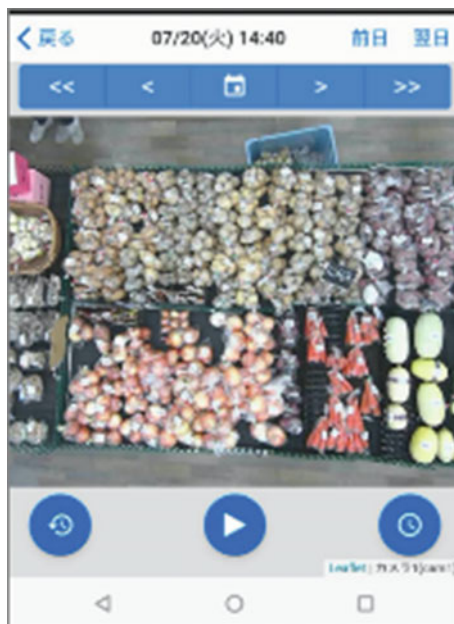
**Determining Inventory Status with a Web Camera.** The equipment shown in Fig. 1 was installed in the ceiling of the store as shown in Fig. 2 to confirm the inventory status of the store. Image data were collected from the four cameras at 5-min intervals during business hours from July to November 2021. The data were transmitted to producers who wished to view the images using an application [2] (Fig. 3). Approximately, 50,000 images were collected, and the inventory information was confirmed by image analysis.

**Fig. 1** Installation equipment



**Fig. 2** 4 webcams



**Fig. 3** Browse by App

**Experiment Information for Producers.** The experiment instructions (Fig. 4) were sent to approximately 1000 producers with registered stores. The producers were notified of information seminars (Fig. 5), and the study participants were thus recruited. A briefing session was also held for the participants to confirm information such as the producer's current situation and requests.

**Managing Shipments to Direct Sales Stores.** Producers who agreed to participate were instructed to use the application and freely view the store conditions. This application enables viewing of the images from the four web cameras in almost real time. It is also possible to check image data of specific dates and times from the recorded data, and ascertain how the store changes throughout the day.

Furthermore, when the producer visited the store to determine the shipment status, the producer was requested to cooperate in the registration of store entry and exit records by having a QR code attached to the name tag required for entry into the store (touch data input) (Fig. 6). Data were also collected on entry to the backyard by installing a human detection sensor (Fig. 7). These images were available for viewing by the producer to confirm if any changes in behaviour had occurred after understanding the inventory status.

**Producer Survey.** Producers who cooperated as participants were surveyed to verify the effect of visualising the store.

**となりの農家高屋店 WEBカメラ実証実験**

**実験協力モニター大募集**

■「実験内容」とは

- 農産物直売所に出品する生産者用のスマートフォン・アプリを導入し効果を検証します。
- 直売所の“いま”の画像を確認することにより、生産者は農産物の売れ行き状況をリアルタイムかつ直感的に把握し、適切な出荷・陳列、出荷作業の効率化を実現できます。
- 過去の売れるタイミングも時間経過とともに確認できます！

お願い


アプリの利用による出荷履歴を確認するため入室管理にご協力いただきます。コロナ禍における直売市での感染発生時の緊急対応の手段としての効果も想定しておりますので、多くの方にご協力いただければ幸いです。また、アプリに対するご意見を簡単なアンケートにてお答えをお願いさせていただきます。詳細は裏面をご確認ください。

■「WEBカメラ」アプリで出来ること

**機能1**：今の店舗の様子をリアルタイムに  
見たいカメラを選択してください。



**機能2**：過去の店舗の様子を  
カレンダーから見たい日を指定してください。



**機能3**：朝・昼・夕方、各時刻の様子を知る



**機能4**：画像を拡大する



Fig. 4 Experiment guide



Fig. 5 Briefing by stores

**Fig. 6** Access control (name tags and readers)



**Fig. 7** Motion sensors



### 3 Visualising the Supply Situation of Local Agricultural Products for Business Operators

A survey was conducted among business operators. For this study, we surveyed business operators that supply school lunches. City A has set a target of expanding the local production for local consumption initiative to include school lunches, based on the Basic Plan for Promotion of Agriculture. However, no specific surveys (calorie-based, weight-based, or monetary-based) have been conducted on the rate of local production for local consumption since 2016, so the current situation is unclear. No specific figures have been listed for the targets either, so it is unclear how far and how the targets should be achieved. Therefore, it is necessary to establish specific targets based on a clear understanding of the current situation.

#### 3.1 Implementation Details

The content of the two items listed below was implemented.

- Conduct interviews to ascertain the current data on school lunches in City A (3 centres) and the method of communication.
- Collate three years' worth of material data on the current situation, understand the rate of local production for local consumption (weight-based), set targets, and disclose information.

**Confirming Data Transmission Status.** The ingredients used for school lunches, from production through to the kitchen, were confirmed. Additionally, when and how this information is communicated, and how this information is used were confirmed. The media used to communicate this information, the type of data transferred, and the time of transmission were also confirmed.

**Confirming Rate of Local Production for Local Consumption.** The rate of local production for local consumption was confirmed by collecting data on the current menu and delivery data based on local purchases. Previously, the Ministry of Education, Culture, Sports, Science and Technology used a calculation method based on the number of ingredients, but the decision was made in the Fourth Basic Plan for Promotion of Shokuiku (Food and Nutrition Education) published in March 2021 to use monetary-based calculations [3]. However, as this survey lacks the complete range of monetary-based data, weight-based calculations were used initially.

## 4 Consumer Behaviour Survey

A survey system was developed, and data were collected to efficiently collect consumer information required by the producers. Previously, direct sales stores conducted paper-based surveys. However, in direct sales store B, only about 100 responses were collected over approximately one year. Conversely, in direct sales store C, as many as three surveys are collected each day. The content of the surveys is limited to asking the respondent about their level of satisfaction with the store, so no information was collected that would be useful to the producers. Therefore, in this study, we conducted interviews with the producers and thereafter conducted surveys classified into the following three approaches.

### 4.1 Implementation Details

The content of the three items listed below was implemented.

- Confirm the demand of people purchasing local ingredients.
- Confirm the demand of people using the direct sales store.
- Confirm the demand of people living in the local area.

**Demand of People Purchasing Local Ingredients.** A sales experiment was conducted targeting consumers, based on consumer information required by the producers. This is because it is necessary to observe consumers' purchasing behaviour to understand their real intentions and motivations that the consumers themselves may be unaware of, which cannot be measured with methods such as questionnaires. In fact, it has been said that consumers are aware of only 5% of their own behaviour and that 95% of their behaviour is unconscious [4]. Therefore, it is

presumed that responses to questionnaires are the person's public stance. Hence, in this study, we confirmed demand by observing behaviours that are difficult to express in a questionnaire system. This approach was implemented by displaying several products comparable in terms of size, appearance, and freshness, observing consumers to determine which products they purchase, and then asking the consumers why they purchased a particular item. In this experiment, consumers were interviewed only regarding purchased products (products placed in the shopping basket), which made it possible to ask about the consumers' true intentions, unlike with normal questionnaires.

**Demand of Users of Direct Sales Stores.** In this study, we developed a survey system that can be easily answered using a tablet, and data were collected. The system was installed in the stores. Unlike with previous paper-based surveys, the data collection rate was improved by installing a system in which questions can be easily answered using a touch panel. The questionnaire consisted of four questions. One question asked how frequently the consumer visits the store, and another asked about their family structure. The remaining two questions were random questions on information desired by producers. People who answered the survey were able to view a fun animation at the end, and respondents were entered into a prize draw with a prize of tea.

**Confirming the Demand of People Living in the Local Area.** Surveys conducted in-store only obtain information on the opinions of consumers visiting the store. Therefore, an online survey was conducted for people living in the area and people involved in the work to clarify the perception of the direct sales store and local vegetables, including that of people who did not visit the store during the survey period, and clarify information requested by the producer.

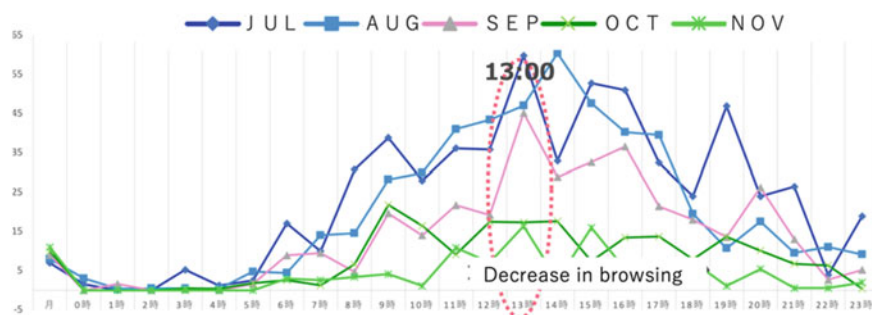
## 5 Results

The possibility of utilising information technology in local production for local consumption of agricultural products was verified through the present survey. Consequently, the following points were clarified for the verification matters for each item.

### 5.1 *Supply Situation at Local Direct Sales Store*

As detailed in Sect. 2, we confirmed the supply status of direct sales stores located in areas where many local producers ship their produce.

In terms of viewing the camera images, the images were viewed during the daytime through to the evening in the three-month period from July to September, but the number of viewers decreased significantly from October onward (Fig. 8). Although



**Fig. 8** Average number of image views per month

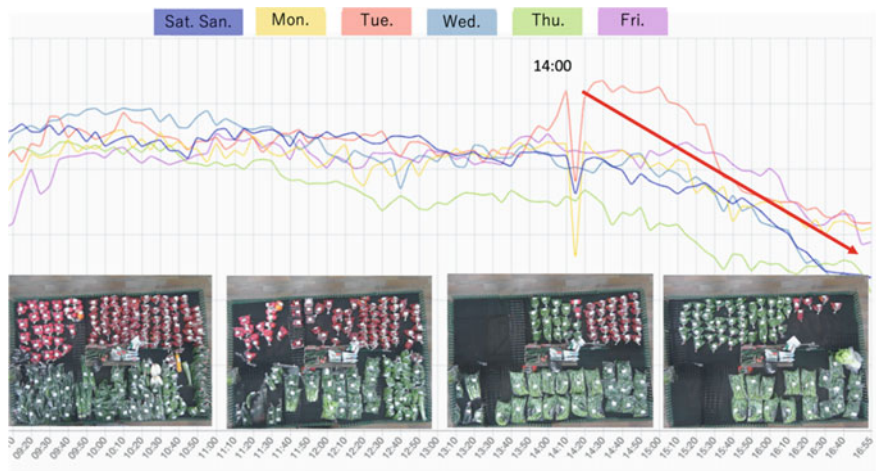
the producers used the system with the expectation of acquiring information on when, to whom, and how their own products were sold, how their products were handled, and how many products they would need to bring the following day, their feedback indicated that they found it difficult to locate their own products owing to the resolution and size of the images. Feedback from the direct sales store personnel included ‘it is difficult to determine what to focus on and what actions should be taken’ because the sales tended to be the same each day, and they were unable to access data analysing anything other than images. This is assumed to be the reason for the decline in the number of producers viewing the images.

As shown in Fig. 9, it was confirmed that the inventory decreases after 14:00. Producers could confirm from the application that store inventory had decreased, but there was no change in the producers’ shipment time (Fig. 10). It is thought that loss of opportunity could be avoided if producers made further shipments once the inventory was low, but producers concentrate their shipments before the store’s opening, and almost never ship products after 14:00, when the inventory is low. It is convenient for the producers to ship products before the store opens owing to the timing of farm and other work, and it is difficult for them to make shipments in response to the inventory situation in the store. The study clarified that cooperation on the store’s part is essential to address these challenges. The results of the questionnaire revealed that additional information required by producers other than the inventory status includes determining the appropriate price for selling the products, as well as the market price and local selling price.

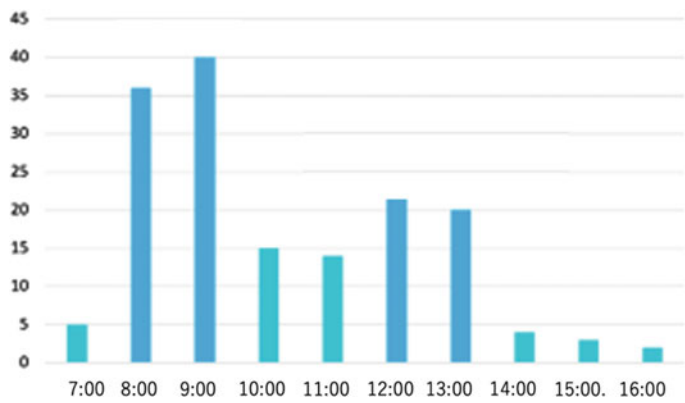
## 5.2 Supply Situation of Local Agricultural Products for Business Operators

As shown in Sect. 3, when the status of business operators in terms of local production for local consumption was confirmed, the following information was clarified.





**Fig. 9** Direct sales inventory information by image identification (weekly average)



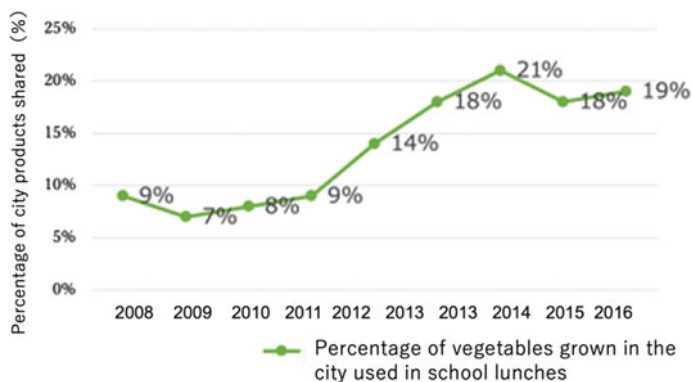
**Fig. 10** Shipping time of the producer's produce (number of people)

This time, personnel at school lunch centres, the local agriculture product coordinator, producer coordinators, and producers, namely the parties involved in school lunches, were interviewed to obtain information on school lunches, including how information is communicated during the process from deciding the ingredients for school lunches through to procuring the materials.

City A has previously issued reports on the results of surveys on the rate of local production for local consumption, but the data cover only until 2016, and it is assumed that these reports are the results of calculations based on ingredient items (Fig. 11).

The local agriculture product coordinator is notified of the ingredients used in school lunches by email one month before use. This information is also shared with market personnel. Each producer coordinator reports to the coordinator information



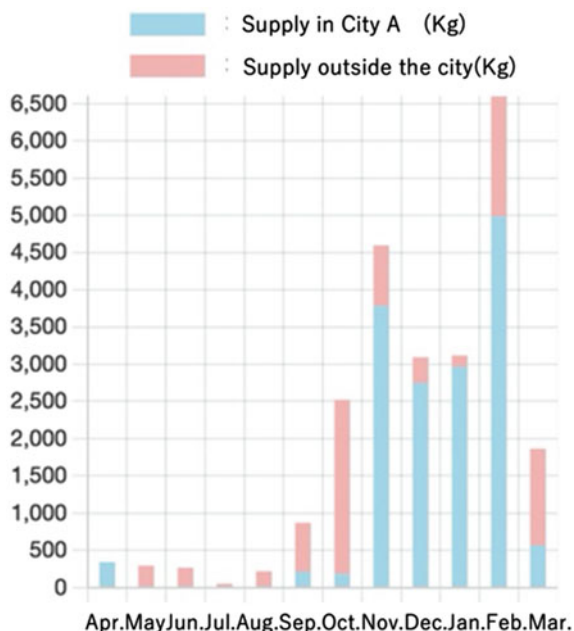


**Fig. 11** Supply of local agricultural products to school meals through 2016

on the vegetables being produced by local producers. Local ingredients are allocated based on this information, and the goods available for delivery are notified to the market personnel. The market personnel purchase the local shortfall of vegetables from suppliers outside the city to secure the required amount. The process involves the producers transporting the vegetables to a designated location, the person in charge of the designated location transporting the vegetables to the market manager, and the contractor in charge for that month delivering the vegetables to the school lunch centres. The process is very complicated, and no one has a grasp of the entire process. Data are not uniformly managed, as vegetable orders and delivery data are managed separately at each location. There are also other data, such as FAX data, that require aggregation.

Next, we tabulated 35 items from the FY2020 data for the main vegetables used in school lunches. Figure 12 shows data on the order volume for each month for a certain vegetable. The usage rate of local produce in the total amount used in the menu is shown in blue, while produce sourced from outside the city is shown in red. This information reveals that the vegetable is used in winter, and in October and March, the local supply is low, so it is sourced from outside the city. City A was unaware of this situation, as were the local producers and concerned parties. Essentially, it was confirmed that the majority of the vegetables are sourced from outside the city even during times when the vegetables can be produced locally. We also confirmed when the vegetables are available locally. If it was known at the time the data were aggregated that the vegetables would be required in March, local producers would have had sufficient time to establish planting plans and harvesting plans, but the information was not communicated well and was not made available. The cause seems to be a lack of communication. Moreover, as the producers were not informed of the requirements for the school lunches, they did not undertake planting with awareness of the school lunches. Hence, rather than adopting a planned approach, the producers simply selected the vegetables from what was currently available in the field and delivered them for the school lunches.

**Fig. 12** Visualisation of local production for local consumption rate



Clarifying the quantity of vegetables that can be produced within the city made it possible to set future targets. It is considered possible to increase the current rate of local production for local consumption by persons involved in school lunches sharing information as they decide on the school lunch menus (required vegetable ingredients). In addition, if school lunch nutritionists are aware of the producers' monthly production volume and varieties, they may be able to incorporate locally produced (surplus) vegetables into the menu. Sharing producer information with nutritionists may contribute to further improving the rate of local production for local consumption. Since further verification is needed to determine specific figures, this information is not published and is omitted.

Producers are keen to tackle food and nutrition education for local children. The advantages of providing vegetables for school lunches were clarified, which include that it is unnecessary to wrap each vegetable individually for delivery for school lunches, which consequently makes shipping easier and increases the profit from this produce as expenses such as packaging costs and labour costs are reduced. However, to improve the rate of local production for local consumption in terms of school lunches, the issues that should be resolved to increase production are also clarified. Lack of communication of information is often cited as an issue that should be resolved, as mentioned earlier. Other issues include determining methods for setting prices, allocating orders, disseminating information on the standards required for delivery, taking weight measurements at delivery, and ensuring guarantees during sudden school closures. An approach that will prevent potential problems is needed. Suitable pricing in particular significantly increases the income of producers and

improves the motivation for production and is considered an important factor for securing successors in the farming sector. Therefore, it is important to create decision-making processes and systems that are satisfactory for everyone. Causes of problems with delivery include the producers being unaware of the specific school lunch situation and the standards required for school lunches. Moreover, vegetables not specifically grown for school lunches sometimes end up being used for school lunches, which is considered to cause problems. In fact, 98% of approximately 200 local producers of vegetables supplied for school lunches were unaware that their vegetables were supplied for school lunches. Problems can be prevented by producers undertaking planned production for school lunches, and this may promote efforts towards highly satisfactory local production for local consumption.

### 5.3 Results of Survey of Consumer Behaviour

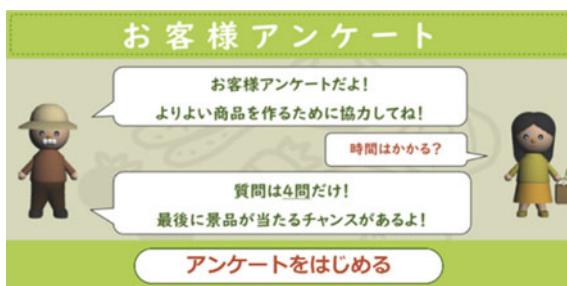
The following points were clarified while confirming consumer behaviour, as presented in Sect. 4.

In this study, we developed a questionnaire system that allowed the questions to be answered easily using a tablet (Figs. 13 and 14). The system was installed in store C for approximately 50 days.

Approximately, 1040 responses (20 responses per day) were obtained using a web-based system. The breakdown of these responses is as follows: 17 responses per day on weekdays and 32 responses per day on weekends and holidays. The system operated with no decline in the number of responses from introduction through to the final week, which was 8 weeks after installation. The results of the responses show that the reason for purchasing locally produced vegetables is the freshness of the vegetables, and the degree of freshness is determined based on their appearance (not shipment date). It has also been clarified that there are different requirements in

**Fig. 13** Development system



**Fig. 14** System top screen

the amounts consumers purchase on weekdays, weekends/holidays, and in different time periods. Consumers purchasing products on weekdays visited the store often, so they only purchased what they would use that day, while consumers visiting the store on the weekends/holidays were often shopping for the week on the weekend, so they tended to purchase a larger quantity of products. Furthermore, people who lived in households with two or fewer people tended to purchase items during the day on weekdays, while people living in households with three or more people often shopped in the evenings, and they tended to purchase in larger volume or larger sized items. Responding to these consumer needs may further improve the perception of local vegetables and provide a satisfying shopping experience.

Feedback on the experiment results was sent to the producers by post. Feedback included 'Viewing the results of even trivial matters, such as numerical values, allowed me to understand outcomes that are normally not visible and that I was unaware of', and 'Processes that I conventionally undertook based on experience and intuition have been supported by the data, which conversely made me realise that I need to change my awareness'.

## 6 Conclusion

The results of this questionnaire clarified that producers sought information on prices before sale of the product rather than information on store inventory. While the inventory information is useful, many producers are unable to make further shipments even if the inventory becomes low during the day, and it seems that producers are unable to change their behaviour. In this study, we considered ways of presenting the price information required by producers, but the applicable stores used unique barcode numbers, which made it difficult to introduce the data with a general system. We also found that producers who proactively utilised this information throughout these initiatives accounted for around 0.5% of all producers, and only around 2% of producers expressed interest in utilising the information. Increasing the number of producers who are interested in utilising digital technology poses a major hurdle.

It is also necessary to improve the mechanism of digitising experience and intuition to create electronic data that are usable information. Furthermore, both parties indicated dissatisfaction and doubts about various aspects, and there was an apparent need for dialogue (communication). It is essential to establish agriculture as a promising and engaging occupation in which young producers can work with hope. This is another reason it is necessary to establish stable agriculture management practices and improve the income of producers by stimulating the demand for local business operators in the city and cooperating with them.

Using a web-based questionnaire system in the stores was an efficient way to collect a large volume of opinions. This initiative suggests the necessity of producing and selling products to suit consumer needs, taking heed of consumers' input, as was implemented in this instance. In this study, the researchers took the lead in registering questions and collecting information on content requested by the producers. However, questions remain about whether producers will utilise the systems themselves to conduct surveys on desired information about consumers, or whether they will simply take independent action.

It was found that a sense of 'value for money' had the greatest effect on the purchasing behaviour of consumers. Therefore, it is necessary for producers to set appropriate prices and clearly demonstrate the rationale for the pricing to consumers. Additionally, stores need to have a mechanism for widely disseminating price information.

Through this study, it has become possible to clarify the status of various situations using information technology, thus also identifying new challenges. Clearly, utilisation of DX in agriculture is indispensable, but introducing systems is only one means of achieving this goal. First, it is vital to design the region, and directly communicate and share the desired goal with all involved parties.

## References

1. Ministry of Agriculture, Forestry and Fisheries (2021) Compilation of the agricultural DX concept, 25 Mar 2021. <https://www.maff.go.jp/j/press/kanbo/joho/210325.html>. Last accessed 15 Sept 2022
2. Nihon Unisys, Tsunagaru Farmers (2022) A sales support service for direct sales outlets. <https://pr.biprogy.com/solution/biz/farmers>. Last accessed 15 Sept 2022
3. Ministry of Agriculture, Forestry and Fisheries (2021) Promotion of local production for local consumption, Apr 2021. [https://www.maff.go.jp/j/shokusan/gizyutu/tisan\\_tisyo/attach/pdf/index-20.pdf](https://www.maff.go.jp/j/shokusan/gizyutu/tisan_tisyo/attach/pdf/index-20.pdf). Last accessed 15 Sept 2022
4. Hosoya G (2019) Neuromarketing to decipher consumers' 'unconsciousness'. MarkeZine, 25 Apr 2019. <https://markezine.jp/article/detail/30888>. Last accessed 15 Sept 2022

# A Data Analysis of Video Game Reviews on Steam



Shuyao Cai , Sunyi Zhang , Lin Zhu , and Yanxia Jia 

**Abstract** The video game industry is currently one of the largest in the global market. Video games have played an important role and made a significant impact on many people's lives, especially the youth. In this paper, we perform an analysis of video games based on players' reviews on Steam, a very popular video gaming platform. We classify the video games in the dataset into five categories based on Steam's classification methods: shooter/First-Person Shooter (FPS), strategy, Role-Playing Games (RPG), sandbox/open world, and sports. We conducted exploratory analysis, sentiment analysis, and topic modeling of the game review data at the individual-game level as well as game category level. Our analysis shows that shooter/FPS and strategy games are the two most popular game categories. Through sentimental analysis, we found that the strategy and shooter/FPS categories show slightly more positive sentiments, and the sports category shows slightly more negative sentiments. Our exploratory analysis shows that players tend to spend significantly more hours in shooter/FPS and strategy games, and these two categories of games contain a substantially higher percentage of outliers, who spend extraordinarily long hours gaming. We also found that the outliers had fewer friends despite their extremely long playing hours. Topic modeling for these two categories reveals some interesting discoveries, such as players' enthusiasm about and the time they spent on the games, as well as their views about the game community and the impact of the games on their lives.

**Keywords** Big data · Natural language processing · Text mining · Topic modeling

---

S. Cai · S. Zhang (✉) · L. Zhu · Y. Jia  
Arcadia University, Glenside, PA 19038, USA  
e-mail: [szhang\\_02@arcadia.edu](mailto:szhang_02@arcadia.edu)

S. Cai  
e-mail: [scai@arcadia.edu](mailto:scai@arcadia.edu)

L. Zhu  
e-mail: [lzhu\\_01@arcadia.edu](mailto:lzhu_01@arcadia.edu)

Y. Jia  
e-mail: [jiay@arcadia.edu](mailto:jiay@arcadia.edu)

## 1 Introduction

The video game industry is currently one of the largest in the global market, with over two billion users worldwide. The video game industry has consistently grown since at least 2015 and expanded 26% from 2019 to 2021, to a record \$191 billion [1]. Video game playing is an extremely popular leisure-time activity, especially among children and adolescents. There has been various work [2–4] addressing the impact of video games. In this paper, we perform an analysis of video games based on players' reviews on Steam, a very popular video gaming platform. In particular, we conduct sentiment analysis of game reviewers and topic modeling of their reviews. Using exploratory data analysis, we analyze outliers, the reviewers who played video games for extremely long hours. Last but not the least, for each of the five game categories, we conduct topic modeling by using the LDA algorithm and pyLDAvis for visualization to gain further insights about each game category.

## 2 Data

The dataset [5] used in this paper was published on October 26, 2015. It includes 79,437 game reviews of eleven different video games from the Steam website, i.e., Arma, Counter Strike, Football Manager, Counter Strike Global, Dota, Civilization, Grand Theft Auto V, Warframe, The Elder Scrolls V, Garry's Mod and Team Fortress. This dataset contains 27 columns, including review text, total game hours, number of friends, number of games owned, and rating. We want to know whether there are differences between various types of games in terms of players' perceptions or comments. For data preprocessing, we remove stop words, expand abbreviations, and perform stemming and lemmatization.

## 3 Research Questions and Methodology

We are interested in the differences between various types of games; therefore, we divide these eleven games into five categories based on the categories on the Steam website, namely shooter/First-Person Shooter (FPS), strategy, sandbox/open worlds, Role-Playing Games (RPG) and sports. Table 1 shows the specific categories.

The research questions that we are interested in answering and the methods we use are described as follows:

**Question 1:** What are the sentiments of the reviewers/players for each game category? To address this, we conduct sentiment analysis, which can help us decipher the mood and emotions embedded in the text.

**Table 1** Five categories of eleven games

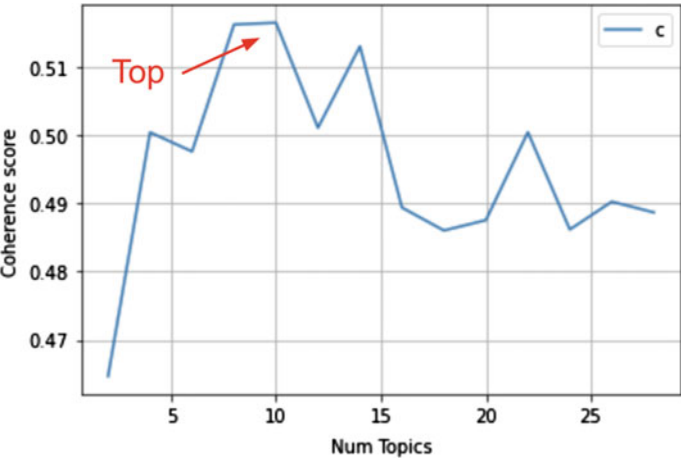
Shooter/FPS	Strategy	Sandbox/Open worlds	RPG	Sports
Arma, Counter_Strike, Counter_Strike Team, Fortress 2	Dota 2, Sid Meiers Civilization 5	Gary’s Mod, Grant Theft Auto V	The Elder Scrolls V, Warframe	Football Manager

**Question 2:** We are interested in the players who played unusually long hours (i.e., the outliers). Specifically, we apply exploratory analysis to address the following questions:

- (1) How are these outliers distributed across the five game categories?
- (2) What is the relationship between the total game hours and other attributes of players, such as the number of friends?

**Question 3:** What are the main topics of the reviews for each game category?

To discover the topics, we use the Latent Dirichlet Allocation (LDA) [6] algorithm. In order to find out the optimal number of topics, we use the UCI coherence score [7] as the performance metric. At last, we use pyLDAvis [8] to dynamically visualize the topic modeling results for a more intuitive understanding of the results (Fig. 1).



**Fig. 1** Determining the optimal number of topics



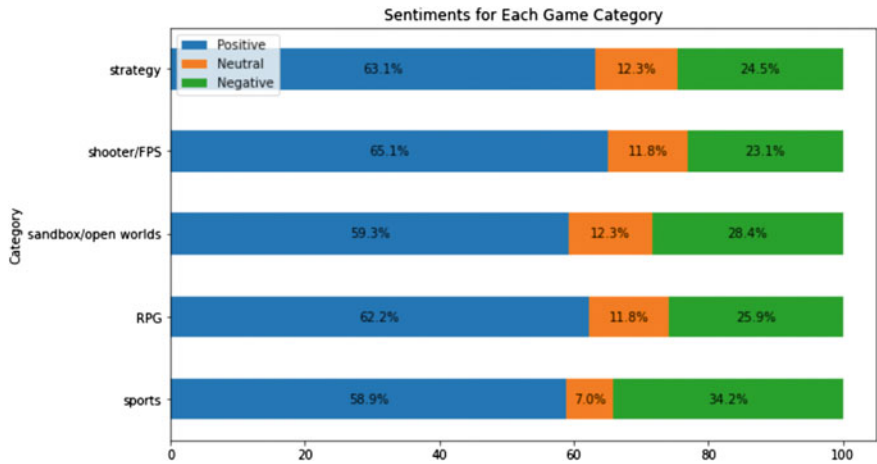


Fig. 2 Reviewer sentiments for each game category

4 Results

4.1 Sentimental Analysis

We use the TextBlob tool to analyze the sentiments for each game category. These sentiments conveyed reviewers’ positive, neutral, or negative attitudes toward the game they were reviewing.

Overall, for all five game categories, the positive sentiment dominates. We believe this is due to the popularity of all of these games. We can also observe from Fig. 2 that the strategy and shooter categories show slightly more positive sentiments.

4.2 Analysis of Total Game Hours

For each game category, we calculate the average of all the players’ total game hours respectively, as shown in Fig. 3. As we can see from the graph, on average players tend to spend significantly more hours on shooter/FPS and strategy games.

4.3 Outlier Analysis

Outliers refer to players who spend extraordinarily long hours playing games. We analyze the outliers of total game hours for each category to see which one attracts the most outliers. There are 7056 outliers in total, with the longest total game hours being



**Fig. 3** Average of all players’ total game hours

16,587.8 h. We calculate the percentage of outliers within each game category to observe how outliers are distributed in all categories. Table 2 shows some descriptive statistics of the dataset, and the boxplot in Fig. 4 shows the distribution of players’ total game hours across all game categories.

As can be observed from the bar chart in Fig. 5, two game categories—strategy and shooter/FPS, have a significantly higher percentage of outliers than the other categories. These two categories attract considerably more outliers, who spend an extraordinarily long time playing these games. This echoes the previous result that reviewers/players show relatively more positive sentiments toward these two kinds of games, and their average total gaming hours are higher.

**Table 2** Descriptive statistics of total game hours

Total game hours	
Count	79,437.000000
Mean	614.530739
Std	1008.248674
Min	0.000000
25%	65.100000
50%	218.500000
75%	761.500000
Max	16,587.800000

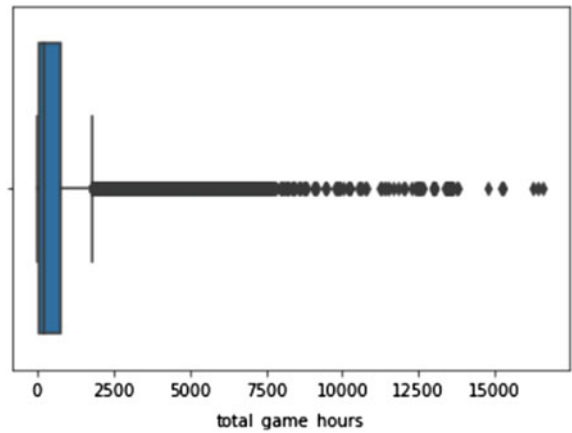


Fig. 4 Box plot of total game hours across all categories of games

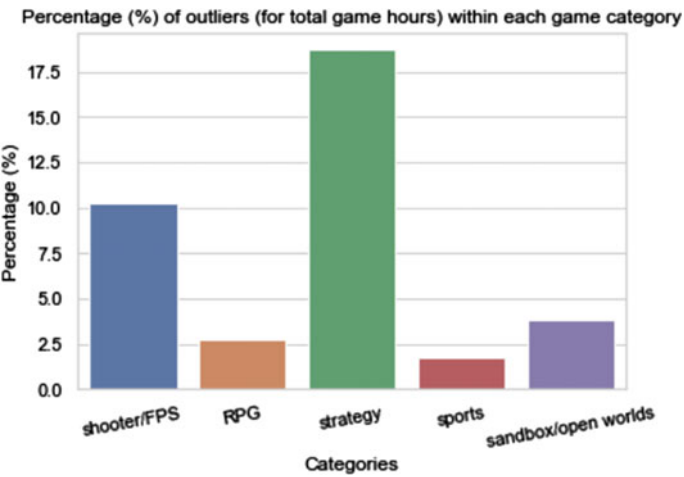
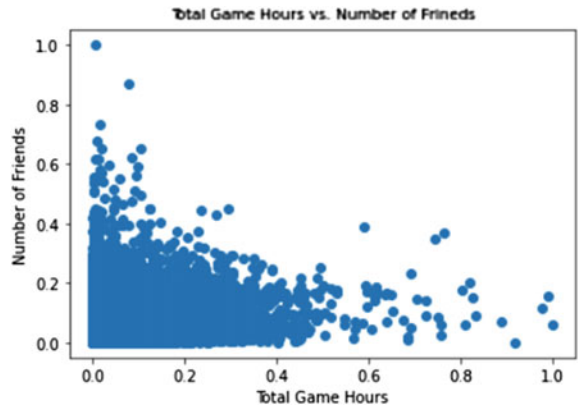


Fig. 5 Percentage of outliers (for total game hours) within each game category

4.4 Total Game Hours Versus Number of Friends

Figure 6 shows a scatter plot of reviewer/player total game hours versus the number of friends. Both the *x*- and *y*-axis are normalized values. The figure shows an interesting phenomenon that the outliers in general have fewer friends on Steam.

**Fig. 6** Reviewer/player total game hours versus the number of friends



4.5 Topic Modeling

For each of the five game categories, we conduct topic modeling by using the LDA algorithm and pyLDAvis for visualization. We also created word clouds to show the most frequent words for each topic. Figure 7 shows the top topics for the sports, sandbox/open world and RPG categories. Figure 8 shows the top topics for the shooter/FPS and strategy categories. It can be observed that the shooter/FPS and strategy categories show a more diverse set of topics than the other three categories.

We focus attention on the shooter/FPS and strategy games, given that these two categories of games demonstrate more interesting patterns, such as having longer game hours and a higher percentage of outlier players (in terms of game play time), as well as slightly higher sentiment scores, than the other three categories.

In order to better understand the meaning of the topics, we look into the reviews for occurrences of the keywords that appear in the word cloud for each topic. In doing so, we are able to see the context where the keywords appeared and therefore



**Fig. 7** Top topics of the sports, sandbox/open worlds, and RPG categories



Fig. 8 Top topics for the shooter/FPS and strategy game category

get a better sense of the meaning of the topics. We classify them into different groups according to their actual meanings.

In the shooter/FPS game category, topics 0, 4, 5, 6, and 9 in Fig. 8a are about various features of the games. For example, players can get through different terrains by vehicles and use a variety of weapons to kill enemies. In these topics, players also report game problems/bugs and even make modifications. Topics 3 and 7 show that most reviewers/players are enthusiastic to play shooter/FPS games. Topic 1 shows that Counter Strike, which is one of the most popular shooter/FPS games, sells hats related to the game, and players of Counter Strike are willing to buy them. Topic 8 shows that players spend large amounts of time playing these games happily. One interesting phenomenon worth noting is that the word “hours” is followed by a large number, as shown in the following examples (Fig. 9).

Topic 2 is about the player community. Mostly the reviewers say positive things about the community; however, negative comments do occur. Due to their relatively low frequency, these negative words are not shown in the word clouds, but the pyLDavis visualization results shown in Fig. 10 reveal that negative words, such as “hate” and “ruin” occur mostly in topic 2, the topic about the gaming community. The word “ruin” most times is followed by the word “life,” which is also among the top words in this topic.

In the strategy game category, as illustrated in Fig. 8b, topics 0, 3, 6, and 7 show various features of the games. For example, players can build their own civilization

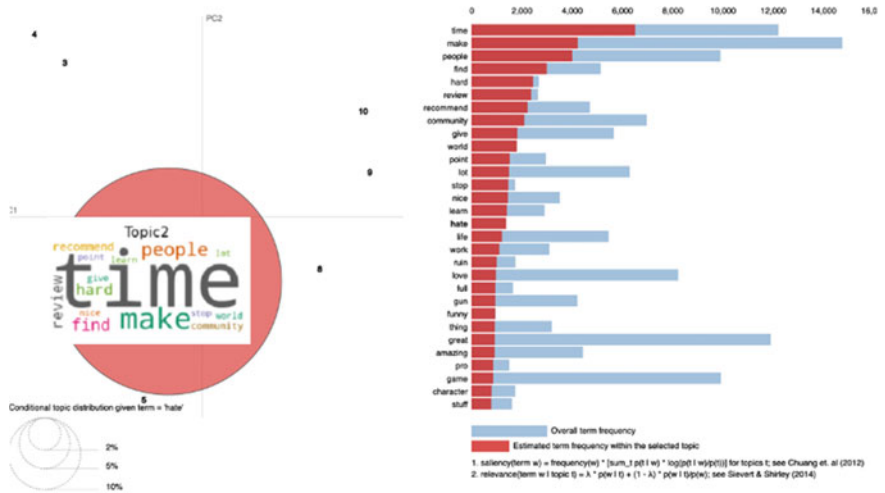
“Why then have I played the game for over 2000 hours? The sensation of winning.”

“It’s like hardcore drug use and it destroys lives. Give it a go, what’s the worst that could happen? Well, you could be here 1000 game hours later hating yourself as you queue for another all pick. Send help.”

“I have played .no I have lived this game for 2690 hours..”

“After 1752 hours of gameplay, 3 created mods.....”

Fig. 9 Word “hours” usually follows a large number



**Fig. 10** Words “hate” and “ruin” occur mostly in topic 2 of the shooter/FPS games

and make grand plans to dominate the world or become heroes and lead their team to victory.

Topic 1 shows that players find these strategy games interesting and recommendable.

Topic 2 shows that strategy games have a community for players to discuss and the impression of players is mostly positive.

Topic 4 shows that players spend much time playing strategy games happily. Interestingly, topics 2 and 4 of the strategy games are similar to topics 2 and 4 for the shooter/FPS games.

Topic 5 contains many negative words with high frequencies, such as “lose,” “ruin,” “life,” “suck,” and “waste,” which shows players’ negative emotions. The word “Russian” and “language” also appear as words of the highest frequencies, and the comments show that there are many Russians playing these games and people who would like to learn the Russian language (Fig. 11).

A careful examination of the context of the word “life” in the reviews reveals that the words “ruin” and “life” are often bounded. The word “life” frequently appears after “ruin,” “change,” and “real,” but the most frequent situation where “life” appears is “ruin my life”. Figure 12 shows some example reviews containing these phrases.

The reviewers’ comments about the strategy games are also a mixture of love with negative comments about their life-changing impact. Interestingly, it can be observed that oftentimes reviewers talked about a game ruining their life, while simultaneously they show enthusiasm about the game by either giving it a very high rating or strongly recommending it.

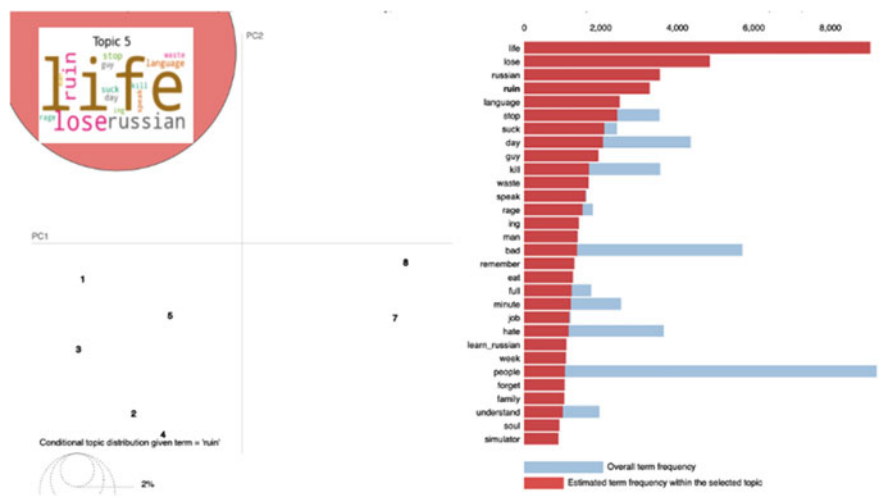


Fig. 11 Words “life,” “lose,” “Russian,” and “ruin” occur mostly in topic 5 of the strategy games

"ruined my life! 10 /10 best game DOTA2"

"After some 1,400 hours on the game, I can say with FULL VALIDITY this game has ruined my life.

"This game has ruined my social life, I spend every day on this game....."

"Successfully ruined my life. 10/10. But in all honesty this is a great game, I recommend it to all my friends....."

Fig. 12 Word “life” often appears after “ruin” in topic 5 of the strategy games

5 Conclusions

In this paper, we use game review data on the Steam website and study the differences between five game categories. We perform exploratory analysis, sentiment analysis, and topic modeling, respectively. We discover that players/reviewers of the shooter/ FPS and strategy games show slightly more positive sentiments toward these games. Players playing these two categories of games spend longer hours playing them than players of the other three game categories playing those games. These two categories also attract significantly more outlier players who spend extraordinarily long hours playing games of these categories. Topic modeling shows interesting discoveries about players’ enthusiasm about and the time they spent on the games, as well as their views about the game community and the impact of the games on their lives.

## 6 Limitations of the Study

There are limitations of our study. First, the dataset covers only eleven popular games in 2015, which is not current and also very limited. Second, the sentiment analysis tool, TextBlob, might not be able to accurately capture the sentiments, compared with some more advanced methods.

## 7 Future Work

Given the above-mentioned limits, we need to find more recent and representative datasets that include a wider range of video games. We can also use more advanced sentiment analysis models to explore the finer-grained emotions of players, for example, happy, sad, worried, etc. In addition, we can conduct hierarchical topic modeling to discover more in-depth insights.

## References

1. Video Game Industry (2022) Wikipedia, 25 June 2022. [https://en.wikipedia.org/wiki/Video\\_game\\_industry](https://en.wikipedia.org/wiki/Video_game_industry)
2. Mark G (2005) Video games and health. *BMJ* 331(7509):122–123. <https://doi.org/10.1136/bmj.331.7509.122>
3. Brilliant D, Nouchi R, Kawashima R (2019) Does video gaming have impacts on the brain: evidence from a systematic review. *Brain Sci*, 25 Sept 2019
4. von der Heiden JM, Braun B, Müller KW, Egloff B (2019) The association between video gaming and psychological functioning. *Front Psychol*, 26 July 2019
5. Mulholland M (2015) Steam review datasets. Github, 16 Nov 2015. [https://github.com/mulhod/steam\\_reviews](https://github.com/mulhod/steam_reviews)
6. Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. *J Mach Learn Res*, 3 Jan 2003
7. Röder M, Both M, Hinneburg H (2015) Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on web search and data mining*, Feb 2015
8. Sievert C, Shirley K (2014) LDAvis: a method for visualizing and interpreting topics. In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, Baltimore, Maryland, USA. Association for Computational Linguistics, pp 63–70



# Non-destructive Technique for Agricultural Seed Classification Using Deep Learning



Deepali B. Koppad, K. V. Suma, N. Nethra, and C. S. Sonali

**Abstract** Good quality seeds result in good quality yield which in turn benefits the farmer. Identifying Good Seeds is generally a manual process which is time-consuming and also prone to errors. Use of computer vision and artificial intelligence can reduce these issues. This paper performs a comparison of various deep learning models for classifying seed X-ray images into good and bad classes. Models such as AlexNet, ResNet, MobileNet and DenseNet are considered along with their variants. Performance metrics considered are classification accuracy, loss and training time. Among the 15 models that were considered in this work, ResNet50 gave the best results with a training accuracy of 98.8%, testing accuracy of 98.8%, training loss of 0.046, validation loss of 0.057 and training time of 15.85 s. This automation of seed sorting would assist in reducing the human errors as well as speed up the process.

**Keywords** X-ray imaging · USB digital microscope · Deep learning · AlexNet · ResNet · MobileNet · DenseNet · Non-destructive testing · Soya seeds

---

D. B. Koppad (✉) · K. V. Suma · C. S. Sonali  
Department of Electronics and Communication, Ramaiah Institute of Technology,  
Bengaluru 560054, India  
e-mail: [deepali.koppad@msrit.edu](mailto:deepali.koppad@msrit.edu)

K. V. Suma  
e-mail: [sumakv@msrit.edu](mailto:sumakv@msrit.edu)

C. S. Sonali  
e-mail: [1ms19ec029@msrit.edu](mailto:1ms19ec029@msrit.edu)

N. Nethra  
National Seed Project, University of Agricultural Sciences, Bengaluru 560065, India

## 1 Introduction

Increasing productivity and upgrading plantation systems are key to accelerating agricultural development. One way to improve the productivity of the crops is by sowing good quality seeds that are likely to result in high quality yield further resulting in better quality seeds. In order to identify the good quality seeds, some amount of segregation is necessary. Traditional seed segregation is done by analysing physical and chemical properties manually with a human panel, which could be time consuming, inaccurate and expensive [1]. Hence, there is a need for reliable, repeatable, economical automation of the process. Computer vision and artificial intelligence (AI) have been used extensively for this purpose [2].

In this study, the implementation of an AI-enabled seed separation system capable of identifying a healthy seed (Good Seed) and an unhealthy seed (Bad Seed) has been carried out. Specifically, convolutional neural network models such as AlexNet, ResNet, MobileNet, DenseNet and their variants are applied and compared. The best model was used to infer some test images. A non-destructive technique of seed imaging namely X-ray has been used [3].

Section 2 discusses the literature survey of related research and work available. Section 3 describes the methodology employed. Section 4 provides the discussion on the results. Section 5 concludes the paper.

## 2 Literature Survey

Patil and Totad [4] described a non-invasive soybean seed analysis which was performed using ML classifiers. Various classifiers such as convolutional neural networks (CNN), Support Vector Machines (SVM) and  $K$  Nearest Neighbours (KNN) were used, which provided an average accuracy of 66.17%.

Benjamaporn and Pornpanomchai [5] described image processing-based methods, which included steps such as image acquisition, image pre-processing, feature extraction and image recognition. Image recognition was implemented by calculating the Euclidean distance to measure the distance between every feature of a sample data set and training data set. The system is claimed to achieve a precision of 95.10% for a matching in a training data set.

Nikam and Kakatkar [6] focus on seed property measurement. Specifically, image analysis is used to evaluate parameters such as seed volume, surface area and sphericity. This could be essentially used to characterise seed samples.

In order to establish a comparison between the CNN models, Zhang et al. [7] have identified that dense summation from the aggregation output provides superior performance to that from a convolutional block output and concluded that DenseNet has produced superior results than ResNet.

Hiremath et al. [8] have used the hardware that they have designed to procure the seed images from top view as well as bottom view and feed these images to a CNN

to classify the seeds as good or bad. Accuracy of 93% is achieved by using only the hardware set-up, but it improved to 96.8% when CNN was used in addition to hardware.

### 3 Methodology

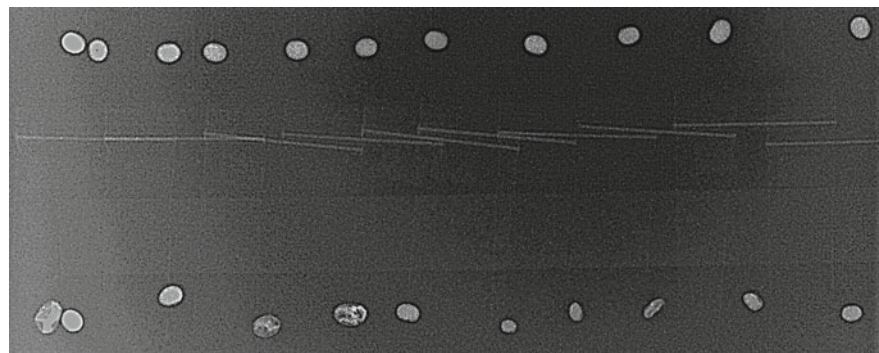
Figure 1 shows the block diagram of the approach used for seed classification. The first step in the process involves obtaining X-ray images of Good and Bad Seeds. These images are pre-processed before using them for training the deep learning models. The last step involves identifying the seeds as good/bad.

#### 3.1 Data Set

The X-ray images were captured using the Amadeo P-100/35HB X-ray unit. The images of the seeds were in JPEG format consisting of multiple seeds in a single image. Each image consisted of 22 seeds, 11 good and 11 bad, with the top row consisting of the Good Seeds and the bottom row consisting of the Bad Seeds. Example of such an X-ray image is shown in Fig. 2.



**Fig. 1** Block diagram of the methodology



**Fig. 2** X-ray image of soya seeds

The seeds were first manually cropped out to obtain images consisting of single seeds, image size  $60 \times 60$  (greyscale). The cropped images were then augmented to increase the number of images in the data set, by flipping the images along the  $x$ -axis,  $y$ -axis and sharpening the images. The images then underwent threshold-based segmentation and median filtering. The images were split into train (240 Bad Seeds and 240 Good Seeds), validation (60 Bad Seeds and 60 Good Seeds) and test (74 Bad Seeds and 90 Good Seeds) sets.

### 3.2 CNN Models

Several CNN models from the Apache MXNet framework were trained to perform a comparative study. MXNet is a framework for deep learning (DL) model used for training and deployment. It is an open-source framework, and it is scalable.

AlexNet [9] is a CNN model having a total of eight layers which consist of five convolution layers, two fully connected layers and one output layer. The input image for the model is of size  $256 \times 256$  pixels. It consists of convolution, max pooling and dense layers.

ResNet [10] is a popular and commonly used DL network for identification of images. The model depth of a ResNet is to be diligently selected because the accuracy of the model prediction is likely to be reduced as the depth increases beyond a certain value. The ResNet model has several variants to it such as ResNet18, 34, 50 and 101. This network solves the problem of vanishing gradient. MobileNet [11] is a CNN algorithm in which reduction of design complexity is prioritised. Convolution kernels in MobileNet are such that the cost calculation is less complex, which will hasten the processing with a trade-off between speed and accuracy. DenseNet [12] is derived from ResNet, and each layer receives data from all previous layers to obtain better accuracy with increasing layers. It utilises dense connections between layers through Dense Block. All these models have different variants; some of which have been evaluated in this work.

### 3.3 Classification

The last phase of the methodology involved classifying the seeds as Good and Bad Seeds. Out of the total image data set, 65% of the images were used for training the models, 15% for validation, and the remaining 20% for testing. There was no overlap of the seed images for the three categories. The output of the CNN models is the two classes: Good and Bad.

## 4 Results and Discussions

### 4.1 Results

Different deep learning models based on convolution neural network were trained and tested for X-ray images of soya seeds. To start with, the batch size in the models was varied from 1 to 4 to check the classification accuracy of the models. Increasing the batch size to 4 showed an improvement in the training accuracy as compared to the model with batch size as 1. To further improve the results, in-training augmentation was introduced. The training and validation data had random left-to-right flips; colour jitter and lighting were introduced in every epoch, effectively helping the model generalise better. This change showed a significant improvement in the training. After introducing the in-training augmentation, the models were trained for 30 epochs with a learning rate of 0.001 and a weight decay of 0.0001. For all models, stochastic gradient descent was the optimizer and soft-max cross entropy was the loss function used.

The models were trained on NVIDIA Tesla V100 12 GB GPU servers, and results obtained were compared. The training and validation results for all the models are given in Table 1.

The AlexNet model was trained in 3.5 s on the GPU server and resulted in a training accuracy of 78.6%. ResNet models with 18 and 50 layers were trained in 7.19 s and 15.84 s, respectively. The accuracy of these models is 94.2% and 98.8%. The DenseNet models 161 and 169 took around 9 s for training and provided accuracy of 83.1% and 80.7%, respectively. MobileNet models 0.5 and 1.0 were trained in 3.9 and 5.1 s to obtain a training accuracy of 92.3% and 98.5%.

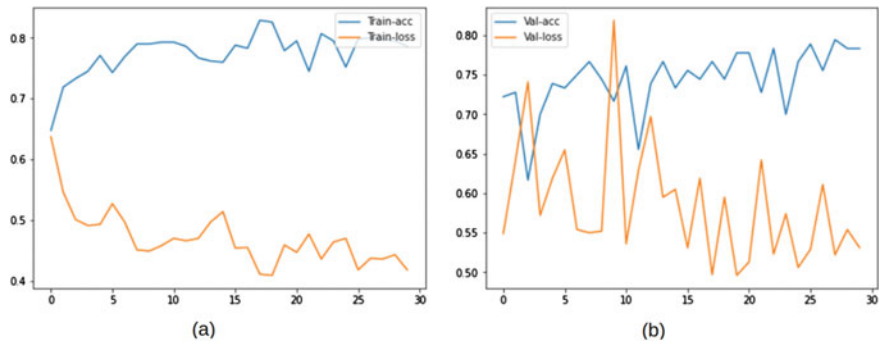
MobileNet 1.0 and ResNet 50 have approximately 98% training accuracy, though MobileNet 1.0 is trained in one third the time of the ResNet50 model.

Figure 3 shows the results for the AlexNet model. Figure 3a depicts the training accuracy and loss. Figure 3b shows the validation accuracy and loss.

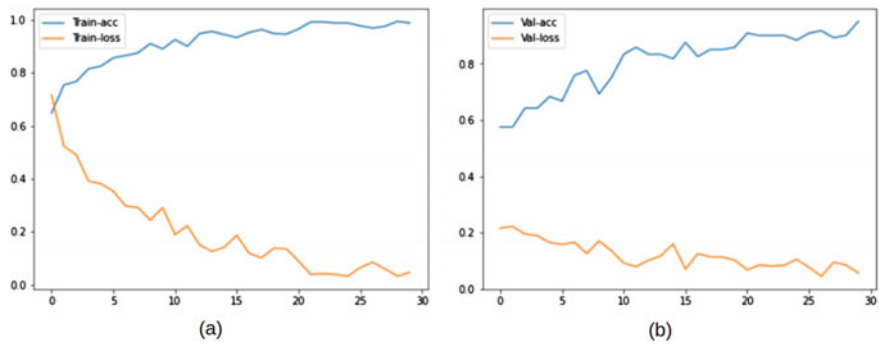
Figure 4 shows the graphical representation of accuracy and loss for the ResNet50 model during training and validation.

**Table 1** Model training and validation results

Model	Training time (s)	Training accuracy (%)	Training loss	Validation accuracy (%)	Validation loss
AlexNet	3.5	78.6	0.418	78.3	0.531
ResNet18	7.19	94.2	0.166	77.5	0.163
ResNet50	15.84	98.8	0.046	95.0	0.057
DenseNet161	9.5	83.1	0.390	80.0	0.447
DenseNet169	8.4	80.7	0.432	75.0	0.524
MobileNet0.5	3.9	92.3	0.160	75.8	0.132
MobileNet1.0	5.1	98.5	0.073	85.0	0.103

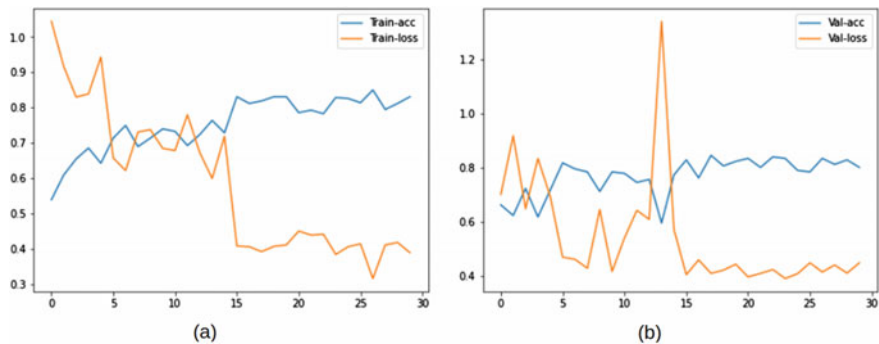


**Fig. 3** AlexNet—**a** training accuracy and loss and **b** validation accuracy and loss

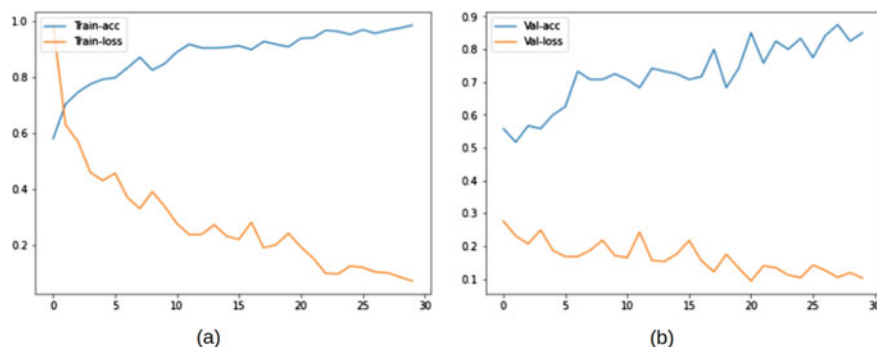


**Fig. 4** ResNet50—**a** training accuracy and loss and **b** validation accuracy and loss

Figures 5 and 6 depict the graphical representation of the training accuracy, training loss, validation accuracy and validation loss for DenseNet161 and Mobile1.0, respectively.



**Fig. 5** DenseNet161—**a** training accuracy and loss and **b** validation accuracy and loss



**Fig. 6** MobileNet1.0—**a** training accuracy and loss and **b** validation accuracy and loss

**Table 2** Classification testing results

Model	Testing accuracy (%)
AlexNet	79.4
ResNet18	97.0
ResNet50	98.8
DenseNet161	88.8
DenseNet169	82.2
MobileNet0.5	76.8
MobileNet1.0	84.8

Table 2 depicts the results for the testing data set images. These images are not used during the training and validation of the models.

Table 2 depicts the accuracy results during testing the various models. As given in the table, the ResNet50 model has the highest accuracy both for training (98.8%) and testing (98.8%), whereas MobileNet0.5 has the least testing accuracy of 76.8%. AlexNet has the least training accuracy of 78.6% and compared to the other models, also has a lower testing accuracy. For the ResNet, DenseNet and MobileNet, multiple variants were trained and results were obtained for training, testing and validation. Variants such as ResNet34, 101, 152, DenseNet 121, 201, MobileNet 0.25 and 0.75 were all considered in this work. Tables 1 and 2 depict the best results among the respective variants.

## 4.2 Discussion

This paper assesses the performance evaluation of various deep learning classification models and their variations for X-ray images of seeds. A similar work was also carried out by the authors for classification of seeds using digital microscope [13]. The soya

seed set used in both the studies was the same, and the imaging technique differed. In [13], CNN models such as ResNet, MobileNet and DenseNet were trained and tested. Parameters such as training and testing accuracy, along with the training times, were observed. The training accuracy for the three models was between 89 and 96%, and the training time was between 20 and 26 s. The testing accuracy in [13] was between 95 and 97%. For the same three CNN models, the current work has a training accuracy between 83 and 98%, training time between 5 and 15 s and a testing accuracy between 84 and 98%. With the use of X-ray images, the model trained faster and also have a marginally higher accuracy. It must be noted that in both the studies, only the external structure of the seeds was considered. The use of X-ray images in future will be to assess the internal structure and understand germination, integrity of the internal tissue and vigour. In [14], a study was carried out for crambe seeds using various CNN-based deep learning models. Seeds of different physiological characteristics were used to train and test the models, and accuracies of 91, 95 and 82% were achieved.

## 5 Conclusion and Future Work

This paper proposes a classification method for soya seed using a non-destructive technique. The X-ray image data set that is obtained is augmented to enhance the classification accuracy since it was a limited image data set. Different deep learning models are used for classification, and the results of the same are presented. A total of 15 deep learning models were trained and tested. Performance of these models was compared in terms of accuracy, loss and model training time. ResNet50 performed best among the 15 models as given in Table 1. Procuring more seed images and developing a large data set without augmentation may improve the analysis. This work can be further extended to other agricultural seeds such as sunflower and kidney bean seeds. Analysis of the internal structure of seeds using the X-ray images can be used for grading of the seeds.

## References

1. Colley M, Stone A, Brewer L (2009) Organic seed processing: threshing, cleaning and storage. Organic Seed Resource Guide, Jan 2009. <https://eorganic.org/node/392>
2. Long W, Jin S, Lu Y, Jia L, Xu H, Jiang L (2021) A review of artificial intelligence methods for seed quality inspection based on spectral imaging and analysis. J Phys Conf Ser 1769:012013. <https://doi.org/10.1088/1742-6596/1769/1/012013>
3. Bruggink H, van Duijn B (2017) X-ray based seed analysis. Seed Test Int 45–50
4. Patil A, Totad SG (2019) Non-invasive soya bean seed analysis using machine learning. Int J Recent Technol Eng 7(5S2). ISSN: 2277-38778
5. Lurstwut B, Pornpanomchai C (2011) Plant seed image recognition system (PSIRS). Int J Eng Technol 3:600–605



6. Nikam SV, Kakatkar MN (2013) Seed property measurement with image analysis. *Int J Sci Eng Res* 4(7)
7. Zhang C et al (2021) ResNet or DenseNet introducing dense shortcuts to ResNet. In: 2021 IEEE Winter conference on applications of computer vision (WACV), pp 3549–3558. <https://doi.org/10.1109/WACV48630.2021.00359>
8. Hiremath SK, Suresh S, Kale S, Ranjana R, Suma KV, Nethra N (2019) Seed segregation using deep learning. In: 2019 Grace Hopper celebration India (GHCI), pp 1–4
9. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, vol 25
10. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. *Comput Vis Pattern Recogn*
11. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861)
12. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4700–4708
13. Suma KV, Koppad D, Awasthi K, Phani Abhiram K, Vikas R (2022) Application of AI models in agriculture. Accepted at International conference on circuits, control, communication and computing (I4C) 21st–23rd Dec 2022, Bengaluru
14. de Medeiros AD, Bernardes RC, da Silva LJ, de Freitas BAL, dos Santos Dias DCF, da Silva CB (2021) Deep learning-based approach using X-ray images for classifying *Crambe abyssinica* seed quality. *Ind Crops Prod* 164:113378

# A Hybrid Strategy for DoS Attacks Detection and Mitigation on SDN Enabled Real Scenarios



Jaime Vergara, Christian Garzón, and Juan Felipe Botero

**Abstract** Software Defined Networking (SDN) enabled the possibility of programming the desired behavior inside network devices, allowing control (logic) and data (forwarding) planes to evolve at a different rate. As a result, SDN has risen rapidly as an enabler of new technological paradigms. However, the research community is still working on defining a secure SDN architecture since the centralization of the control plane potentially adds a new point of failure. Also, there has always been an interest in using OpenFlow (OF) capabilities to detect and mitigate threats against devices that are part of SDN architectures. This work proposes a hybrid strategy to protect controllers and hosts against denial of service (DoS) attacks, combining of tools and machine learning (ML) algorithms inside a real testbed.

**Keywords** SDN · Security · Attacks · DoS · OpenFlow · Machine learning

## 1 Introduction

During the past decade, an emerging paradigm known as Software Defined Networking (SDN) introduced the possibility of programming the network, thus allowing the definition of the desired functionality in different devices. Even though SDN began as an academic initiative [1], the industry has played an essential role in the adoption and development of SDN since many vendors of commercial network equipment include support for SDN enabling protocols. Also, corporations like Google, Facebook, Yahoo, Microsoft, Verizon, and Deutsche Telecom financed the Open Networking Foundation (ONF) [2] to promote the evolution and growth of SDN through open standards.

The increasing interest in SDN, where control and data planes are allowed to evolve separately, led to the developing of new network features. Threat detection and mitigation applications can be included in this group of new features. In this case,

---

J. Vergara (✉) · C. Garzón · J. F. Botero  
Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia  
e-mail: [jalberto.vergara@udea.edu.co](mailto:jalberto.vergara@udea.edu.co)

network application programmers can take advantage of southbound API capabilities (e.g., OpenFlow protocol) to identify and block cyber-attacks [3].

Even though SDN evolves rapidly and rises as an enabler of new networking paradigms (e.g., Network Functions Virtualization), it is far from being considered secure. In recent years, research has been conducted to facilitate the implementation of an SDN secure architecture. However, the logical centralization of the control plane creates a new single point of failure that compromise the entire network and the services it provides [3].

The contribution of this paper is twofold. Firstly, we propose a hybrid alternative, including an attack detection system that protects both the controller and the hosts, using machine learning techniques and tools contained in the OpenFlow (OF) protocol. Secondly, we implement this strategy in a testbed that incorporates real devices and prove its functionality.

This article is structured as follows: The second section describes the proposed solution, specifying the architecture of the experimentation environment. Section IV shows the results of the functional tests, followed by a review of related topics on SDN security implementations in real scenarios. Finally, the results and future work are discussed.

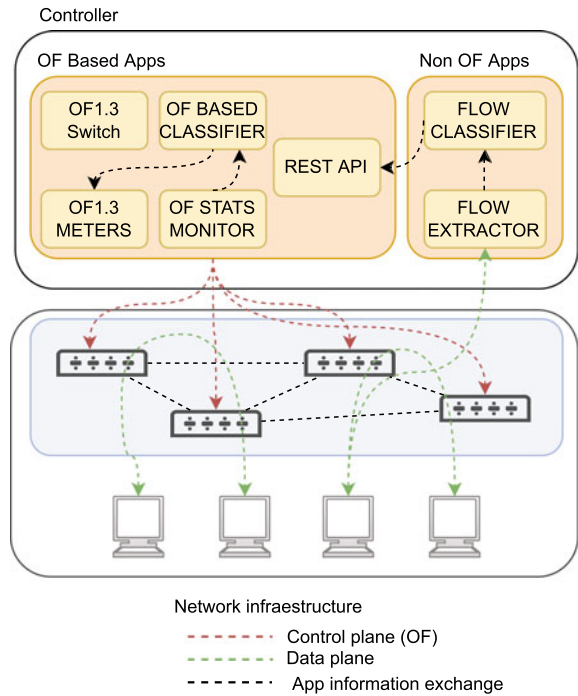
## 2 Proposed Solution

The evolution and growth of network security threats derive new challenges in designing Intrusion Detection Systems. Kaspersky, in its IT risks survey report [4], informed that both DoS and DDoS are the most common type of cyber-attacks, and in 2018, 50% of these threats lead to service disruption, and over 18 million users were affected. SDN-based architectures can leverage attack detection and mitigation of the attacks mentioned above. However, these architectures are vulnerable to DoS attacks. Therefore, we proposed a hybrid strategy, where OF statistics detect attacks between hosts, and a flow dissector is installed in the controller to protect it from DoS threats. In both scenarios, OF is used to mitigate cyber-attacks. Since OF statistics offer a limited number of features to train classification models, we consider additional protection based on a flow dissector installed on critical infrastructure points (e.g., the controller). The drawback of such strategies is the need of tracing all traffic to extract flows and calculate different features, implementing resource-demanding techniques like port mirroring. The general architecture of the solution is shown in Fig. 1. Its components play the following roles.

### 2.1 Controller Protection

This part of the proposal aims to protect the controller from DoS attacks. It comprises a flow extractor and a flow classifier that install entries in OF flow tables inside the

Fig. 1 Solution architecture



switches through an API the controller provides. All these applications run inside the controller.

**Flow classifier:** We trained a random forest (RF) algorithm with samples made of network flows. A flow is a group of packets sharing, during a period, the same values for the following 5-tuple of protocol headers: origin and destination IP addresses, origin and destination ports, and transport layer protocol. We used the Canadian Institute for Cybersecurity (CIC) intrusion detection evaluation dataset [5], using one of the sets that hold almost 7,00,000 samples of regular and DDoS traffic, where nearly 3,00,000 flows are labeled as attacks. Unfortunately, the majority of DoS samples were created using the Hulk DoS tool [6], which can lead to a bias toward the patterns exhibited by the flows extracted from this tool. These datasets have proven helpful in several works; however, most proposed algorithms are tested in offline classification tasks, where the features are already extracted from previously captured files to build train and test datasets. In our case, we added a test phase where real traffic is generated and classified online in a real scenario. We solved the two-class problem with relatively good results. Table 1 displays the confusion matrix of the trained model, showing good performance distinguishing between both classes, a recurrent result in works related to this dataset.

**Table 1** Confusion matrix

	Normal	Attack	Accuracy	F1-score	Sensitivity	Specificity	Precision
Normal	110432	30	99.9999%	99.9999%	99.9999%	99.9999%	99.9999%
Attack	24	62366					

**Flow extractor:** This part of the solution uses the tool CICFlow [7] (which is also supported by CIC) to extract per-flow features such as duration, number, and length of packets, and the number of bytes, separately in forward and backward directions.

**REST API:** We used the endpoint exposed by the controller to send HTTP requests to install flows with no action matching the IPs that are marked as malicious. Since both the flow extractor and the classifier are non-OF applications, we decided to keep them outside the Network Operating System (NOS) and use the REST API provided by the OF controller.

**2.2 Hosts Protection**

For this part of the proposal, we use OF statistics to detect DoS attacks and implement two mitigation strategies, including meters, a feature first described in OF 1.3.

**OF 1.3 Switch:** This application mimics the typical functioning of a traditional switch, where a table of port and MAC address correspondence is built to enhance packet switching. In our case, we modified this OF application to install flows in switches’ tables to match IP addresses and transport layer ports, thus having more detailed statistics for traffic classification.

**OF stats monitor:** We modify the controller stats module to periodically ask for information about each entry in flow tables on network devices. A controller can request individual and aggregate flow statistics to each switch [8].

**OF-based classifier:** After retrieving per-flow statistics, we calculate the features suggested in [9]. These statistics are calculated over all the flows retrieved over a period.

- Median packets and bytes per flow: DoS attacks could use IP spoofing, leading to the generation of flows with a small number of packets. Normal traffic usually involves a higher amount of packets. Also, DoS attacks commonly use a small packet size.
- Median duration of flow: This feature measures the mean time a flow spends in the flow table of a switch.
- Percentage of pair flows: Registers the relation between a pair flow (bidirectional) and the total amount of flows during a certain time interval.

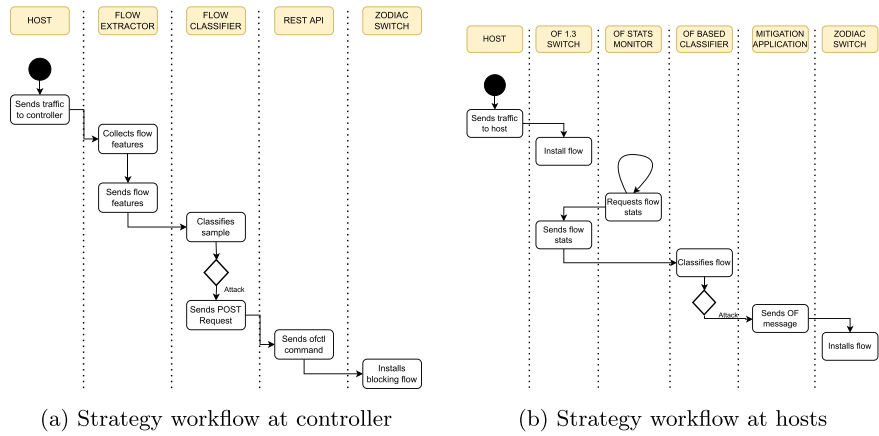


Fig. 2 Strategies workflows

- Growth of single flows and ports: At the beginning of a DoS attack, the number of flows can increase rapidly. Also, an attacker can generate flows with random source ports. Both statistics will grow when an attack is conducted.

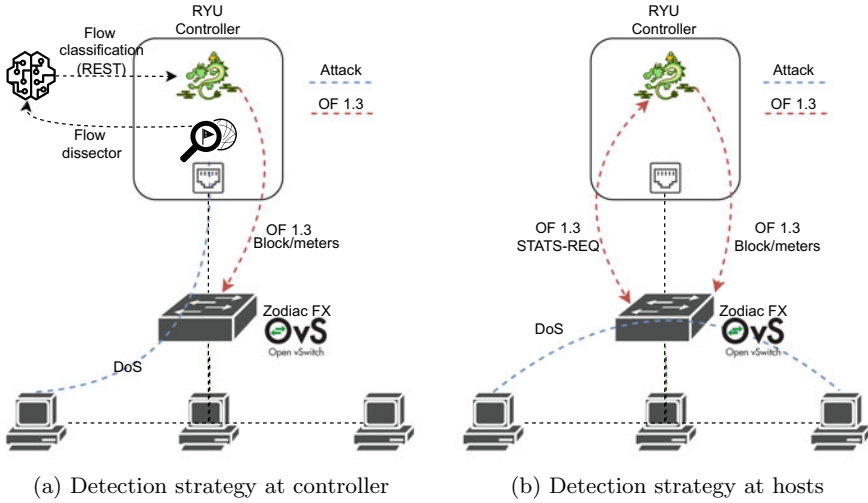
When a possible attack is detected, the classifier will analyze individual flow statistics to block the incoming traffic from the attacker based on predefined thresholds. In our case, we analyze the number of bytes and packets and the duration of each flow. We tuned the threshold values by generating regular and DoS traffic using the Hulk tool; this will hold a bias toward the patterns detected in the traffic we are using to adjust the algorithm; however, this research intends to test the usability of the strategy over a real testbed.

**Mitigation strategies:** When a potential DoS attack is detected, two different actions can be carried out: blocking the IP source of the flow or implementing an OF meter [8]. Meters are tools included in OF 1.3 and allow running simple QoS operations. In our case, we can limit the rate of a particular flow identified as hazardous.

### 3 Experimental Setup and Results

#### 3.1 Testbed

The strategy was deployed in a testbed that consists of an assemble of three OrangePi single board computers (SBC), a Zodiac FX switch that runs OpenVSwitch v.2.7, and a computer (with an Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz and 8GB of RAM) that executes the SDN-enabled RYU controller, the modified flow extractor, and classifier. One of the SBCs works as an attacker, running a Hulk DoS generator,



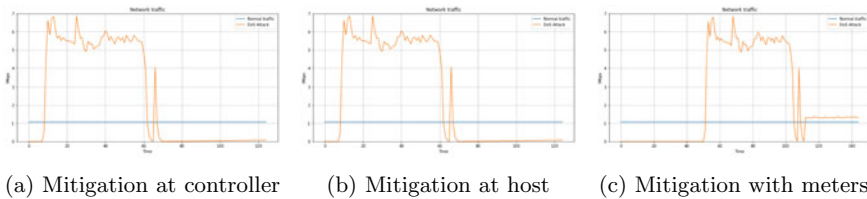
**Fig. 3** Detection strategies

while the others generate regular traffic using iperf. All the devices are connected to the same switch. Two different tests were conducted to evaluate the hybrid strategy. The connections and general workflows for both scenarios are shown in Figs. 2 and 3.

### 3.2 Attack Mitigation at the Controller

In this scenario, an SBC generates a DoS attack against the controller’s IP and the controller’s destination port to listen to OF requests. Packets that go through the controller network interface are dissected and grouped into flows by the CICFlow application (flow extractor). The features of each flow are sent to the classifier. The duration of a flow depends on TCP SYN/FIN flags or an arbitrary time in seconds. In our case, we used a maximum flow duration of 60 s since it is the default parameter configured in the feature extraction tool. Suppose the classifier identifies a flow as an attack. In that case, it sends a POST request to the controller, installs a flow matching the source IP without any action, and drops packets corresponding to that entry in the flow table. This flow is detailed in Figs. 2a and 3a.

We used the tool iPerf to generate an average flow of 1 Mbps to the controller, and at the same time, a DoS attack was sent to the same IP using the Hulk tool. After the flows are extracted and classified, the source IP of the malicious traffic is blocked in the switch, and the controller stops receiving requests from the attacker. This behavior is shown in Fig. 4a, where the bandwidth consumption drops after the entry in the switch’s table is installed.



**Fig. 4** Bandwidth monitoring

### 3.3 Attack Mitigation at Hosts

In this case, the SBC starts a DoS attack on another SBC in the topology using the HTTP port (a web server was previously instantiated at the destination). The controller periodically requests OF stats for each flow and calculates the defined thresholds. In a case where a threshold is surpassed, the controller analyzes per-flow statistics and sends an OF message to install an entry at the switch that blocks the IP from the attacker (see Figs. 2b and 3b). Again, iPerf was used to emulate normal traffic to the vulnerable host, and the DoS attack was generated with Hulk. Figure 4b shows how the attack traffic is wholly blocked, while in Fig. 4c the traffic rate of the particular flow is limited using OF Meters.

## 4 Related Work

Adding a single point of failure poses new challenges for securing SDN-enabled architectures. DoS attacks are included among these threats, where an attacker tries to wear out the available resources to disable services and affect network performance. In SDN, DoS attacks could flood both the control and data plane by filling up the memory that a switch uses for flow tables, affecting the installation of more flow rules. Also, a growing amount of packets that do not match a particular entry could end up filling up the switch's buffer and overloading the control channel with packet-in messages to the controller. Finally, a direct attack on the controller could affect all the devices, preventing the operation of network applications.

According to [10], solutions that address DoS attacks can be broadly classified between intrinsic solutions (focus on components of SDN and their functionalities) and extrinsic solutions (where network flows are analyzed through feature extraction processes). Intrinsic solutions could include strategies that, for example, deal with the limitations of flow tables inside a switch. In [11], every flow that does not match with an entry on the switch's table is stored in a temporary space in memory; if this count reaches a threshold, an alert is sent to the controller to implement a mitigation action. Some solutions look for the controller's protection by defining different processing queues served using a scheduling algorithm [12]. Additionally, among intrinsic solutions, some initiatives include new modules for SDN architecture. As



an example of this, Floodguard [13] adds a flow analyzer and a packet migrator. Both components, in conjunction, install flow rules to mitigate DoS attacks and handle flow misses in OpenFlow tables.

On the other hand, extrinsic solutions include approaches that search to protect SDN components by collecting and analyzing statistics associated with flows. For example, entropy-based attack detection is a statistical approach to protect the controller against attacks. In [14], authors proposed a distributed approach where entropy is calculated over edge switches; if entropy variations are detected among hosts, the attack source can be identified. These entropy-based alternatives are relatively lightweight approaches for gaining information about the network but typically involve threshold values that must be carefully established. There is also a growing interest in ML implementations, where the defense mechanism is an algorithm trained on labeled flows to separate benign from malicious traffic. Several papers include traditional ML algorithms, including support vector machines (SVM), random forests (RF), and also deep learning approaches like dense neural networks [10]. However, most works have focused on binary classification problems, where they try to distinguish normal traffic from anomalous traffic [15, 16]. This approach typically results in reported accuracies over 90%. Conversely, the multi-class classification problem usually poses more demanding challenges since datasets like the one used in this work are heavily imbalanced. Only a few papers addressed the imbalance of the dataset and considered metrics beyond the accuracy [17–21], which is a crucial issue to be addressed in the design of an IDS, considering the ratio of normal and abnormal traffic. An algorithm installed on an IDS could report high accuracies, but this could be biased toward the majority class, which usually corresponds to normal traffic.

The reviewed articles evaluated their methods using different tools. Most are based on emulators like mininet or even discrete event simulators like OMNet++. Fewer approaches tested their strategies over real scenarios, implementing topology-agnostic solutions that work on physical devices; this is also stated in surveys, where the hardware implementations are significantly fewer compared to the software alternatives [10]. Most strategies focus on detection, leaving open challenges in mitigation, where OpenFlow can be used as an alternative. Also, several papers concentrate on developing one kind of strategy to protect, for example, only one part of the architecture. In this case, a hybrid approach that analyzes statistics to detect patterns could use OF to implement mitigation strategies.

## 5 Conclusions and Future Work

We have presented a hybrid solution to detect and mitigate DoS attacks in SDN architectures. It consists of two functional modules, one of which serves as a protection for the controller, grouping packets in flows and extracting their features that an ML algorithm uses to detect DoS attacks. The other uses OF protocol statistics to identify DoS attacks among the hosts connected to OF-enabled switches. Both alternatives use OF to mitigate DoS attacks by blocking or limiting the attacker's traffic rate.

The hybrid alternative was tested in a real scenario and can be easily extended to more complex topologies since it periodically gathers OF statistics for each switch. On one side, the controller protection strategy can be migrated to other controllers in case of a distributed architecture. On the other side, and taking advantage of the centralized control, OF applications included in the strategy can install and monitor the flows in every switch to detect and mitigate attacks between hosts. In the future work, this will be tested on scenarios with more OF switches and emulated scenarios to prove its scalability.

We are pursuing the implementation of a better-tuned algorithm for the OF-based classifier, using different DoS attack tools to avoid biases in the detection since, as was stated before, the databases used for training are heavily built over a particular attack generation tool. Also, we expect to test the strategy in more complex topologies.

**Acknowledgements** This paper has been supported by the Ibero-American Science and Technology Program CYTED (Project: 519RT0580), and the General System of Royalties from Colombia (BPIN code 2020000100381)

## References

1. Kreutz D, Ramos FMV, Veríssimo PE, Rothenberg CE, Azodolmolky S, Uhlig S (2015) Software-defined networking: a comprehensive survey. *Proc IEEE* 103(1):14–76. . ISSN 1558-2256. <https://doi.org/10.1109/JPROC.2014.2371999>
2. Open Networking Foundation (2014) ONF member listing. <https://www.opennetworking.org/member-listing>
3. Chica JCC, Imbachi JC, Botero Vega JF (2020) Security in SDN: a comprehensive survey. *J Network Comput Appl* 159:102595. ISSN 1084-8045. <https://doi.org/10.1016/j.jnca.2020.102595>. <https://www.sciencedirect.com/science/article/pii/S1084804520300692>
4. Kaspersky Lab (2018) Denial of service: how businesses evaluate the threat OD DDoS attacks. <https://www.shorturl.at/>
5. Sharafaldin I, Lashkari AH, Ghorbani A (2018) Toward generating a new intrusion detection dataset and intrusion traffic characterization. 01:108–116. <https://doi.org/10.5220/0006639801080116>
6. Shteiman B (2017) Hulk dos tool. <https://github.com/grafov/hulk>
7. Canadian Institute for Cybersecurity (2017) Cicflow. <https://github.com/CanadianInstituteForCybersecurity/CICFlowMeter>
8. Open Networking Foundation (2012) ONF openflow specification. <https://opennetworking.org/wp-content/uploads/2014/10/openflow-spec-v1.3.0.pdf>
9. Braga R, Mota E, Passito A (2010) Lightweight DDoS flooding attack detection using NOX/OpenFlow. In: *IEEE local computer network conference*, pp 408–415. IEEE
10. Eliyan LF, Di Pietro R (2021) Dos and DDoS attacks in software defined networks: a survey of existing solutions and research challenges. *Future Gene Comput Syst* 122:149–171
11. Wang M, Zhou H, Chen J, Tong B (2015b) An approach for protecting the openflow switch from the saturation attack. In: *2015 4th national conference on electrical, electronics and computer engineering*, pp 729–734. Atlantis Press
12. Hsu S-W, Chen T-Y, Chang Y-C, Chen S-H, Chao H-C, Lin T-Y, Shih W-K (2015) Design a hash-based control mechanism in vSwitch for software-defined networking environment. In: *2015 IEEE international conference on cluster computing*, pp 498–499. IEEE

13. Wang H, Xu L, Gu G (2015a) Floodguard: a dos attack prevention extension in software-defined networks. In: 2015 45th annual IEEE/IFIP international conference on dependable systems and networks, pp 239–250. IEEE
14. Wang R, Jia Z, Ju L (2015c) An entropy-based distributed DDoS detection mechanism in software-defined networking. In: 2015 IEEE trustcom/BigDataSE/ISPA, vol 1, pp 310–317. IEEE
15. Benmessahel I, Xie K, Chellal M (2018) A new evolutionary neural networks based on intrusion detection systems using multiverse optimization. *Appl Intell* 48(8):2315–2327. ISSN 0924-669X. <https://doi.org/10.1007/s10489-017-1085-y>
16. Nawir MB, Amir A, Yaakob N, Ong B (2018) Effective and efficient network anomaly detection system using machine learning algorithm. p 12
17. Muhammad A, Qaiser R, Muhammad Z, Hasan T, Syed H, Muhammad K (2021) Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using UNSW-NB15 data-set. *EURASIP J Wirel Commun Network* 2021(01). <https://doi.org/10.1186/s13638-021-01893-8>
18. Aleesa A, Thanoun M, Mohammed A, Sahar N (2021) Deep-intrusion detection system with enhanced UNSW-NB15 dataset based on deep learning techniques. *J Eng Sci Technol* 16:711–727
19. Jing D, Chen H-B (2019) SVM based network intrusion detection for the UNSW-NB15 dataset. In: 2019 IEEE 13th international conference on ASIC (ASICON), pp 1–4. <https://doi.org/10.1109/ASICON47005.2019.8983598>
20. Kumar V, Sinha D, Das A, Pandey S, Goswami R (2020) An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset. *Cluster Comput* 23. <https://doi.org/10.1007/s10586-019-03008-x>
21. Soulaïman M, Khaldoun K, Asef J (2021) Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset. *Comput Intell Neurosci* (06):1–13. <https://doi.org/10.1155/2021/5557577>

# 1001 Games a Night—Continuous Evaluation of an Intelligent Multi-agent-Based System



Eicke Godehardt, Mohamed Amine Allani, Alexander Julian Vieth,  
and Thomas Gabel

**Abstract** The goal of the presented approach is to improve the stability of our RoboCup team code by providing an improved continuous integration software engineering process. As big and even small changes in our code base cannot be judged by just a couple of games, roughly 1000 games were run each night to have a good feeling whether changes were for the better or for worse. In addition, it is now possible to analyze the output to gain even deeper understanding of different approaches and parameters. This is supported by interactive visualization techniques. As a brute force approach will collect way too much data every night, it is necessary to condense the output and keep just a very small fraction of the detailed log data for further analysis. To decide which log files to keep different outlier detection algorithms are compared and optimized.

**Keywords** RoboCup 2D · Machine learning · Multi-agent systems · Continuous integration

## 1 Introduction

This paper describes the effort done to setup a continuous integration and development environment to judge the effects of changes as well as enable deeper analysis of the team's performance. The object of investigation is the RoboCup team FRA-UNited of the Frankfurt University of Applied Sciences [1]. This team participates in

---

E. Godehardt (✉) · M. A. Allani · A. J. Vieth · T. Gabel  
Frankfurt University of Applied Sciences, Niebelungenplatz 1, 60318 Frankfurt, Germany  
e-mail: [godehardt@fb2.fra-uas.de](mailto:godehardt@fb2.fra-uas.de)

M. A. Allani  
e-mail: [allani@stud.fra-uas.de](mailto:allani@stud.fra-uas.de)

A. J. Vieth  
e-mail: [avieth@stud.fra-uas.de](mailto:avieth@stud.fra-uas.de)

T. Gabel  
e-mail: [tgabel@fb2.fra-uas.de](mailto:tgabel@fb2.fra-uas.de)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_59](https://doi.org/10.1007/978-981-99-3091-3_59)

the 2D simulation league [2], where twelve autonomous agents (eleven players and one coach) make individual decisions every 100 milliseconds. One central server is coordinating the teams and provides the calculated update of the world back to all agents.

The FRA-UNited team has quite a long history which led to a pretty big C++ code base. This makes it very hard to make confident changes without a safety net. A continuous integration and deployment strategy can help. The main issue is that it is hard to judge the team's performance by a couple of games. It is rather necessary to run around 1000 games to really get a robust result whether the last changes were for the good or for worse.

In Sect. 2, we will discuss related work, based on which the chosen approach is explained in Sect. 3. Section 4 presents the results of an empirical evaluation of the presented approach, and Sect. 5 concludes.

## 2 Related Work

Two techniques are actually in the focus of this paper. First of all the overall development environment including the execution of a great number of simulated soccer games. On the other hand the outlier detection process as a central element inside of this environment. Here, we quickly go over related work in both areas.

At least one other team of RoboCup 2D league—team Helios—has a similar approach of running many games in an automated manner. In [3], they describe a performance evaluation system used for their team utilizing SlackBot, Amazon S3, and Google Sheets. A user can create a job with branch and opponent information as well as the number of games to a SlackBot. A server then assigns client PCs with the jobs depending on CPU load. Client PCs with high load are considered *busy* and do not get jobs assigned. Client PCs run the assigned games, analyze the resulting log files into a CSV file, and push them to a shared storage (Dropbox and Amazon S3). The resulting CSV files are aggregated into a Google Sheet, which the user can use to evaluate the performance.

The main difference is the automatic approach presented here combined with local only resources and the focus on more graphical data analysis possibilities. In addition, the authors of this paper have no information whether real log files are obtained by the other approach.

## 3 Approach

While the paper at hand provides a coarse overview of the used techniques and approaches, the interested reader is referred to [4, 5] for more details. In [4], special features of the architecture are highlighted, whereas a deeper investigation of the applied outlier detection process is described in [5].

### 3.1 Architecture

The CI system consists of six components:

- a Git server
- a Jenkins server providing team builds and version info's
- a web server (using Grails) to collect and serve match data and statistics
- an interactive React web application
- 20 test computers running shell and python scripts
- RoboCup 2D tournament software [2, 6].

The only interaction a user has with the system is mainly by pushing a new commit to the git repository and by looking at the visualized results on the React web application. As the name suggest, the CI system functions mostly autonomously, with the process of fetching the newest FRA-UNited build, playing/analyzing games and pushing summaries back to the web server is all initiated by the test computers, which is different to the approach taken in [3]. This has the advantage of the collecting server only being loosely coupled to the test computers, as the server can be oblivious to the source of incoming matches. Scaling this system up or down requires only minimum effort. To enable the over-night running of matches, the test computers are configured with a cronjob, which fetches the newest version of the CI scripts and executes a specific entry point script.

One important feature is that one can find optimal configuration values based on A-B testing. For this, the number of games can be split, e.g., in half, to see the impact of different configuration values of the agents' behavior.

### 3.2 Outlier Detection

One really important component in the whole setup is the outlier detection to minimize the number of kept log files for further investigation and deserves its own research. This is quite important as you may gain some understanding from the condensed log information, but can no longer deeper investigate and replay the actual log files to see real decisions in a game. So the main idea is to keep log files of a couple “interesting” games—or in other words anormal games. Here, outlier detection comes into play. Outlier detection is a wide field and many reviews discuss research in that area [7, 8].

In a given dataset, some samples may differ from most observations to an extent where it would be considerable that they were generated by a different process. These samples are known as *anomalies* and their characteristics are different from those of normal samples [9]. The proportion of outliers in any given dataset is referred to as *contamination*.

Under anomalies, one could differentiate between *outliers* and *novelties*. Generally, the amount of these anomalies represents a small portion of a dataset. When

their proportion is lower than 5%, these anomalies are called outliers[10]. Therefore, the process of detecting outliers is called *outlier detection*. In this context, the training data already contains outliers that are far from the other observations, also known as *inliers*. On the other hand, novelties are completely new observations; the training data does not contain any anomalies. *Novelty detection* is then the process of detecting novelties, and it occurs after the training phase.

All algorithms mentioned in this paper work according to a common principle: The model defines normal data points by identifying dense clusters containing the most similar points in the dataset. This occurs during the training phase. The next step is to select outliers, either by selecting the outliers that lie in different regions by calculating an *anomaly score* representing the “outlierness” of a data point for each sample. According to this score, outliers can be selected by either i) *block-testing* which consists in defining a threshold value based on a given degree of contamination and selecting the points that have a score higher or lower than the threshold or by ii) *consecutive testing* which consists in sorting the samples in descending/ascending order with respect to their anomaly score and with the most suspicious observations at the beginning, then evaluating each sample individually until a normal observation is reached.

The main differences between outlier detection algorithms lies in the way how each method calculates an anomaly score [10]. Under these methods, one could differentiate between four groups:

- nearest neighbor-based algorithms
- isolation algorithms
- probabilistic algorithms
- domain-based algorithms.

In this approach, we focus on the first two groups with one example from either group based on the applicability in the given scenario.

*Local Outlier Factor* Local Outlier Factor (LOF) is an outlier detection algorithm based on analyzing the direct neighborhood of each data point to determine its outlierness [11, 12]. Indeed, the density of the neighborhood is taken into account and is used to calculate a score, namely the Local Outlier Factor. A point is defined as outlier if its score is higher or lower than a certain threshold.

*Isolation Forest* The other kind of outlier detection applicable in the given scenario is Isolation Forest (IF) [13]. It is based on random forests and computes an isolation score for all instances in the dataset. The algorithm builds binary trees which have random split values on each of their nodes. The result is a “forest” of such trees which processes and assigns an isolation score to each data point. The calculated score is based on the average path length from root to leaf in every tree. The shorter the path, the more likely the point is an outlier.

**The Gold Standard** As both algorithms are instances of unsupervised learning algorithms, a way to judge the results is highly needed. In order to do so, we asked a RoboCup expert in the team to look manually for outliers in our data set. This leads to the gold standard, both algorithms can be measured by.

Since the algorithms should work on batches of 1000 samples, the labeled dataset has to contain 1000 samples. However, since labeling the data is very time-consuming, a way to reduce this number was needed. The 1000 unlabeled samples were used to train both algorithms using the default hyperparameters. As the top labeled outliers seem quite good examples of outliers, only the top 100 samples (50 first samples from every algorithm) were selected for manual labeling. The rest of the samples were assumed to be inliers.

To label the samples, the expert based his choice on the following criteria and features:

1. `goals_<team>`: Goals are the major feature used by the expert to assess a game. For most selected outliers, there is a considerably big difference of at least three between `goals_l` and `goals_r`.
2. `shots_on_target_<team>`: In some samples selected as outliers, the team FRA-UNited had 0 for shots on target. Furthermore, when there is a discrepancy, i.e., a huge gap between `shots_on_target_<team>` and `goals_<team>`, the sample was marked as outlier.
3. `possession_<team>`: A higher possession for FRA-UNited is improbable. However, a very small value for possession for FRA-UNited resulting in lowers values for other features like passes is also considered unlikely.

Considering the 50 samples that obtained the largest outlier scores by IF and LOF, 7 and 5 of them were classified as outliers by the expert, respectively. Under the assumption that none of the remaining samples were outliers, this yields a degree of contamination of 0.007 for IF and of 0.005 for LOF.

## 4 Evaluation

In this section, the hyperparameters which have the highest Fowlkes-Mallows index (FMI) score will be selected for both algorithms and compared to one another (FMI score is chosen, because of its resilience to class imbalance). In that context, the IF evaluation with the optimal hyperparameters of the previous experiment ( $t = 90$  and  $\psi = 256$ ) and the LOF algorithm with the optimal hyperparameter  $k = 150$  are selected. The one on one comparison of their metrics can be seen in Table 1.

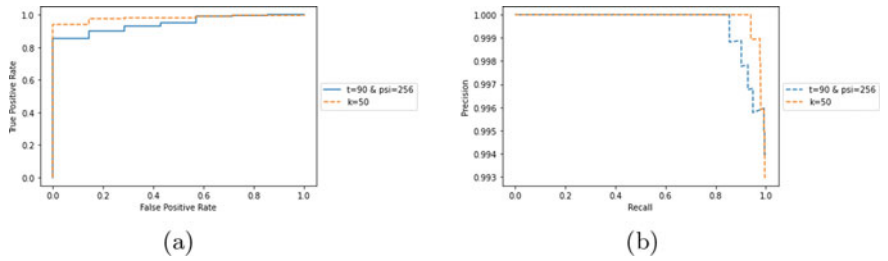
As can be seen in Table 1, although LOF has higher values for ROC-AUC and PR-AUC than IF, both algorithms have equal performance according to the other metrics (F1, AP, and FMI). This can be explained by examining their ROC and PR curves (cf. Fig. 1).

Figure 1 and Table 1 show that the AUC of LOF is higher for both, ROC and PR. Indeed, the LOF curve seems to be higher than the IF curve, in the region between

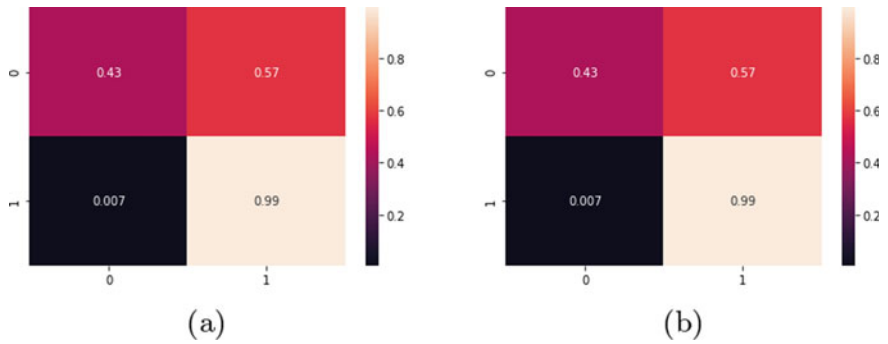


**Table 1** Results of the evaluation of the IF and LOF algorithms with their respective optimal hyperparameters (AUC: area under curve, ROC: receiver operating characteristics, PR: precision-recall, AP: average precision, F1: harmonic mean of the precision and recall, FMI: Fowlkes-Mallows index)

Hyperparameters	ROC-AUC	PR-AUC	AP	F1	FMI	Time (s)
IF: $t = 90, \psi = 256$	0.94605	0.99960	0.42857	0.99445	0.98893	0.32307
LOF: $k = 150$	<b>0.98000</b>	<b>0.99986</b>	0.42857	0.99445	0.98893	0.08500



**Fig. 1** Comparison of IF and LOF with their respective optimal parameters: **a** ROC curve, **b** PR curve



**Fig. 2** Confusion matrices: optimal parameters LOF (a) and IF (b)

0.0 and approx. 0.6 FPR in the ROC curve and approx. 0.8 and 1 in the PR curve. However, both curves overlap in the other regions. This can be explained by the process that generates these curves and the way the thresholds for these curves are chosen. LOF seems to be performing better than IF when different classification thresholds are chosen. However for the threshold corresponding to the predefined 0.01 contamination ratio, both algorithms have the same performance. This can be demonstrated by the values of the metrics F1-score, AP, and FMI and by examining the confusion matrices in Fig. 2.

Both matrices in Fig. 2 are identical for both algorithms which suggests that both algorithms perform the same when the threshold is set according to the contamination score 0.01.

Additionally, it is important to note that the detected (hand-marked) outliers are exactly the same for both algorithms; however, not all predicted outliers are the same. This has repercussions on how both algorithms will be used in the final version of the system.

## 5 Conclusion

The RoboCup 2D simulation league continues to evolve, with a steady increase in the capabilities on how the teams play soccer. As every match of 2D soccer is tainted to various degrees by randomness, a need to play a large number of matches arises, to limit the impact of chance on the overall results. For this reason, a continuous integration (CI) system has been implemented. This paper discussed various aspects of the system, which is used by the team of the Frankfurt University of Applied Sciences' RoboCup team—FRA-UNITed, with the goal to assess its performance. The main points discussed were

- Risk of overfitting, by only playing against the same team
- No interaction and influence on the CI system besides new commits
- Unused potential for configuration testing.

In this work, a new CI system was demonstrated, which allows the managing and usage of arbitrary team binaries, along with the ability to freely specify configuration values for arbitrary config files.

Evaluating outlier detection algorithms in the context of analyzing RoboCup 2D games is not trivial. Indeed, the gold standard may be defined according to subjective beliefs that assess the data. Furthermore, the patterns that are used to identify these outliers can be complicated, and there is a high possibility that more feature are needed in order to score a higher true negative rate. Therefore, the results of the evaluation can be considered satisfactory, knowing that almost half of the outliers specified by the golden standard are successfully detected.

## References

1. Gabel T, Eren Y, Sommer F, Vieth A, Godehardt E (2022) RoboCup 2022: robot soccer world cup XXVI. Springer, CD supplement, Bangkok
2. Noda I, Matsubara H, Hiraki K, Frank I (1998) Applied artificial intelligence. 12(2–3):233
3. Yamaguchi M, Kuga R, Omori H, Fukushima T, Nakashima, Akiyama H (2021) RoboCup 2021 symposium and competitions, worldwide
4. Vieth AJ (2022) A continous integration system with diversified opponents and dynamic team configurations for RoboCup 2D. Master's thesis, Frankfurt University of Applied Sciences

5. Allani MA (2021) Analysis and visualization of RoboCup games using machine learning (orig German: Analyse und Darstellung von RoboCup Spielen mit Maschinellem Lernen). Master's thesis, Frankfurt University of Applied Sciences
6. Hechenblaickner A, Hech League Manager (2004–2010)
7. Ord K (1996) Probability Judgmental Forecasting. *Int J Forecast* 12(1):175
8. Alghushairy O, Alsini R, Soule T, Ma X (2020) Big data and cognitive computing 5(1):1
9. Hawkins D (1980) Identification of outliers. In: *Monographs on applied probability and statistic*. Chapman and Hall, London [u.a.]
10. Domingues R, Filippone M, Michiardi P, Zouaoui J (2018) *Pattern recognition* 74:406
11. Mishra S, Chawla M (2019) Emerging technologies. In: Abraham A, Dutta P, Mandal JK, Bhattacharya A, Dutta S (eds) *Data mining and information security*. Springer Singapore, Singapore, pp 347–356
12. Breunig MM, Kriegel HP, Ng RT, Sander J (2000) In: *Proceedings of the 2000 ACM SIGMOD international conference on management of data*. Association for Computing Machinery, New York, NY, USA, 2000, SIGMOD '00, pp 93–104. <https://doi.org/10.1145/342009.335388>
13. Liu FT, Ting KM, Zhou ZH (2008) In: *2008 Eighth IEEE international conference on data mining*. IEEE, pp 413–422

# Real-Time Hand Action Detection and Classification Based on YOLOv7 from Egocentric Videos



Van-Hung Le

**Abstract** Hand detection and classification still exists many challenges when detecting and classifying the hands-on data received from the first-person camera/egocentric vision, because the data of the hand is obscured by the viewpoint or other objects. On this dataset, the fingers are almost completely obscured, and only the data area of the hand palm is visible. In this paper, we propose to apply YOLOv7 and its variations for fine-tuning the hand action detection, and classification model on the RGB image of egocentric vision dataset as First-Person Hand Action Benchmark (FPHAB) dataset, HOI4D dataset. We have prepared the ground-truth data of the hand action by manual operation for the hand action detection evaluation on the FPHAB and HOI4D datasets. The results are compared with different YOLO variations (YOLOv7, YOLOv7-w6, YOLOv7-X) and Mediapipe on the FPHAB and HOI4D datasets. We did a very diverse evaluation with two configurations (*Conf.* #123, *Conf.* #213) of data on the FPHAB and a configuration of the HOI4D dataset. YOLOv7 and its variants are more than 95% accurate across all evaluation configurations with an IOU threshold of 0.95.

**Keywords** Hand detection · Hand classification · YOLOv7 · FPHAB dataset · HOI4D dataset · Convolutional neural networks

## 1 Introduction

3D hand pose estimation and hand action recognition are the studies strongly applied to human-machine interaction, controlling smart home appliances [2]. The results of 3D hand pose estimation and hand action recognition are highly dependent on the data area of the hand detected in the image results. Therefore, the detection of the hand action is an important pre-processing step in building applications based on the estimation and recognition of the action of the hand. If the results detect the

---

V.-H. Le (✉)  
Tan Trao University, Tuyen Quang 2200, Vietnam  
e-mail: [van-hung.le@mica.edu.vn](mailto:van-hung.le@mica.edu.vn)

wrong hand, further studies are not possible. Another problem is this step is to limit the data area of hand posture estimation and hand activity recognition on this data area. The FPHAB [4] dataset is the dataset collected from the camera from a first-person/egocentric vision. In this dataset, the fingers are obscured by the back of the hand or the objects that the hand performs grasping. Recently, Google Mediapipe (GM) has achieved many impressive results in detecting and estimating human and hand pose, and human and hand activity recognition [1, 6]. However, when applying GM to detect the hands-on color images of the FPHAB database, there are many hands in the image that are not detected, especially the hands that perform interactive actions holding objects. The HOI4D [7] dataset is the dataset collected from the camera from a first-person/egocentric vision. In this dataset, the fingers are obscured by the back of the hand or the objects that the hand performs grasping. To carry out studies on 2D and 3D hand posture estimation, the hand on the image needs to be detected and positioned in the image. In particular, this step needs to be performed in real time and can be performed on the CPU so that its computation time does not greatly affect the processing time of both the estimation and recognition process of the hand. Recently, the YOLOv7 model was proposed by Wang et al. [12]. YOLOv7-E6 [12] is more accurate and faster than SWINL Cascade-Mask R-CNN [3] by 2% and 509%, respectively. In this paper, we propose to use YOLOv7 and its variations to fine-tune the model for hand detection and classification in the FPHAB and HOI4D datasets. The hand detection and classification model that we pre-trained is performed on the training data of the configuration #123 (*Conf.* #123), configuration #213 (*Conf.* #213) of FPHAB dataset and training set (70%) of the HOI4D dataset. We also evaluated on the testing set of the configurations of the FPHAB dataset and (30%) HOI4D dataset. The main contributions of the paper are as follows:

- We have proposed using YOLOv7 and its variants (YOLOv7, YOLOv7-w6, YOLOv7-X) for fine-tuning hand action detection and classification of the model on two first-person viewpoint datasets (FPHAB and HOI4D datasets).
- We have made hand data area marking perform the manual operation for the evaluation of the performance hand detection results on the FPHAB dataset and extracted the hand bounding box of the HOI4D dataset.
- Experiments on hand action detection and classification are presented in detail; the results of hand action detection and classification are compared with state-of-the-art methods on the FPHAB and HOI4D datasets.

## 2 YOLOv7 for Hand Action Detection and Classification

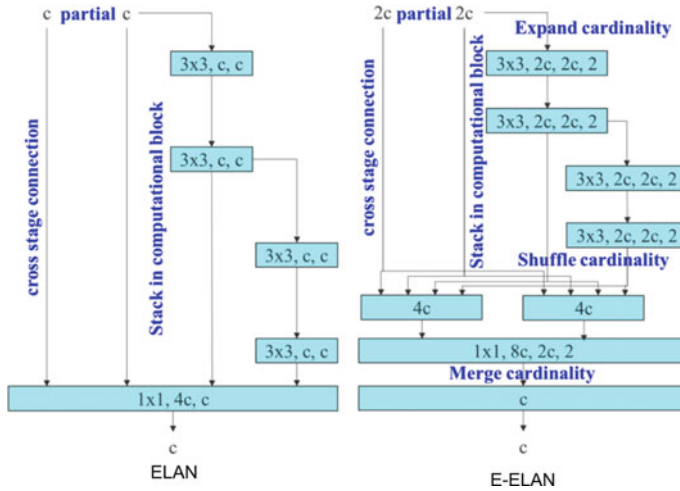
Based on the results of research on object detection in the RGB images of MS COCO 2017 database, YOLOv7 is having the best results both in accuracy and computational time. In this paper, we propose to use YOLOv7 for fine-tuning the hand action detection model of the FPHAB and HOI4D datasets. As discussed in Sect. 1, the hand action, left-hand and right-hand data of the FPHAB and HOI4D datasets contain

challenges that pre-trained models based on CNNs are undetectable. In YOLOv7, the authors made some major improvements. First, the efficient layer aggregation networks (ELAN) are proposed to expand to extended efficient layer aggregation networks (E-ELAN), where the strategy that learns at more depth with the shortest and longest derivatives along the slope will have a higher probability of convergence. This means not changing the gradient transmission path of the original architecture, but increasing the group of convolutional layers of the added features, and combining the features of different groups by mixing and merging the cardinality manner, as presented in Fig. 1. This way of working can improve the learning efficiency of learned solid maps and improve the use of parameters and calculations. This process increases the accuracy of the learned model without increasing complexity and computational resources. Second is the proposed Model Scaling for concatenation-based models (MSCM), The main idea of MSCM is based on scaled-YOLOv4 [10] to adjust the number of stages. When increasing the depth of a transition layer which is immediately after a concatenation-based computational block will increase, as illustrated in Fig. 1a, b, it means the input width of the subsequent transmission layer to increase. Therefore, the model scaling on concatenation-based models is proposed. This process only makes the depth in a computational block need to be scaled, and the remaining transmission layer is performed with corresponding width scaling, as illustrated in Fig. 1c. Third is to reduce the number of parameters and computation for object detection; YOLOv7 is used the re-parameterized to combine with a different network. This work can reduce about 40% parameters and 50% computation of object detector, and the detection will be faster and more accurate. Fourth is a new label assignment method that guides both the auxiliary head and lead head by the lead head prediction. This method used lead head prediction as guidance to generate coarse-to-fine hierarchical labels. YOLOv7 and its variants still use Darknet-53 [9] as a backbone for training models to detect the object in the images. With the advantages of YOLOv3 and YOLOv5 in object detection whose backbone is Darknet-53 and features pyramid (FPN, PAN), Ge et al. [5] proposed YOLO-X with the following two improvements: Replacing YOLO's head with a decoupled one greatly improved the converging speed. The decoupled head is essential to the end-to-end version of YOLO. In [12], the E-ELAN and MSCM methods are applied to YOLO-X, called YOLOv7-X. Another variant of YOLOv7 presented in this paper is YOLOv7-W6. The E-ELAN and MSCM methods also are applied to YOLOR-W6 [11].

## 3 Experimental Results

### 3.1 Datasets

The dataset on which we experimented is the First-Person Hand Action Benchmark (FPHAB) [4]. This dataset is captured from an Intel RealSense SR300 RGB-D camera attached to the shoulder of the subject. The resolutions of the color and



**Fig. 1** Illustration of ELAN and E-ELAN architectures of YOLOv7 [12]

depth images are  $1920 \times 1080$  pixels and  $640 \times 480$  pixels, respectively. The hand pose is captured using six magnetic sensors; it provides 3D hand pose annotation and intrinsic parameters for converting 2D hand pose annotation. There are several subjects (6 in total) performing multiple activities from 3 to 9 times 45 hand actions. In this paper, we used two configurations for training and testing, presented as follows. The first configuration (*Conf. #123*) used the 1st sequence in each subject (*Subj.*) from *Subj. #1* to *Subj. #6* for testing, the 2nd sequence in each *Subj.* for validation and the remaining sequence for training. The second configuration (*Conf. #213*) used the 2nd sequence in each *Subj.* from *Subj. #1* to *Subj. #6* for testing, the 1st sequence in each *Subj.* for validation and the remaining sequence for training.

**HOI4D** [7] is collected and synchronized based on a Kinect v2 RGB-D sensor, and an Intel RealSense D455 RGB-D sensor. This is a large-scale 4D egocentric dataset with rich annotation for category-level human-object interaction. HOI4D includes 2.4M RGB-D frames of egocentric vision with over 4000 sequences. It is collected from 9 participants interacting with 800 different object instances from 16 categories over 610 different indoor rooms. To get the ground-truth data which is the bounding boxes of the hand on the image for evaluation, we rely on the hand key points and named “kps2D” in the 3D hand pose annotation. We take the bounding box of 21 hand key point annotations. This process is demonstrated in the source code of the “*get\_2D\_boundingbox\_hand\_annotation.py*” file in link.<sup>1</sup> In this paper, we pre-trained on the YOLOv7 and its variants (YOLOv7-X, YOLOv7-w6) on the training set of the *Conf. #123* of the FPHAB dataset and evaluated it on the validation set and testing set of configurations (*Conf. #123*, *Conf. #213*). We also use 70% of the frames for training and 30% of frames for testing the hand detection and classification model of

<sup>1</sup> <https://drive.google.com/drive/folders/1yzhg5NsalPkOHI6CMkAE07yv5rY63tI7?usp=sharing>.

the HOI4D dataset. This data is split randomly and based on the proportions in [7]. We use the  $\text{Thesh}_{\text{IOU}}$  to evaluate as follows: 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95. In this paper, we used a server with an NVIDIA GeForce RTX 2080 Ti, 12GB for fine-tuning, training, and testing. The programs were performed in the Python language ( $\geq 3.7$  version) with the support of the CUDA 11.2/cuDNN 8.1.0 libraries. We also perform a Google Mediapipe (GM) evaluation for hand detection on images with a laptop: CPU 11th Gen Intel Core i5-1135G7, 2.4GHz, RAM LPDDR4X 16GB.

### 3.2 Evaluation Metrics

Similar to the evaluation of object detection and classification on images, we perform the calculation of the IOU value according to the formula (1).

$$\text{IOU} = \frac{B_g \cap B_p}{B_g \cup B_p} \quad (1)$$

where  $B_g$  is the ground-truth bounding box of hand action,  $B_p$  is the predicted bounding box of hand. To determine whether the bounding box is a true finding, we use a threshold  $\text{Thesh}_{\text{IOU}}$  for the evaluation. If IOU is greater than or equal to  $\text{Thesh}_{\text{IOU}}$  then it is a true detection otherwise it is false. In this paper, we also distinguish between the hand action and the background, so we also use the formulas for precision ( $P$ ), Recall ( $R$ ) and  $F_1$ -Score ( $F_1$ ) (Eq. 2) to evaluate the analysis results of hand action classification on the image.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}; R = \frac{\text{TP}}{\text{TP} + \text{FN}}; F_1 = \frac{2 * (R * P)}{(R + P)} \quad (2)$$

where TP is true positives, TN is true negative, FP is false positive, FN is false negative. We pre-trained each network (YOLOv7, YOLOv7-w6, YOLOv7-X) with 50 epochs and batch size = 2 frames, the size of the image can be  $\text{img\_size} = 640 \times 640$  or  $\text{img\_size} = 1280 \times 1280$ ,  $\text{conf\_thres} = 0.001$ . There are also some other parameters shown in Table 1.

### 3.3 Hand Action Detection and Classification Results

The results of hand action detection and classification on the *Conf. #123* of the FPHAB dataset are shown in Table 2. The results of hand action detection and classification on the *Conf. #213* of the FPHAB dataset are shown in Table 3.

The results of Tables 2 and 3 can be seen that the hand action detection and classification results on the FPHAB dataset are very accurate, the results are greater



**Table 1** The list of parameters of YOLOv7 and its variants [12], resulted in the processing time of the networks when evaluated on the testing set of the FPHAB database

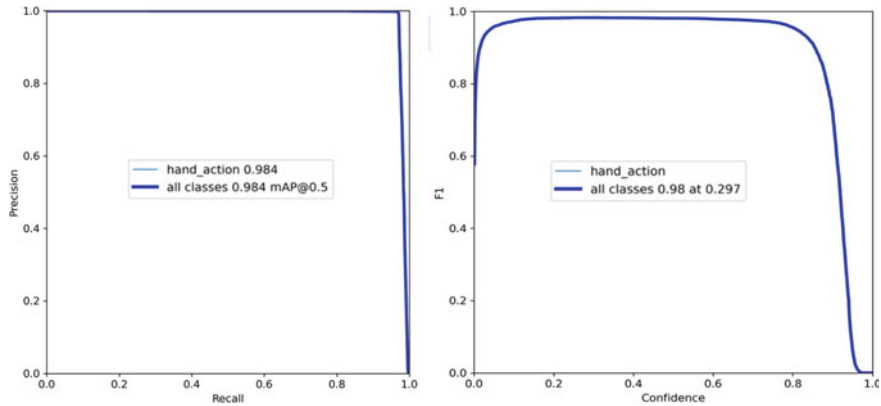
Method	Image size (pixel)	Number of layers	Number of GFLOPS	Parameters	Number of epoch	Processing time for testing (fps)
YOLOv7	640 × 640	314	103.2	36481772	50	133
YOLOv7-X	640 × 640	362	188.0	70782444	50	98
YOLOv7-w6	1280 × 1280	370	101.8	80909336	50	60

**Table 2** The results of hand action detection and classification on the *Conf.* #123 of the FPHAB dataset

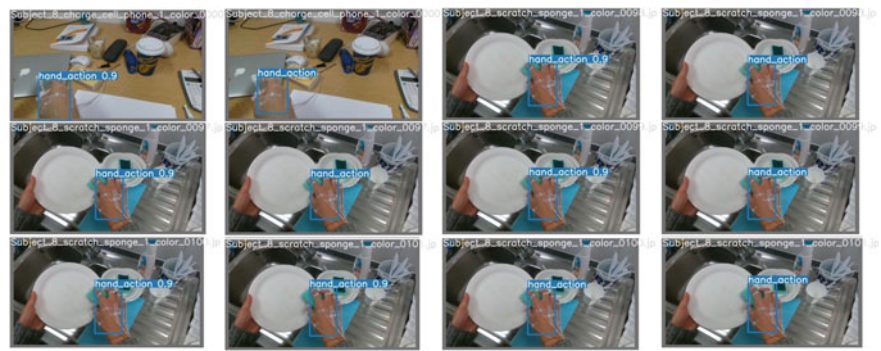
IOU threshold/Methods/Precision (P) (%), Recall (R) (%)		YOLOv7		YOLOv7-w6		YOLOv7-X	
		P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Conf. #123	$A_{50}^{val} / A_{50}^{test}$	99.3/99.7	98.1/96.9	99.3/99.7	97.8/96.9	99.3/99.2	98.5/98.7
	$A_{55}^{val} / A_{55}^{test}$	99.2/99.7	98.1/96.9	99.3/99.7	97.8/96.9	99.3/99.2	98.5/98.7
	$A_{60}^{val} / A_{60}^{test}$	99.2/99.6	98.1/97	99.3/99.7	97.8/96.9	99.3/99.2	98.5/98.7
	$A_{65}^{val} / A_{65}^{test}$	99.2/99.6	98.1/96.9	99.3/99.7	97.8/96.9	99.3/99.1	98.5/98.7
	$A_{70}^{val} / A_{70}^{test}$	99.2/99.7	98.1/96.8	99.3/99.7	97.8/96.9	99.3/99.1	98.4/98.5
	$A_{75}^{val} / A_{75}^{test}$	99.4/99.4	97.9/96.9	99.3/99.7	97.8/96.9	99.3/99.2	98.4/98.3
	$A_{80}^{val} / A_{80}^{test}$	99.4/99.5	97.9/96.8	99.3/99.7	97.8/96.9	99.3/99.5	98.4/97.8
	$A_{85}^{val} / A_{85}^{test}$	99.1/99.5	97.8/96.5	99.4/99.5	97.7/96.8	99.3/99.5	98.4/97.3
	$A_{90}^{val} / A_{90}^{test}$	99.1/99.1	97.8/96.2	99.1/99.4	97.7/96.5	99.1/99.6	98.1/96.3
	$A_{95}^{val} / A_{95}^{test}$	96.6/97	94.3/93.9	98.4/98	95.5/94.4	97.7/97.5	95.5/94.5

**Table 3** Results of hand action detection and classification on the *Conf.* #213 of the FPHAB dataset

IOU threshold/Methods/Precision (P) (%), Recall (R) (%)		YOLOv7		YOLOv7-w6		YOLOv7-X	
		P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Conf. #213	$A_{50}^{val} / A_{50}^{test}$	99.7/99.3	96.9/98.1	99.7/99.3	96.9/97.8	99.2/99.3	98.7/98.5
	$A_{55}^{val} / A_{55}^{test}$	99.7/99.2	96.9/98.1	99.7/99.3	96.9/97.8	99.2/99.3	98.7/98.5
	$A_{60}^{val} / A_{60}^{test}$	99.6/99.2	97/98.1	99.7/99.3	96.9/97.8	99.2/99.3	98.7/98.5
	$A_{65}^{val} / A_{65}^{test}$	99.6/99.2	96.9/98.1	99.7/99.3	96.9/97.8	99.1/99.3	98.7/98.5
	$A_{70}^{val} / A_{70}^{test}$	99.7/99.2	96.8/98.1	99.7/99.3	96.9/97.8	99.1/99.3	98.5/98.4
	$A_{75}^{val} / A_{75}^{test}$	99.7/99.4	96.9/97.9	99.5/99.3	96.8/97.8	99.2/99.3	98.3/98.4
	$A_{80}^{val} / A_{80}^{test}$	99.4/99.4	96.9/97.8	99.7/99.3	96.9/97.8	99.2/99.3	98.3/98.4
	$A_{85}^{val} / A_{85}^{test}$	99.5/99.1	96.8/97.8	99.5/99.4	96.8/97.7	99.5/99.3	97.8/98.4
	$A_{90}^{val} / A_{90}^{test}$	99.5/99	96.5/97.1	99.4/99.1	96.5/97.7	99.5/99.1	97.3/98.1
	$A_{95}^{val} / A_{95}^{test}$	97/97.6	93.9/95.4	98/98.4	94.4/95.5	99.6/97.7	96.3/95.5



**Fig. 2** The distribution of precision, recall, and  $F_1$ -score of hand action detection on the test set of Conf. #123 of FPHAB dataset when  $\text{Thesh}_{\text{IOU}} = 0.5$



**Fig. 3** Illustrating some results of hand action detection and classification on the testing set of Conf. #123 of FPHAB dataset when  $\text{Thesh}_{\text{IOU}} = 0.5$

than 95%, even if the  $\text{Thesh}_{\text{IOU}} = 0.95$ , which is close to absolute accuracy. Tables 2 and 3 also show that  $P$  is usually greater than  $R$  in most cases. This is because in the image of the FPHAB dataset there can be two hands, and as a result, there are many background data areas that are mistakenly detected as the hand action, so the FN increases. Therefore,  $R$  is smaller than  $P$  in many cases. The processing time of the hand action detection and classification process is shown in Table 1; it is also very fast to ensure the pre-processing step without much impact on the processing time of the construction applications.

Figure 2 shows the results on precision, recall,  $F_1$ -score, and confusion matrix on the hand action detection on the testing set of the FPHAB dataset when  $\text{Thesh}_{\text{IOU}} = 0.5$ .

Figure 3 illustrates some results of hand action detection and classification on the testing set of Conf. #123 of FPHAB dataset when  $\text{Thesh}_{\text{IOU}} = 0.5$ .

**Table 4** Results of hand classification on the HO4D dataset [7] by the GM and YOLOv7

Hands/Measurements/Hand sub-data		Sub-data			
		ZY2021 0800001	ZY2021 0800002	ZY2021 0800003	ZY2021 0800004
<i>GM</i>					
Left hand	<i>P</i>	63.0	66.7	51.18	–
	<i>R</i>	65.22	69.8	54.19	–
	<i>F<sub>1</sub>_score</i>	64.09	68.21	52.64	–
Right hand	<i>P</i>	34.8	47.75	33.59	42.63
	<i>R</i>	38.44	52.6	36.2	46.34
	<i>F<sub>1</sub>_score</i>	36.53	50.06	34.85	44.41
<i>YOLOv7</i>					
Left hand	<i>P</i>	95.0	97.2	94.8	–
	<i>R</i>	96.21	98.7	94.78	–
	<i>F<sub>1</sub>_score</i>	95.6	97.9	94.79	–
Right hand	<i>P</i>	94.6	92.5	93.66	92.3
	<i>R</i>	96.25	94.8	96.4	96.3
	<i>F<sub>1</sub>_score</i>	95.42	93.64	95.01	94.26

The results of classification of the left hand, right hand, and background on the HO4D dataset [7] based on the GM and YOLOv7 are shown in Table 4. In Table 4, *P* of the left hand is from 51.18 to 66.7%, *P* of the right hand is from 33.59 to 47.75% by the GM [8]. This result is low because in the HOI4D dataset, there are many objects that the manipulation hand has a similar color to the skin of the hand.

The classification and hand detection results in Tables 4 and 5 are very high; the results are greater than 95% even when the  $Thesh_{IOU}$  is 0.95. The processing time of hand detection and classification is about 35 fps by the GM. The programming source code is developed on the Python programming language and compiled on the Visual Studio Code software with the computer configuration in Sect. 3.1.

## 4 Conclusion

Estimating 2D and 3D hand postures on data obtained from egocentric vision can build visual applications such as assisting the blind, human–machine interaction, playing games, etc. To carry out these studies, the hand in the image must first be detected and classified. In this paper, we propose to use YOLOv7 and its variants (YOLOv7-w6, YOLOv7-X) to fine-tune the hand action detection and classification model on the FPHAB and HOI4D datasets. The research results show very impressive results of YOLOv7 and its variants for hand action detection and classification. The processing time is very fast and does not affect the calculation time of the application-

**Table 5** Results of hand detection on the HO4D dataset [7] by the GM and YOLOv7

Methods/Average precision (AP)/Hand sub-data			Sub-data			
			ZY2021 0800001	ZY2021 0800002	ZY2021 0800003	ZY2021 0800004
GM						
Left hand	Average precision (AP) (%)	AP <sub>50</sub>	37.6	60.59	29.74	–
		AP <sub>55</sub>	32.1	58.04	26.04	–
		AP <sub>60</sub>	26.1	54.09	22.17	–
		AP <sub>65</sub>	14.5	48.67	18.27	–
		AP <sub>70</sub>	20.11	42.01	13.24	–
		AP <sub>75</sub>	9.7	33.22	8.63	–
		AP <sub>80</sub>	5.92	24.72	5.31	–
		AP <sub>85</sub>	2.87	14.48	2.46	–
		AP <sub>90</sub>	0.83	4.3	0.79	–
Right hand	Average precision (AP) (%)	AP <sub>50</sub>	92.22	97.35	89.43	90.66
		AP <sub>55</sub>	90.47	96.59	87.43	88.81
		AP <sub>60</sub>	87.26	95.33	84.49	85.93
		AP <sub>65</sub>	81.95	93.08	80.28	81.31
		AP <sub>70</sub>	74.29	89.11	73.55	73.74
		AP <sub>75</sub>	63.69	82.1	62.87	62.61
		AP <sub>80</sub>	49.22	69.13	46.89	46.97
		AP <sub>85</sub>	30.89	47.0	28.2	27.6
		AP <sub>90</sub>	13.82	20.05	10.45	10.03
YOLOv7						
Left hand	Average precision (AP) (%)	AP <sub>50</sub>	99.9	99.6	99.7	–
		AP <sub>55</sub>	99.65	99.45	99.04	–
		AP <sub>60</sub>	99.6	99.48	99.17	–
		AP <sub>65</sub>	99.4	99.38	99.12	–
		AP <sub>70</sub>	98.79	99.16	99.04	–
		AP <sub>75</sub>	98.42	98.82	98.75	–
		AP <sub>80</sub>	98.38	98.49	98.6	–
		AP <sub>85</sub>	97.6	96.25	97.2	–
		AP <sub>90</sub>	96.4	96.1	96.9	–
Right hand	Average precision (AP) (%)	AP <sub>50</sub>	99.8	99.82	99.69	99.66
		AP <sub>55</sub>	99.76	99.48	99.4	99.52
		AP <sub>60</sub>	99.43	99.2	99.34	99.31
		AP <sub>65</sub>	99.28	99.12	99.15	99.13
		AP <sub>70</sub>	98.9	98.27	98.6	98.74
		AP <sub>75</sub>	98.2	97.9	97.7	97.6
		AP <sub>80</sub>	97.4	96.75	97.4	96.9
		AP <sub>85</sub>	97.12	96.47	96.7	96.8
		AP <sub>90</sub>	96.82	96.05	96.49	96.3
		AP <sub>95</sub>	95.9	95.7	96.21	95.5

building process based on 2D and 3D hand pose estimation. In the future, we will carry out further studies on the estimation and recognition of hand activity based on the results of hand detection and classification.

**Acknowledgements** This research is supported by Tan Trao University.

## References

1. Allena CD, De Leon RC, Wong YH (2022) Easy hand gesture control of a ROS-car using google MediaPipe for surveillance use. In: Fui-Hoon Nah F, Siau K (eds) HCI in business, government and organizations. Springer International Publishing, Cham, pp 247–260
2. Ansar H, Ksibi A, Jalal A, Shorfuzzaman M, Alsufyani A, Alsuhibany SA, Park J (2022) Dynamic hand gesture recognition for smart lifecare routines via K-Ary tree hashing classifier. *Appl Sci (Switz)* 12(13). <https://doi.org/10.3390/app12136481>
3. Cai Z, Vasconcelos N (2021) Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans Pattern Anal Mach Intell* 43(5):1483–1498, 1906.09756. <https://doi.org/10.1109/TPAMI.2019.2956516>
4. Garcia-Hernando G, Yuan S, Baek S, Kim TK (2018) First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In: Proceedings of computer vision and pattern recognition (CVPR)
5. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) YOLOx: exceeding yolo series in 2021. arXiv preprint [arXiv:2107.08430](https://arxiv.org/abs/2107.08430)
6. Indriani, Harris M, Agoes AS (2021) Applying hand gesture recognition for user guide application using MediaPipe. In: Proceedings of the 2nd international seminar of science and applied technology (ISSAT 2021), vol 207, pp 101–108. <https://doi.org/10.2991/aer.k.211106.017>
7. Liu Y, Liu Y, Jiang C, Lyu K, Wan W, Shen H, Liang B, Fu Z, Wang H, Yi L (2022) hoi4d: a 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 21013–21022
8. MediaPipe (2022) MediaPipe hands [Online]. <https://google.github.io/mediapipe/solutions/hands>. Accessed 25 Oct 2022
9. Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement, pp 1–6. [arXiv:1804.02767](https://arxiv.org/abs/1804.02767), <https://pjreddie.com/media/files/papers/YOLOv3.pdf>
10. Wang CY, Bochkovskiy A, Liao HYM (2021) Scaled-YOLOv4: scaling cross stage partial network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 13029–13038
11. Wang CY, Yeh IH, Liao HYM (2021) You only learn one representation: unified network for multiple tasks. [arXiv:2105.04206](https://arxiv.org/abs/2105.04206)
12. Wang CY, Bochkovskiy A, Liao HYM (2022) YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, pp 1–15. [arXiv:2207.02696](https://arxiv.org/abs/2207.02696)

# Interaction-Driven Design: A Case Study of Interactive Lighting



Cun Li and Qiao Liang

**Abstract** Due to the rapid development of screen-based interaction, the word “interaction” is interpreted narrowly as a screen or touching interaction. In this paper, we present six interactive lighting design cases adopting the “Interaction-driven design”, which is initiated and driven by interaction behaviours rather than the user’s needs or technology. Each design case consists of design background, interactive process, and prototyping. Based on the exploitation of the six interactive lighting design cases, we aim at expanding the interaction methods and boundaries, increasing the bonding between users and products, and creating a brand-new user experience. In the discussion section, we reflect on design considerations of interaction-driven design, including conducting lighting design aiming at creating a new type of user behaviour, the input behaviour, as the core of the interaction, creating peak experience, and reflections on the relationship between interaction behaviour and product form, and the evaluation criteria for the interaction-driven pattern.

**Keywords** Interaction-driven design · Lighting design · Human–computer interaction

## 1 Introduction

In 2004, when people used a Motorola V3 cell phone, they usually talked about the product itself, like “it is super thin and light”, “it is a flip phone”, and “the black material seems so special”. However, when it comes to 2007, when people used the first generation of iPhone, besides talking about the unique style due to a giant screen, it is more about how to operate the screen. For example, “I can use my finger to touch those App icons”, “unlocking it by sliding the icon from left to right”, and

---

C. Li (✉) · Q. Liang

School of Design, Jiangnan University, 1800 Lihu Blvd, Bin Hu Qu, Wu Xi Shi 214126, Jiang Su Sheng, China

e-mail: [CunQ.Design@outlook.com](mailto:CunQ.Design@outlook.com)

“using two fingers to zoom a picture”. This represents a shift from classic industrial design to interactive design, from physical attributes to behavioural attributes, from product to service, and, last but not least, from function to user experience. Since then, interaction design has made significant progress in cell phones and PC based on the universal usage of screens. However, due to the rapid development of screen-based interaction, the word “interaction” is interpreted narrowly as a screen or touching interaction. The truth is that interaction is a much broader concept involving software and hardware. For example, a lighting that can be turned on by a blow or a loudspeaker whose volume can be adjusted by hand gesture is closer to the essence of the interaction.

The design practices in this paper take “Interactive Lighting” as the theme and use intelligent hardware to create an interactive behaviour. It tries to expand the interaction methods and boundaries, increase the bonding between users and products, and create a brand-new user experience. Based on this topic, different project teams have conceived various ways to interact with light, including but not limited to obtaining light, adjusting light, and setting light. An intelligent prototype is built with the help of the Arduino platform and related sensors, and the input and output are realized through rounds of debugging iteration.

Therefore, this paper’s study aim is to explore “interaction-driven design”, which means that it is initiated and driven by interaction behaviours rather than the user’s needs or technology. The detailed concept of “interaction-driven design” and related studies will be elaborated on in the next section.

## **2 Related Work**

The study project links the following areas of research: interaction-driven design, non-functional interaction, and lighting design.

### ***2.1 Interaction-Driven Design***

In Maeng et al.’s study, they define the concept of “interaction-driven design” [1], which focuses on the movements in interaction. The interaction-driven design is parallel with the user-driven product and the technology-driven product, which means that the product design is initiated and driven by interaction behaviours rather than the user’s needs or technology. The concept arises from the background that interactivity is being defined as a factor that is independent of users and technology [2]. While in Lee et al.’s study, they also point out that interaction should be defined as an independent factor, rather than being subordinate to user needs and technology [3]. In summary, the “interaction-driven design” can be characterized as a non-functional perspective dealing with movements of user behaviour.

## 2.2 *Non-Functional Interaction*

The research field related to “interaction-driven design” is “non-functional interaction”. For instance, *Designing Behaviour in Interaction* [4]—movements (behaviours) are created by the choreographer and applied to the actual design product (lights). *Interaction Relabelling* [5]—when designing product interactions, a method of extracting ideas by relating the interactions (mechanisms) of two unrelated products. *Rich Interaction* [6]—the buttons for manipulating the product are removed, and a way of interaction is proposed that realizes its function through a higher rate of physical movement of the product and the structure. *Soft(n)* [7]—explores the concept of somaesthetics to express interaction in design. This concept was brought to life through *soft(n) design*, an interactive tangible art installation developed. *Interactivity Attributes*—a new approach to conceptualizing and designing interactive artefacts is presented, emphasizing the importance of elucidating the complex qualities of interactivity to facilitate aesthetic interaction design [8].

In summary, in the current research on interactive product design, non-functional interaction can be classified into the following two categories. The first category is to explore new interaction styles for the current function, including the above *Designing Behaviour in Interaction*, *Interaction Relabelling*, *Rich Interaction*. The second category is to identify the interaction quality, including the above *soft(n)*, and *Interactivity Attributes*.

## 2.3 *Lighting Design*

The lighting design is an extensive research field. Related studies cover various fields, for instance, therapeutic lighting design for the elderly [9], architectural lighting design [10], and Hotel Guestroom Lighting Design [11]. In Jay et al.’s study, they conduct a review study on subjective criteria for lighting design [12].

As technology develops, the concept of smart lighting has become one of the developing trends in lighting design [13] focuses on facilitating street lighting infrastructure integrating sensors, control and communication capabilities in the context of smart cities. Ivan et al. also define the concept of smart lighting: a heterogeneous and multidisciplinary field in lighting management, with the potential to integrate a wide range of sensor and control technologies, to achieve greater efficiency and lower negative impacts [14].

Among which the most relevant research area is interactive lighting: van de Werff et al.’s research defines design considerations for interface properties, shared control, and hybrid control. This work helps realize the potential benefits of interactive office lighting [15]. Aliakseyeu’s study bring together a community of researchers, designers, and technologists to explore the potential of interactive city lighting and how it could support or enhance the lives of those living in a city [16]. Alexander Wiethoff et al. report on preliminary insights on how users perceived the interaction



with an interactive lighting system in front of a large, public audience. Paredes et al.'s research defines design considerations for interface properties, shared control, and hybrid control, which helps realize the potential benefits of interactive office lighting [17]. Nacsca et al.'s study explores the relationship between interaction design and system design by building interactive illuminated objects as part of a shared meaning system [18].

### 3 Six Cases of Interactive Lighting

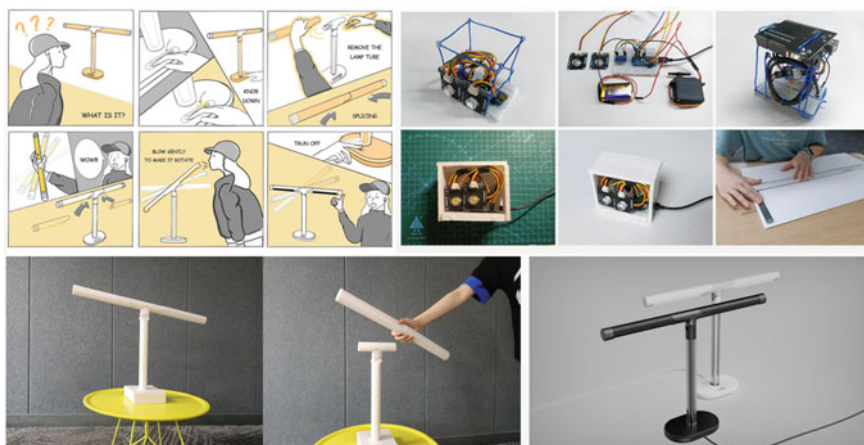
In this section, we present six interactive light designs, which take “Interactive Lighting” as the theme and use intelligent hardware to create an interactive behaviour.

#### 3.1 Case 1: Visualight

**Design background.** Light is invisible, borderless, and free of gravity. But what if light has weight? How should this weight be presented and further interacted with? Based on such thinking, the team members conducted product research, technical scanning and brainstorming, and finally came up with the concept of “light with weight”. **Interactive Process.** The interaction process with light is divided into four steps: capturing, distributing, weighing, and adjusting. Firstly, light can be captured by taking down and combining the lamp tubes and shaking them in the hand. Then, tilting the combined lamp can re-distribute the brightness of the light between the two lamp tubes. After being dismantled and put back on the lamp holder, the lamp will naturally tilt according to the brightness of the light—just like those lights have weight. If users want to adjust the brightness of the light again, they can blow gently at one end. **Prototyping.** Team members use three-axis acceleration sensor, airflow sensor and RFID to complete the interaction process and 3D printing to make the prototype. Then sketches, moodboard, 3D digital model, and 3D renderings are made to deliver the final design result (Fig. 1).

#### 3.2 Case 2: Jane Light

**Design Background.** The explosive information of the Internet has dispersed the time of concentration. How to combine lighting with time management methods such as Pomodoro Technique to improve the concentration of work and study become the design problem of this project. In ancient China, oil lamps were widely used, and people extend the lighting time by adding oil into the lamp. To transfer this user behaviour and make it intelligent is the main purpose of this design. **Interactive**



**Fig. 1** Storyboard and prototype (above), physical mock-up and 3D rendering (below)

**Process.** The final product consists of base, sticky notes, pot, and light. There are two kinds of sticky notes on the product base. The wide one is used to write the work task, and the narrow one is used to set the time. There are three colours of narrow ones which are red, blue, and green, representing 20 min, 40 min, and 60 min, respectively. Users write words on the wide one, tear a narrow one to set time, and then clip them to a special position of the product. Then users take down the pot and point the bottom at the narrow note. The pot will be lightened up to a certain brightness, achieving the effect of “light absorption”. After user “pour” the light onto the top of the product, the main lamp will light up and be set to the according time. After the set time is up, the main light turns dark to remind users to set the next task or rest period. **Prototyping.** Using Hall sensor, ultrasonic sensor, colour recognition sensor, wireless module and other modules, the technical prototype was built after iteration. A paper model is made to simulate further product details (Fig. 2).

### 3.3 Case 3: Wan Xiang

**Design Background.** The kerosene lamp/oil lamp of the last century is a shared memory of the previous generation. By combining the imitation of the tumbler with a lamp, this design tries to convey a spirit which is to be strong and calm in facing hardship in life. **Interactive Process.** Like a tumbler, the product can stand at any angle and position. There are two ways to interact: one is to change the brightness of the lamp by adjusting the angle at which the lampstands, and the other is to change the colour of the lamp by changing the direction of the lamp in reverse. Combining these two ways together, users get the ideal lighting by playing with it. **Prototyping.**

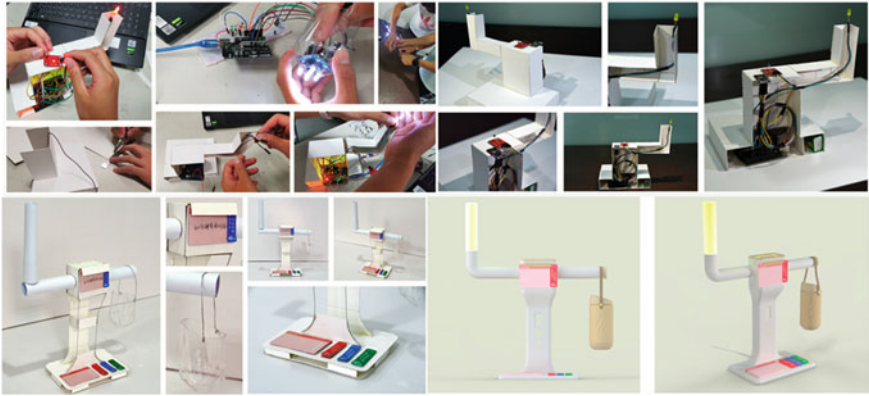


Fig. 2 Physical mock-up and 3D rendering of Jan light

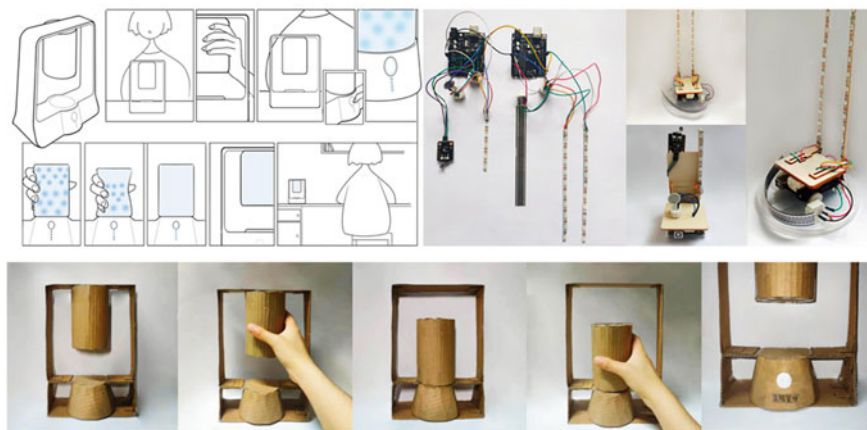


Fig. 3 Prototype, physical mock-up and 3D rendering of Wan Xiang

Ultrasonic sensor, MPU6050 sensor and other modules are used to make the technical prototype. A paper prototype is also made to mock-up the styling (Fig. 3).

### 3.4 Case 4: Washing Light

**Design Background.** Living in a fast-paced era, life is inevitably trifling. People need to find a touch of pure land and a clear spring in the hustle and bustle, so that the impetuous state of mind can return to indifference. **Design Concept.** The interactive inspiration comes from people’s daily behaviour of washing and drying. Water washes away stains and cleans fabrics; during kneading and squeezing, stray light gradually settles, leaving a clear and white lighting. While realizing the lighting



**Fig. 4** Storyboard, inner structure (above), prototype and rendering (below)

function of conventional lightings, it pays more attention to people's inner world and spiritual needs, and pays attention to the mood of life: while washing away stray light, it also washes away depression and troubles in the heart and creates a peaceful and comfortable learning, working or living atmosphere for users. **Interactive Process.** (1) Initial state: the light is hung above the base by magnetic attraction, (2) take the lighting and place it: remove the lighting and place it on the circular boss below to achieve the pre-conditions for cleaning, (3) turn on the light: turn on the light by pressing the button on the base, at this time white light and miscellaneous lights are all there, and (4) washing state: kneading the lighting shade, cleaning the miscellaneous light, and making it sink to the base, as the white light becomes purer, the light representing the miscellaneous light on the base gradually lights up, and the miscellaneous light disappears after a while. Complete process. After 25 min, the stray light will cover the white light and need to be cleaned again; press the base button again to turn off the light (Fig. 4).

### 3.5 Case 5: MUSELAMP

**Design background.** This interactive light brings a more visual experience. We consider combining it with hearing and touch and realize the control of lights through the interaction between people and lamps in different traditional ways. **Design concept.** MUSELAMP is an interesting interactive lamp that controls the colour, brightness, and presentation of light with sound and colour. In the colour mode, the user can control the lighting effect and music by putting a coloured paper tape in the recognition area and pulling it; in the sound mode, the user can control the lighting effect by playing music or spoken language. **Interactive Process.** There are

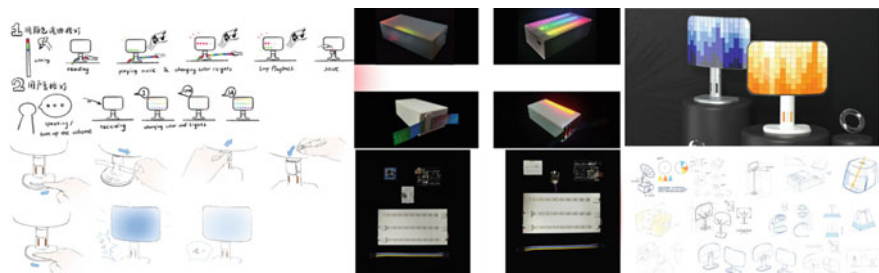


Fig. 5 Storyboard, prototype, rendering, and sketch of the MUSELAMP

two modes: the first one is to control the change of music and light colour through the rhythm and colour changes of the paper tape. The second one is to realize the diversification of lights through external sounds. At the same time, it gives a beautiful and reasonable shape and appearance (Fig. 5).

3.6 Case 6: Match Light

**Design background.** With the development of technology, interaction design has gradually entered our daily life. Good interactive behaviour can bring the relationship between products and users closer, give users a better experience, and add interesting interactions or warmth between people and products. This interactive lamp organically combines natural interactive behaviour with intelligent technology, aiming to allow users to experience more comfortable human–computer interaction (Fig. 6).

**Interactive Process.** The interactive behaviour of this light design draws inspiration of drilling wood to make fire. The specific interactive process is as follows: (1) set the light colour. Move the paddle on the colour palette to choose the favourite colour. Turn the lamp to the position of the colour wheel, align the selected colour to pick the colour, (2) set the brightness of the light: rub the rotatable part of the lamp left and right, the larger the angle, the brighter the light, (3) set the lighting time: rotate the matchstick downwards, (4) turn off the lights when the time limit is reached: rotate the matchstick (5) upwards and turn off the lights when the time limit expires.

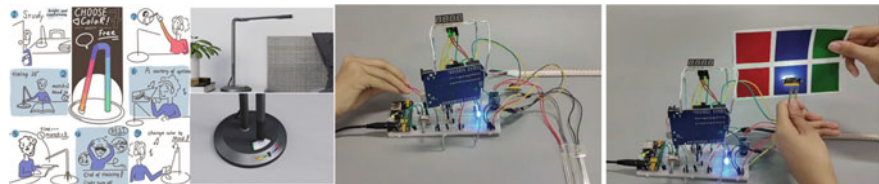


Fig. 6 Storyboard, rendering, prototype of Match Light

The effect of breathing lights will appear, and then the lights will go out. Rotate the matchstick up.

## 4 Discussion

**Conduct lighting design aiming at creating a new type of user behaviour.** In most cases, the starting point of the design is to find and define a user problem or user scenario. However, in this project, designers were asked to design lighting design from interaction behaviour to create a new way of interaction or a new kind of user behaviour. Also, the realization of user interaction strongly relies on possible smart technology, meaning those designs are interaction-driven and technology-based rather than problem-driven. Taking Case 1 as an example, it might look the same as a traditional lighting by form, but the way of interaction has differed profoundly. Through behaviour such as “swing” and “blowing”, a brand-new way of interaction is created, which improves the intelligence and interactivity of traditional lights, making them a better carrier of a smart home. Those design cases represent an attempt at a new design pattern and provide new thinking of product design, which can be widely applied in other product categories.

**The core of the interaction is the input behaviour, and creating peak experience is necessary.** When talking about interaction, inputs and outputs are the keys. On the one hand, the variety of interactive behaviour is of great significance to the enrichment of user experience. It can even be said that the core of the design is neither styling nor technology but the input behaviour itself. On the other hand, what kind of feedback the interactive output gives the user significantly impacts the stickiness, pleasure, and experience of the user’s behaviour. So it is essential to create some moments with peak experience, such as joy, glory, and surprising moments, which elicit specific emotions and create a memorable experience. For example, in Case 3, the adjustment of brightness and colour is connected with playing with a tumbler. In Case 5, the creation of light is based on the attribute of colour.

**Reflection on the relationship between interaction behaviour and product form.** The relationship between interaction behaviour and product form is quite interesting. In some cases, they are interrelated, while others are independent. For the former, the realization of behaviour remarkably impacts product form or even decides the form to some extent. For example, if the behaviour “swing” is used, the stick-like form is usually used for ergonomic reasons. If the user needs to “pour”, the product will likely look like a cup or container. For the latter, the form is more flexible as user behaviour has good transferability. For example, “blowing” only requires a place on the product to mount the airflow sensor, meaning that the product can be designed into various shapes. Compared to attributes of classic industrial design such as styling, and CMF, the advantage of transferability creates larger design space for designers, who can create a flexible product based on the same behaviour.

**Reflection on the evaluation criteria for the interaction-driven pattern.** If the product is designed following a problem-driven pattern, it is usually evaluated on whether it fulfils the defined function, the form, and other related attributes. However, the criteria for the interaction-driven pattern are totally different. Firstly, whether the interaction behaviour is consistent with the design goal. The most important attribute evaluated is the behaviour rather than the product. Moreover, the design goal the designer usually generates is not about function, but about a particular experience, which will be met by interaction design. Secondly, the realization of the prototype. As interaction design, especially those emphasizing hardware interaction, usually rely on technology or platform like Arduino and sensors, the quality of the prototype decides whether users can experience the expected behaviour properly. Thirdly, the smoothness of interaction and the delicacy of experience. When talking about input and output, feedback on behaviour is crucial. This is closely related to the technical level, involving the setting of the parameter. For example, if the input is “blowing”, and the output is “spinning”, how many milliseconds is set for “delay” will have a totally different impact on user experience.

## 5 Conclusion and Future Work

We have reported the six interactive lighting design cases adopting the “interaction-driven design”. As mentioned, we try to expand the interaction methods and boundaries, increase the bonding between users and products, and create a brand-new user experience. Therefore, we conclude the paper with the following reflections: conducting lighting design aiming at creating a new type of user behaviour, the input behaviour, as the core of the interaction, creating peak experience, and reflections on the relationship between interaction behaviour and product form, and the evaluation criteria for the interaction-driven pattern.

As for the future work, since lighting products are relatively simple, the types and quantities of behaviours that are generated in this project are limited. However, due to the excellent transferability, it is possible that the product category can be extended in the future, hopefully finding and defining more abundant user behaviour. Besides, as the realization of interaction behaviour relies heavily on technology, for instance, the types of sensors, it is necessary that designers have more solid technic support so that designers have more freedom in user behaviour design.

**Acknowledgements** This research was funded by Humanities and Social Sciences Foundation project funded by the Ministry of Education of China, grant number 22YJC760031. Philosophy and Social Science Re-search in Colleges And Universities In Jiangsu Province, grant number 2022SJYB0938. All of the cases in this article are designed in a course named *Intelligent Product Development* supervised by the authors. We would like to thank all of the 26 students from Class ID2002, School of Design, Jiangnan University. It was by their great teamwork that the topic of “Interactive Lighting” has solid and interesting cases to be analysed.



## References

1. Maeng S, Lim Y, Lee K (2012) Interaction-driven design: a new approach for interactive product development. In: Proceedings of the designing interactive systems conference on—DIS '12, p 448. ACM Press, Newcastle Upon Tyne, United Kingdom. <https://doi.org/10.1145/2317956.2318022>
2. Lim Y, Stolterman E, Jung H, Donaldson J (2007) Interaction gestalt and the design of aesthetic interactions. In: Proceedings of the 2007 conference on designing pleasurable products and interfaces, pp 239–254
3. Lee S-S, Kim S, Jin B, Choi E, Kim B, Jia X, Kim D, Lee K (2010) How users manipulate deformable displays as input devices. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1647–1656
4. Ross PR, Wensveen SA (2010) Designing aesthetics of behavior in interaction: using aesthetic experience as a mechanism for design. *Int J Des* 4:3–13
5. Djajadiningrat JP, Gaver WW, Fres J (2000) Interaction relabelling and extreme characters: methods for exploring aesthetic interactions. In: Proceedings of the 3rd conference on designing interactive systems: processes, practices, methods, and techniques, pp 66–71
6. Frens JW (2006) Designing for rich interaction: Integrating form, interaction, and function. In: conference; 3rd symposium of design research; 2006-11-17; 2006-11-18, pp 91–106. Swiss Design Network
7. Schiphorst T (2009) soft (n) Toward a somaesthetics of touch. In: CHI'09 extended abstracts on human factors in computing systems, pp 2427–2438
8. Lim Y, Lee S-S, Kim D (2011) Interactivity attributes for expression-oriented interaction design. *Int J Design* 5
9. Shikder S, Mourshed M, Price A (2012) Therapeutic lighting design for the elderly: a review. *Perspect Public Health* 132:282–291
10. Mansfield K (2018) Architectural lighting design: a research review over 50 years. *Light Res Technol* 50:80–97
11. Park N-K, Pae JY, Meneely J (2010) Cultural preferences in hotel guestroom lighting design. *J Inter Des* 36:21–34
12. Jay P (2002) Subjective criteria for lighting design. *Light Res Technol* 34:87–96
13. Castro M, Jara AJ, Skarmeta AF (2013) Smart lighting solutions for smart cities. In: 2013 27th international conference on advanced information networking and applications workshops. IEEE, pp 1374–1379
14. Chew I, Karunatilaka D, Tan CP, Kalavally V (2017) Smart lighting: the way forward? Reviewing the past to shape the future. *Energy Buildings* 149:180–191
15. van de Werff T, van Lotringen C, van Essen H, Eggen B (2019) Design considerations for interactive office lighting: interface characteristics, shared and hybrid control. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–14
16. Aliakseyeu D, van Essen H, Lucero A, Mason J, Meerbeek B, den Ouden E, Wiethoff A (2013) Interactive city lighting. In: CHI'13 extended abstracts on human factors in computing systems, pp 3191–3194
17. Paredes P, Ko R, Calle-Ortiz E, Canny J, Hartmann B, Niemeyer G (2016) Fiat-lux: interactive urban lights for combining positive emotion and efficiency. In: Proceedings of the 2016 ACM conference on designing interactive systems, pp 785–795
18. Nacsa J, Barakova E, Frens J (2011) Sharing meaning and physical activity through a tangible interactive lighting object. In: Proceedings of the second conference on creativity and innovation in design, pp 227–232



# Sustainable Technologies for Environment-Friendly and Ecological Resilience



Paul M. Cabacungan, Khim Cathleen M. Saddi,  
Maria Theresa Joy G. Rocamora, Reymond P. Cao, Salvador P. Granada,  
Paul Ryan A. Santiago, Neil Angelo M. Mercado, Carlos M. Oppus,  
Cristina F. Gonzales, Nathaniel Joseph C. Libatique, Emma E. Porio,  
and Gregory L. Tangonan

**Abstract** We were able to deploy 450 W solar photovoltaic (PV) components for three houses in Naga City, Philippines, that powered solar fans, LED lighting, Clean Water System, NearCloud device, and image transmission via radio. Each clean water system gives 200 L of drinking water after one hour of cleaning. Vegetable growth parameters in hydroponics setup are sent through LoRa-based sensors and are remotely monitored through IoT analytics platform. The domestic liquid waste treatment system reduces the bacteria to 10% which makes it suitable for watering of plants. Partner schools use the Ateneo Innovation Center's (AIC) NearCloud device for wireless dissemination of learning modules to students. The image transmission device can send pictures during emergencies, without a need for a signal. The brick stove and oven use residual heat for water heating while cooking. This Sustainable Technologies for Environment-Friendly and Ecological Resilience (STEER) project gives opportunities for communities to use seven innovative technologies with efficient and economical devices that are neatly incorporated in an eco-resilient house.

**Keywords** Net zero construction · Anticipatory technologies · Disaster resilience · Clean water system · Evacuation center design

---

P. M. Cabacungan (✉) · M. T. J. G. Rocamora · R. P. Cao · S. P. Granada · P. R. A. Santiago ·  
N. A. M. Mercado · C. M. Oppus · C. F. Gonzales · N. J. C. Libatique · E. E. Porio ·  
G. L. Tangonan  
Ateneo de Manila University, Quezon City, Philippines  
e-mail: [aic.sose@ateneo.edu](mailto:aic.sose@ateneo.edu)

K. C. M. Saddi  
Ateneo de Naga, Camarines Sur, Philippines  
e-mail: [civildept@gbox.adnu.edu.ph](mailto:civildept@gbox.adnu.edu.ph)

# 1 Introduction

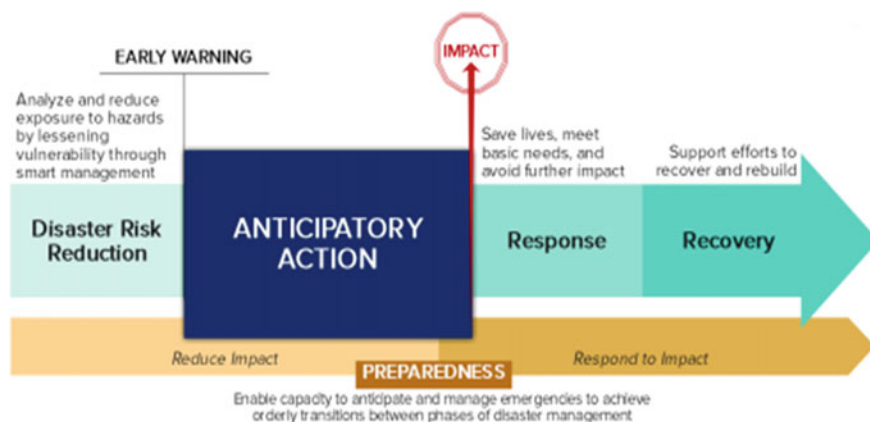
In 2013, a World Bank report estimated that weather-related disasters in the Philippines caused an average of 0.7% Gross Domestic Product (GDP) losses and 90% of damages. Given the 60% exposure of the archipelago to multiple hazards, 74% of its population is considered at-risk [1]. For effective disaster management, the short-term expansion of social protection systems is a decisive factor [2]. There is a need to focus on protecting society via affordable and sustainable technologies, while reducing people's susceptibility to risks. It is important to find a means of merging the medium to long-term merits of sustainable anticipatory technologies and the short-term benefits of effective social protection. This current research project called Sustainable Technologies for Environment-friendly and Eco-Resilient (STEER) Household Model aims to build a resilient house equipped with technologies for climate change adaptation and disaster risk reduction. The innovative architectural design not only reflects 'Green Construction' or sustainability, but also showcased a platform for the proof-of-concept of AIC technologies.

## 1.1 Shelter Design Innovations

In this study, we conducted an energy consumption and offset analysis of our Sustainable Shelter Design that aims to achieve a net zero energy. We used recycled concrete and plastic waste materials in house construction to gain energy credits. For the outdoor kitchen and cooking area, we constructed a highly efficient oven that uses residual heat from cooking to clean and sterilize water for drinking and sanitation. We installed a rain catchment facility to store water for use during the dry season. We used natural techniques for wastewater treatment through a constructed wetland system to achieve zero liquid waste. We used the integrated framework of Environmental Livelihood Security (ELS) that explicitly encapsulates the water-energy-food nexus thinking as a conceptual tool for achieving sustainable development [3].

## 1.2 Anticipatory Technologies

Tozier de la Poterie, et al. described that multi-hazard considerations need to be taken into account in order to provide better interventions [4]. Anticipatory technologies, such as the ones the AIC has produced, contribute to reduced impact and reduced risk in the face of impending hazards, assisting in faster and more efficient disaster response and recovery. Pre-positioning of assets, institution of zoning, as well as creating multi-purpose shelters will assist in reducing the effects of cascading



**Fig. 1** This diagram encapsulates the need for an anticipatory action plan to increase resiliency of vulnerable communities [4]

disasters, making the recovery process quicker. In the event that these anticipatory processes are insufficient, pre-positioned assets should then be available as summarized in Fig. 1.

## 2 Literature Review

Roopnarine in 2013 advocates for the pre-positioning of resources and explains that storing supplies locally brings economic benefits and allows delivery of emergency assistance at maximum speed and minimum cost [5].

### 2.1 Resilient Low-Cost Rural Houses

Post-disaster shelter responses of Bangladesh emphasized the need for more contextual approaches to develop disaster-resilient low-cost rural houses. Properties of the local construction materials were ascertained from laboratory tests. Based on Finite Element Method (FEM) analyses, design needs to consider local affordability and the service and environmental loads [6]. Different treatment schemes for increasing the durability of materials were employed to study their effectiveness [7]. Another study found that as the volume of the Recycled Concrete Aggregates (RCA) in concrete increases, its flexural strength decreases. The flexural strength of concrete with RCA is comparable to the flexural strength of concrete with Natural Concrete Aggregates (NCA) [8]. Saiyari in 2015 utilized Concrete Waste (CW) as RCA, as an alternative to NCA in producing concrete Lego® blocks. It was found that RCA with 50%

replacement yields the highest strength compared to all concrete mixes, which has a percentage increase of 27% compared to Philippine National Standards (PNS). With these illustrated properties, concrete waste possesses potential properties as partial replacement to natural aggregates for non-load bearing concrete blocks [9].

In 2016, Hashemia et al. suggested that a north–south building orientation with the main openings on the north side is recommended as the most appropriate building orientation to reduce the risk of overheating and thermal discomfort. Avoid large openings on the eastern and western side of the house to prevent direct sunlight from coming in and generating too much heat inside. There are various day lighting technologies like skylights, tubular daylighting devices, and minimize excessive solar heat gain [10]. Natural sunlight is solar concentrators [11]. The use of recycled water bottles can provide skylight roofing panel systems.

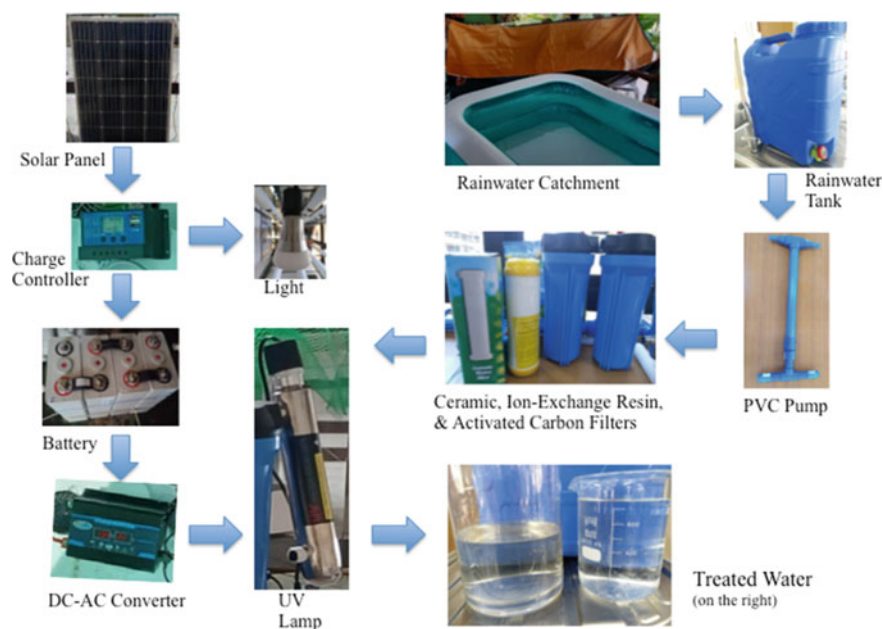
Incorporation of water recycling is important as it allows water reuse for toilet flushing and landscape irrigation. Such a method has been showcased in the zero-wastewater building in the Ateneo de Manila University campus. It has a modular design and proves to be scalable [12]. Domestic liquid waste treatment systems support the Clean Water Act, Ecological Solid Waste Management Act, and Clean Energy Act [13].

Hydroponics is a form of protected agriculture that reduces the use of pesticides and saves water of up to 70–90% while using low-value space in cities such as rooftops [14]. Some features could be added like a temperature and relative humidity monitoring system, which also projects the real-time data in a web server in 2014 [15] and monitoring system for acidity (pH), electrical conductivity (EC), water level, dissolved oxygen, temperature, moisture and humidity, with data displayed on a dashboard in 2019. The indoor setup used a hybrid powering system and grew approximately 100 to 120 high value green leafy plants such as lettuce, arugula, basil, stevia, and parsley [16].

In 2012, Esporlas et al., together with AIC, drew specific details of a brick oven, taken from a design of a Filipino food chain, which addressed the need of a rehabilitation village in Cagayan de Oro City after a disastrous flooding. Copper pipes were implanted inside the oven to use residual heat to clean water passing through the pipes [17].

## 2.2 *Technology Innovations*

The AIC promotes innovative technologies through multidisciplinary teaming and strategic long-term partnerships with the industry. Several technologies of AIC were integrated, one of which was the Water-Electricity-Lighting System (WELS). WELS is a portable clean water system, as shown in Fig. 2, with provision for lighting and cellular phone charging. It can be connected to a rainwater harvesting facility. P. Cabacungan et al. presented a model of engaging the stakeholders that brings mutual benefit to both university and community in WELS deployments [18, 19].



**Fig. 2** Water-electricity-lighting system (WELS) was developed to address the need for clean water in far-flung areas [11, 18]

The AIC and its affiliates developed the AIC NearCloud, a mobile cloud digital library. Contents were preloaded on each NearCloud device and accessed with a smartphone, or other devices. Content can also be updated with low-bandwidth Internet through hotspot or Wi-Fi router [20–24]. Its main component is the Raspberry Pi—a low-cost, credit card-sized single-board microcomputer that offers various interfaces and has considerable processing power that can process and store data. Its ad-hoc Wi-Fi network feature makes the system suitable for remote deployment.

The AIC NearCloud is a low-cost, accessible, and adaptable remote learning platform. Each node has a simple browser-ready interface allowing cross-platform access, with Kolibri installed for content management, creation and distribution, and interactive content. Various contents such as Khan Academy tutorials, PhET interactive labs, and DepEd Commons are available and cached in the device. Advanced educational applications can also be bundled in desktop mode like NetLogo and graphing calculators [22–24].

The NearCloud system is a powerful caching device designed as part of the disaster communication system which consists of LoRa beacons, Unmanned Aerial Vehicles (UAVs), ground responders, and command and control center. Beacon messages and multimedia files received by the aggregator were uploaded in the mobile cloud. The system cached with HD maps, disaster and medical information, and facial network served as an aid in disaster response [22–24].

When telecommunication infrastructures are down, radio communication systems can be utilized with Fast Light Digital or FLDIGI—a free and open-source application which allows an ordinary device’s sound card to be used as a simple two-way data modem. The software connects the microphone and headphone connections of an amateur radio SSB transceiver or an FM two-way radio to the computer’s headphone and microphone connections, respectively [25–27].

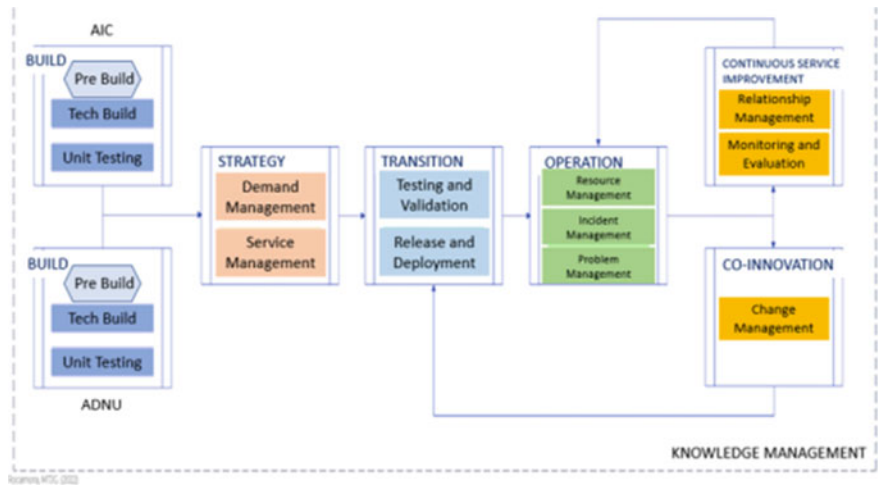
We employed LoRa and NearCloud technologies to address the post-disaster communication needs. Solpico et al. [28] and dela Cruz et al. [21] demonstrated a disaster-resilient communication system that uses a UAV equipped with a LoRa node as data aggregator to gather information from survivors and rescuers who are also carrying LoRa beacon nodes and send it to a mobile NearCloud server. The research shows that Mobile Cloud wireless mesh network offers an ICT system for information gathering, transmission, and processing.

### 3 Methods

The process of using anticipatory technologies in the resilient housing project, as done by the AIC and Ateneo de Naga University (AdNU), employs a service model approach which promotes co-innovation and co-ownership between partners. Through its phases, responsibility with the maintenance, monitoring, and evaluation of the system is distributed between both parties, while ensuring that the homeowners/beneficiaries, who will be trained in the usage of the technologies and its maintenance, will also have a hand for co-innovation. Additionally, this presents them with the chance to transform the technologies handed over to them into sustainable livelihood opportunities. The process, as clearly presented in Fig. 3, is self-sustaining, while still allowing for the liberty of further innovation and adaptation based on their environment and available resources.

#### 3.1 *Pre-build Consultations and Community Awareness*

The Ateneo de Naga University’s Civil Engineering Department designed and supervised the construction of the proposed houses. The project site was determined based on the results of the 2019 Naga City Climate Disaster Risk Reduction Assessment (CDRA) report and other data from the local government units (LGU) with special focus on the main hazards found in Naga City. While professional advice is important in designing the house, input from the future homeowners plays an important role. Finite Element Method (FEM) was used in the structural analysis.



**Fig. 3** Service model process flow for co-innovators is the framework used in co-developing and co-innovating technologies with our partners

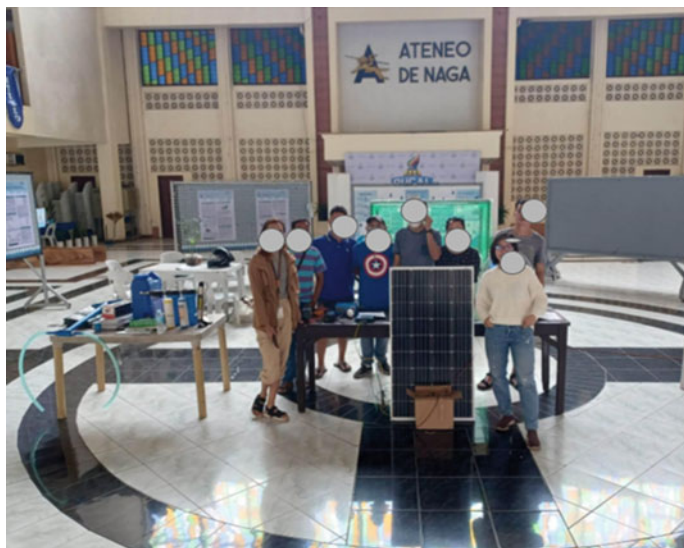
**3.2 Three-Phase Workshop**

Three different workshops were conducted. Workshop 1 was co-designing with the community prior to the preparation of the technical designs. We implemented the principle of co-generation of knowledge, seen in Fig. 4, as we engaged the community in the discussion of brainstormed sketches. Aligned to our advocacy for the environment and sustainability, we gave particular emphasis on natural ventilation, the use of solar fans, the presence of vegetation in hydroponics setup and plants nurtured by the waste treatment system.

Workshop 2 was a capacity building related to the production of building materials prior to the bayanihan construction, as clearly seen in Fig. 5. We engaged the community in the production of construction materials. Local skilled workers were tapped for employment opportunities. The construction framework was community-based and simplified for ease of replication. Post-typhoon structural and social assessment was also done after the typhoon season.

Workshop 3 in Fig. 6a provided hands-on training of local residents on the use, maintenance and repair of incorporated technologies. As shown in Fig. 6b, we provided instructional manuals which eventually led to co-ownership of AIC Technologies. We paid special attention to gender and livelihood issues that are integral to the design.





**Fig. 4** Workshop 1 is conducted with emphasis on our principle of co-generation of knowledge



**Fig. 5** Workshop 2 practices the bayanihan culture for constructing the shelter

### ***3.3 Technology Incorporation***

The Water-Electricity-Light System (WELS) was incorporated in the house build. It integrates water filtering and disinfecting systems, lighting and cellular phone charging. Single outlet was utilized to avoid ‘octopus connection’ so mobile phone





**Fig. 6** **a** Workshop 3 includes hands-on training of local residents. **b** Workshop 3 includes providing of instructional manuals

charging is to be done alternately with the use of ultraviolet (UV) disinfection lamps. A two-way radio was also installed for communication during emergencies. Solar fans were put up for air circulation and ventilation.

A rainwater catchment facility was provided to make water supply available. A pre-filtration system was placed on the roof downspout and gutter. Due to the non-load bearing house structure, a water collecting container in the form of an inflatable pool was used as a water storage facility. A do-it-yourself polyvinyl chloride (DIY PVC) pump sucks in water and pushes it into the system for cleaning without any need for electrical power. Filtration and disinfection happen through an off-the-shelf water cleaning kit with three-stage filtration. This filtration system was attached to a UV lamp for ultraviolet disinfection. This UV lamp was enclosed in a stainless canister and consumes only 6 watts of power and has a 254-nm peak wavelength to optimally kill microorganisms. An off-the-shelf water sensor was used to monitor and measure the acidity (pH), electric conductivity (EC), turbidity, and total dissolved solids (TDS) in water.

AIC also used water test kits to measure e-coli and coliform levels in water. These helped maintain water quality in conformance with the Philippine National Standards for Drinking Water (PNSDW).

Wastewater treatment, using constructed wetlands with papyrus plants and off-the-shelf filters and UV lamps, was also included in the house construction. It recycles wastewater and uses it for watering plants and for flushing the toilet.

For energy storage, a 100-Ah, 12 V lithium-ion phosphate battery was used. Off-the-shelf 500-W inverter with built-in 5-V DC phone charger with better heat dissipation, due to its aluminum body and wider operational power range, was used to utilize the stored energy in the battery. Rotated charging of batteries through solar panels ensured optimal use of solar energy and 24/7 power supply. 7-W DC light-emitting-diode (LED) lights were used in the eco-house so they could be powered directly using the batteries and did not have to run through a converter. An efficient

smart 5-W bulb with a built-in battery can also be used as it can readily serve as a flashlight.

An efficient cooking station that allowed heating of food, as well as drinking water, was provided through a brick stove. The hydroponics provided locally produced vegetables. The house elevation, which provides safety measures for flooding, also opened up the possibility of raising livestock in the vacant space below the house. Putting a trellis at the back of the house to provide shade and serve as frame for vines and plants to crawl also controlled the temperature of the microclimate.

A wireless long- and short-range monitoring sensor uses a multi-parameter sensor interface with LoRa nodes with ESP32 chips for smart hydroponics farming. Temperature, moisture, pH, NPK, conductivity, total dissolved solids, and salinity can all be measured using the sensor probe [29]. The LoRa nodes are based on the ESP32 SoC chip, which provides Wi-Fi and Bluetooth communication. It can transmit sensor data to the cloud over a Wi-Fi network and analyze collected data using an IoT platform. Parameters can be monitored remotely and in real time, allowing for efficient utilization of agricultural input resources. A monitoring system incorporating IoT technologies can make farming methods more sustainable.

A pre-positioned information system with the use of NearCloud technology was made available. This served as a local repository of vital data for community use, like maps, important family documents and identification securities.

All the installed systems underwent stress tests through continuous operation of 24 h or more, while monitoring their performance. When efficiently operational, these can also provide an option for a domestic enterprise by selling clean drinking water, vegetables, or papyrus plants. Livelihood options contribute to the sustainability of the systems.

## 4 Results and Discussions

### 4.1 *Pre-build Consultations and Community Awareness*

A total of 450-W solar photovoltaic (PV) components were installed for the three houses in Naga City. Each house consists of a 150-W solar panel, Lithium-Ion battery, 30A max solar charge controller, DC-12 V 15-W solar fan, AC 7-W bulb, 12VDC-220VAC 400-W converter. The expected energy from four hours of sunlight is 600 W-h.

The stored energy is projected at 600 W-h on a sunny day with 4 h of sunlight. Half of this energy can run the UV light, 6 LED bulbs, and DC fan for four hours during night time. The other half can be used for emergencies, like charging mobile phones and/or radio communication devices.

## 4.2 Clean Water System

Each house was installed with a rainwater harvester. A 3-m  $\times$  3-m in size rain catchment canvas tarpaulin leads to the inflatable pool, serving as water storage, which measures 2.00 m  $\times$  1.46 m  $\times$  0.48 m. The harvested rainwater is pushed through the Clean Water System, shown in Fig. 7, by a do-it-yourself (DIY) PVC pump. The particulate filter removes all suspended solids bigger than 0.9  $\mu$ m diameter (pore size of ceramic filter). The ion-exchange resin removes most of the nitrates and sulfates in rainwater and the activated carbon eradicates the odor. The 6-W ultraviolet (UV) lamp kills the viruses and bacteria. The clean water system addresses the need for a safe clean water supply without chlorination.

The use of compact dry coliform test kits visually shows our partner communities the quality of their drinking water that may pose health risks and the water quality after passing through the Clean Water System. Test kits detect the presence of coliform in water after 24 h. Water testing is an integral part of the training to local stakeholders for it builds a strong commitment to maintain safe drinking water quality.

The treated water was sold at PHP1.00/l. Cleaning for 5 h, our partner communities earned PHP 1000/day or PHP 30,000/month which ensured the return of investment



**Fig. 7** Clean water system not only ensures the community of a coliform free drinking water, but also an alternative source of livelihood

(ROI). This alternative livelihood empowers the community and motivates them to maintain the system for sustainability.

### 4.3 *Hydroponics with Sensor and Monitoring System*

The community partners harvested 405 cups of lettuce in a 3-m<sup>2</sup> area in 35–40 days which they sold for PHP 20/cup, with gross earning of PHP 8100/harvest. With this earning, they could reach a return of investment (ROI) after their 6th harvest, profit can start on the 7th harvest. The upfront costs include nine (9) styrofoam boxes, 405 styrofoam cups, 9 kg coco peat, 1000 ml nutrient solution, lettuce seeds pack, slotted angle bar for framing, 70% sunshade garden net and 0.8 mm thick UV plastic. Recycled components can be used. The styrofoam box can last for 3–4 years, while cups can last for 4 months. The excess nutrient solution can be filtered by cheesecloth and be reused again. Used coco peat can be disinfected by domestic bleach, washed and sundried for reuse. With the success of lettuce growing, seen in Fig. 8, hydroponic lettuce has a promising future in the market.

We also developed a Long-Range Radio (LoRa)-based sensor and monitoring system as shown in Fig. 9. The sensor is connected to one LoRa module that forms a sensor node. This node is sending data to the receiving LoRa gateway via 868 MHz frequency antennas. The gateway is connected to a USB Wi-Fi dongle to send the data to ThingSpeak. This IoT analytics platform service aggregates, visualizes, and analyzes data in near real time. It showed growing parameters of lettuce in hydroponics setup and the status of nutrient solution in graphical form for easy remote monitoring. This system is powered by two 20-A-h battery banks and a 100-A-h Li-ion battery. These are charged by a 150-W solar photovoltaic (PV) panel.



**Fig. 8** Hydroponic setup in growing lettuce on a schoolyard offers young students an endeavor in gardening or farming



Fig. 9 These are the components of ThingSpeak near real-time data

Different parameters—electrical conductivity, nitrogen, pH Level, and phosphorus—of the AIC hydroponics can be measured and the 24-h worth of data are shown in Fig. 10.

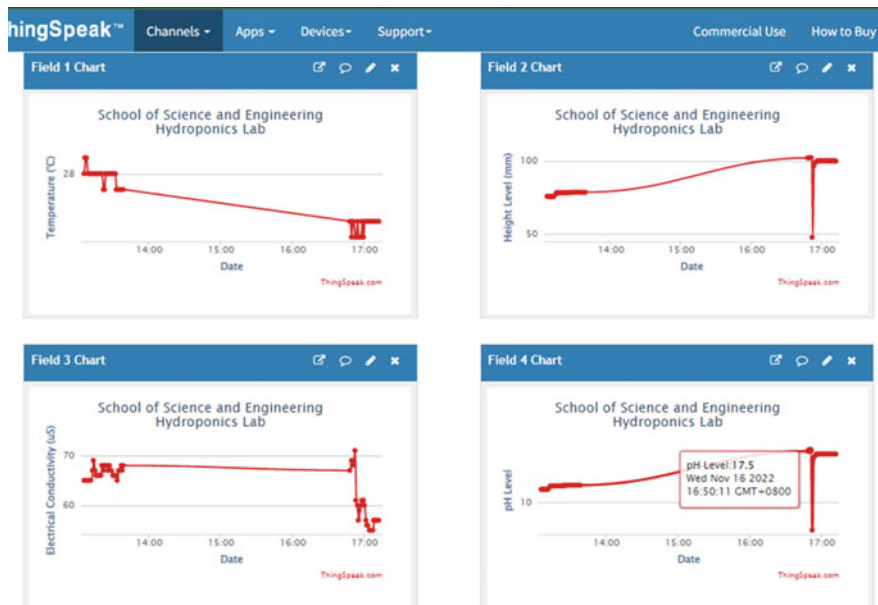


Fig. 10 Free open-source ThingSpeak platform provides near real-time remote data monitoring of growth parameters in protective agriculture



This device can transmit data up to 2.6 km line of sight, and we were able to test it only up to 50-m distance, with a glass and a concrete wall obstruction. The sensor recorded the NPK levels, temperature, moisture, electrical conductivity, acidity, and salinity of the nutrient solution of the hydroponics system. We were able to monitor the nutrient contents of the mixture especially during the rainy days when the solution was diluted by rainwater. Because of this, we were able to immediately readjust the nutrient and water dilution. For easy, better, and more accurate monitoring of the lettuce plants, we used a LoRa module and sensor system.

#### ***4.4 NearCloud System***

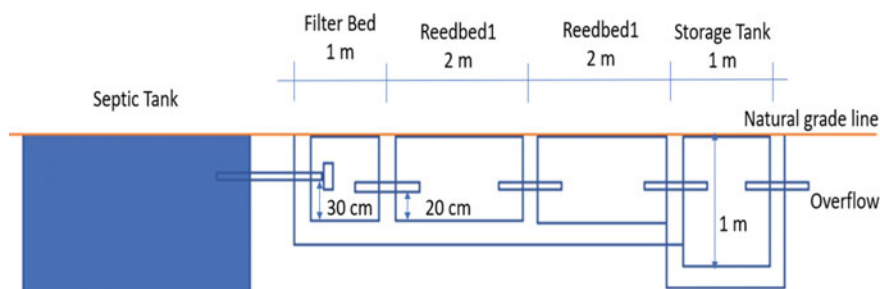
A NearCloud system is a pocket size computer with terabyte storage powered by a small power bank or a solar PV system. It serves as a locally accessible drive which does not require an internet connection. Its LAN port and USB Wi-Fi dongle can be used to connect to the internet. Cached content can be consumed by different users simultaneously without delay. The NearCloud system or EduCloud can help address the learning constraints brought about by insufficient modular materials in schools. Teachers in Brgy. Panicuason were able to digitally share lesson plans and other teaching materials using the AIC NearCloud. Folders can also be created to limit access to certain groups of students or faculty.

#### ***4.5 Image Transmission System***

Two pairs of 433 MHz transceiver radios were turned over to volunteer groups in Brgy. Panicuason. This technology consists of 2 pairs of smartphones and 2 pairs of transceiver radios. The smartphone was loaded with a Fast Light Digital (FLDigi) application that was downloaded for free and can convert pictures to audio signals. This audio signal can be relayed via transceiver radios using the microphone to the speaker and vice versa. Once the audio signal reaches the receiver phone, it will decode back to its original picture file. This image transmission technology makes evacuation and emergency response quick and efficient without depending on cell phone signal. The information received by the Barangay officials can be immediately relayed to Municipal officers for quick response and action.

#### ***4.6 Waste Treatment***

The constructed domestic wastewater treatment system was able to filter 90% of the bacteria from the common septic tank of the three houses. The diagram in Fig. 11 illustrates how the facility is able to catch the septic overflow liquid from a 10 cm



**Fig. 11** Domestic liquid waste treatment system filters 90% of the bacteria and converts the treated water to an agricultural grade

diameter PVC pipe and connect it to our 0.5 m<sup>3</sup> concrete filter tank. This filter consists of 0.06 m<sup>3</sup> gravel, 0.03 m<sup>3</sup> charcoal, and 0.06 m<sup>3</sup> sand. It is connected to a 4 m × 1 m × 0.5 m concrete reed bed system. This reed bed is made up of 8.26 cm, 0.8 m<sup>3</sup> gravel (alternative aggregate recycled concrete), 0.4 m<sup>3</sup> sand, 7 bags cement, 150 pcs, 10 cm × 20 cm × 40 cm CHB, 20 pcs. of #10 round bar, 10pcs. #9 round bar, a concrete filter tank cover and waterproofing. The reed bed contains 1.9 cm (¾ in.), 2.5 m<sup>3</sup> gravel as the papyrus plant grow bed. We used 64 papyrus plants at regular distances to reduce the bacteria to 10%. The flowrate of the system is 400 L for 24 h. After the reed bed, there is a 500-L concrete tank to store the treated water. It is pumped out to water the plants using a DIY PVC hand pump.

The partner community can sell small papyrus plants for PHP 300/piece, either for use as an ornamental plant for landscaping or for basket making. Release of agricultural grade or class-C water from the septic tanks lessens water pollution in rivers and soil contamination.

#### 4.7 Brick Stove and Oven Equipment

We constructed a brick stove and oven using the following materials: old two-burner stove, 20 pieces 5-cm (2-inch) × 10-cm (4-in.) × 20-cm (8-in.) brick stone, 2 burner stove, 1 sack cement, 2 sacks of fine sand, 3 m #10 round bar, and 2 m 0.75 cm (½-in.) diameter copper tube. This system enables the community partners to cook their food, toast bread, and boil water simultaneously. It uses scrap wood and charcoal to fuel it. The bricks and the fire chamber cover minimize the heat loss as shown in Fig. 12. There is a small hole at the fire chamber cover to allow air to come in and a chimney to release fumes. The firebrick cooker is permanently mounted on a concrete stand shared by the three families. It is an economical cooking equipment which uses readily available materials without the need for an expensive gas tank nor electrical supply. Local technicians appreciate the new knowledge of constructing it. They see the potential of using this for a food business in the future.



**Fig. 12** Brick stove and oven with water heating provision is an energy efficient cooking equipment

The team made instructional manuals of all these technologies using the local language. Detailed specifications of each component were provided, as well as the projected costs, and assembly instructions. Demonstration and hands-on training of all technologies were conducted to community recipients, including women and children, to ensure the recipients can maintain and operate these systems. We insisted that the mothers and teenagers are properly trained because most of the fathers are working in the city. We also trained AdNU students on some of these technologies, and they helped in the installation. This strategy of involving students adds to their skills and knowledge and ensures the maintenance of the technologies.

## 5 Conclusion

The research team recognizes the importance of prepositioning assets before an emergency or series of cascading events strike. These assets are aligned with the United Nations Sustainable Development Goals (UN SDG) #s 2, 4, 5, 6, 7, 11, and 17. Addressing the lack of electricity, water, food, communication, educational devices, environment care, and cooking facilities in times of need will definitely make communities more resilient to calamities. These partner communities are owning these sustainable technologies as they learned how to operate, maintain and earn from them. They were not only trained in the basic know-how of the technologies, they also learned how to gain profit in an environment-friendly and ecological resilient



way. This STEER project includes the community recipients as partners in the use, administration, maintenance, and management of these technologies. This incorporation of seven innovative technologies will be beneficial to their daily activities and contribute to their livelihood and calamity resilience.

As the Philippines engineers disaster resilience in Local Government Units and impacted communities, we expect to implement the ‘net zero’ green construction techniques that reuse energy-intensive materials and incorporate the Anticipatory Technologies (Energy Hub, Clean Water Systems, and NearCloud Information and Communications) into new designs of safe zones and evacuation centers.

**Acknowledgements** We would like to thank the Ateneo University Research Council and Ateneo School of Science and Engineering for funding this project. We are also grateful to our partners—Coastal Cities at Risk Philippines (CCARPH), Center for Community Development and Civil Engineering Department of Ateneo de Naga University for their full support.

## References

1. World Bank (2013) Building resilience: integrating climate and disaster risk into development. Lessons from World Bank Group experience, The World Bank, Washington DC
2. Aleksandrova M, Balasko S, Kaltenborn M, Malerba D, Mucke P, Neuschäfer O, Radtke K, Prütz R, Strupat C, Weller D, Wiebe N (2021) WorldRiskReport 2021—focus: social protection. Bündnis Entwicklung Hilft. ISBN 978-3-946785-12-5
3. Biggs EM, Bruce E, Boruff B, Duncan JMA, Horsley J, Pauli N, McNeill K, Neef A, Van Ogtrop F, Curnow J, Haworth B, Duce S, Imanari Y (2015) Sustainable development and the water–energy–food nexus: a perspective on livelihoods. <https://www.sciencedirect.com/science/article/pii/S1462901115300563?via%3Dihub>. Last accessed 2022/08/08
4. Tozier De La Poterie A, Clatworthy Y, Easton-Calabria E, Coughlan De Perez E, Lux S, van Aalst M (2021) Managing multiple hazards: lessons from anticipatory humanitarian action for climate disasters during COVID-19. *Clim Dev* 1–15. <https://doi.org/10.1080/17565529.2021.1927659>. Last accessed 2022/08/11
5. Roopnarine L (2013) How pre-positioning can make emergency relief more effective. *The Guardian*, 17 Jan 2013. [www.theguardian.com/global-development-professionals-network/2013/jan/17/prepositioning-emergency-relief-work](http://www.theguardian.com/global-development-professionals-network/2013/jan/17/prepositioning-emergency-relief-work). Last accessed 2022/08/11
6. LGA (2018) Operation L!STO: disaster preparedness manual v3 for city and municipal Local Government Units, Local Government Academy [Online]. Available: [https://v2v.lga.gov.ph/media/uploads/2/Knowledge%20Exchange/Operation%20Listo%20Manual/Listo%20Manual%20City%20Municipal%20LGUs\\_Final%20Version%202018.pdf](https://v2v.lga.gov.ph/media/uploads/2/Knowledge%20Exchange/Operation%20Listo%20Manual/Listo%20Manual%20City%20Municipal%20LGUs_Final%20Version%202018.pdf). Last accessed 2022/08/05
7. Islam M, Hossain T, Moles O, Podder R, Caimi A, Nayar F (2015) Disaster resilient rural house designs for different geographic regions of Bangladesh
8. Belarmino LY, Diaz JRG, Niedo RO, Ezraphil JB (2012) Recycled concrete as an alternative coarse aggregate for concrete on grade. PULSAR 1.1. Web. 04 Feb 2016
9. Saiyari D (2015) Utilization of concrete waste as partial replacement to aggregates for nonload bearing concrete Lego® block. In: International conference on environmental quality concern, control and conservation
10. Hashemia A, Khatamic N (2016) Effects of solar shading on thermal comfort in low-income tropical housing. In: 8th International conference on sustainability in energy and buildings
11. Sharp F, Lindsey D, Dols J, Coker J (2014) The use and environmental impact of daylighting. *J Cleaner Prod* 462–471

12. Granada S Jr, Libatique N, Tangonan G (2011) Design and development of integrated wastewater treatment using natural system UV irradiation and PV cells. Unpublished thesis, Ateneo de Manila University
13. Oppus C (2020) Waste management and treatment in Ateneo de Manila University, Ateneo Innovation Center, School of Science and Engineering
14. Urban Greens (2019) Hydroponic workshop 101
15. Ado R, Brin M, Guzon G, Juen A (2014) Solar-powered aquaponic system (SAP) with mobile application: an aquaculture system with hydroponics in a symbiotic environment for faster production of vegetables (unpublished)
16. Van Cutsem T. Urban hydroponics research for high value product (unpublished)
17. Esporlas L, Flores S, Limpin J, Paloma M, Ng S, Zabala K (2012) Brick oven implementation (unpublished)
18. Cabacungan PM, Tangonan GL, Cabacungan NG (2020) Water-electricity-lightsystem: technology innovations. *Int J Recent Technol Eng* (in press). <http://ijrte.org/download/volume-8-issue-6/>. Last accessed 2022/08/05
19. Cabacungan P, Cabacungan N, Tangonan G (2020) University-community partnership for water technology deployment and co-innovation: a decade of engagement. *Int J Adv Res Publ* 4(5). Available:<http://www.ijarp.org/published-research-papers/may2020/Universitycommunity-Partnership-For-Water-Technology-Deployment-And-Co-innovation-A-Decade-Of-Engagement.pdf>. Last accessed 2022/08/05
20. Talusan JP et al (2018) Mobile cloud: low-cost low-power cloud implementation for rural connectivity and data processing. In: 2018 IEEE 42nd Annual computer software and applications conference (COMPSAC), Tokyo, Japan, pp 622–627
21. dela Cruz JA, Libatique NJ, Tangonan G (2019) Design of a disaster information system using mobile cloud wireless mesh with delay tolerant network. In: 2019 IEEE Global humanitarian technology conference (GHTC), Seattle, WA, USA, pp 1–8. <https://doi.org/10.1109/GHTC46095.2019.9033450>
22. Mercado N, Mamaradlo J, dela Cruz J, Cabacungan P, Libatique N, Tangonan G (2020) Caching strategies with mobile cloud wireless mesh: a study in disaster risk and resilience. In: IEEE Global humanitarian technology conference (GHTC), pp 589–596. ISBN 978-1-7281-7388-720
23. Mamaradlo JP et al (2020) University campus 5G testbed and use case deployments in the Philippines, in *Broadband Access Communication Technologies XIV*, vol 11307. International Society for Optics and Photonics. SPIE, pp 15–30
24. Cruz XMM, Honrado JLE, Libatique NJC, Tangonan GL, Oppus CM, Cabacungan PM, Mamaradlo JP, Mercado NM, Dela Cruz JA, Dela Cruz JA, Cruz JV (2022) Design and demonstration of a resilient content distribution and remote asynchronous learning platform. In: *Adjunct proceedings of the 2021 international conference on distributed computing and networking (ICDCN'21)*. Association for Computing Machinery, New York, NY, USA, pp 98–103. <https://doi.org/10.1145/3427477.3428190>. Last accessed 2022/08/05
25. Documentation/FAQ—fldigi. [Fedorahosted.org](http://Fedorahosted.org)
26. Rolling your own with digital Amateur radio. *Linux Journal*. [www.linuxjournal.com](http://www.linuxjournal.com). Last accessed 2022/08/05
27. An Amateur radio survival guide for Linux users. *Linux Journal*. [www.linuxjournal.com](http://www.linuxjournal.com)
28. Solpico D et al (2019) Application of the V-HUB Standard using LoRa Beacons, Mobile Cloud, UAVs, and DTN for disaster-resilient communications. In: 2019 IEEE global humanitarian technology conference (GHTC), pp 1–8. <https://doi.org/10.1109/GHTC46095.2019.9033139>
29. Co J, Tiausas FJ, Domer PA, Guico ML, Monje JC, Oppus C (2018) Design of a long-short range soil monitoring wireless sensor network for medium-scale deployment. In: *TENCON 2018—2018 IEEE region 10 conference*, pp 1371–1376. <https://doi.org/10.1109/TENCON.2018.8650541>

# Preliminary Investigation into a Security Approach for Infrastructure as Code



Ammar Zeini, Ruth G. Lennon , and Patrick Lennon

**Abstract** IaC is relatively a novel technology, with the result that many security frameworks don't have a clear strategy for risk management or threat modelling for infrastructure when implementing IaC techniques. In DevOps, infrastructure is initialized, prepared, managed, and configured with a left-shift on quality. The DevOps methodology increases the integrity and stability of the deployment. IaC works best with DevOps practices for code quality, scalability, security, and reliability. Infrastructure as Code (IaC) promotes managing knowledge and experience through reusable scripts of infrastructure code, instead of the traditional method of manual labour technique, which is typically slow and time-consuming. This research determines some security risks that should be considered during the IaC development process. It further defines the main security practices that should be added into Infrastructure as Code life cycle to fill the gap in the SDLC for IaC. An initial proposal to secure pipelines for IaC is presented.

**Keywords** Infrastructure as code · Security as code · Security development life cycle frameworks · DevOps

---

A. Zeini (✉) · R. G. Lennon · P. Lennon  
Atlantic Technological University, Letterkenny, Ireland  
e-mail: [L00170932@atu.ie](mailto:L00170932@atu.ie)

R. G. Lennon  
e-mail: [ruth.lennon@atu.ie](mailto:ruth.lennon@atu.ie)

P. Lennon  
e-mail: [patrick.lennon@atu.ie](mailto:patrick.lennon@atu.ie)

A. Zeini · R. G. Lennon  
Lero—The Irish Software Engineering Research Centre, University of Limerick, Limerick, Ireland

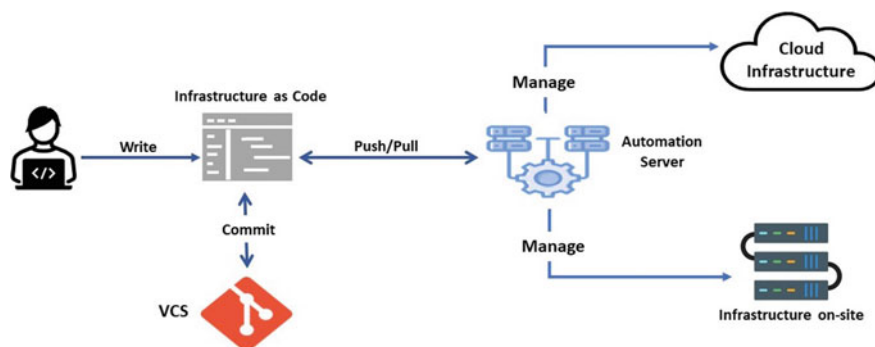
# 1 Open Challenges

With the fast growth of IT infrastructure, Infrastructure as Code (IaC) becomes the trend for most of the companies and domains to provisioning their infrastructure. Domains such as embedded system for medical sector, marine safety, transportation networks, electricity networks, and other are depending on IaC to build their infrastructure. By implementing our suggested security framework for IaC, these sectors could be improved and secured to be resistant to infrastructure attacks.

DevOps is a development methodology that enhances agile concepts to resolve communication conflicts between development, test, security, operations, and other teams. Improving communications maximizes efforts through achieving the continuous integration and continuous delivery (CI/CD). Software delivery services such as Software as a Service (SaaS) make DevOps a vital methodology [1]. To achieve CI/CD, the DevOps process applies automated tasks such as packaging of the application with its internal libraries and dependencies, containerization, and provisioning of the production and test environments. It is noted in ISO/IEC/IEEE 32675 [2] that automation is one of the pillars of DevOps. IaC ensures that the deployment process and infrastructure creation are more consistently controlled and can be easily replicated [3]. Consider particularly files specifying environment settings such as libraries or RAM for each virtual machine (VM) [4] which can be used to consistently create environments. Infrastructure as Code is based on practices from software development emphasizing consistent, repeatable routines for provisioning and changing systems and their configuration. In automating tests and changes to your systems, reliability is greatly improved [5]. The need for this technique has stemmed from the rapid growth of cloud computing, with particular emphasis on programmatic provisioning, configuration, and management of computational resources [6] in increasing the speed of obtaining resources.

The speed of development of this technique has resulted in many open challenges. Sharing IaC security and measurements knowledge from the beginning of the project for example. It has been argued [7] that DevOps shares the tools but doesn't share the knowledge. However, the ISO/IEC/IEEE 32675 DevOps standard [2] recommends the use of techniques and tools for management of knowledge. Lack of consistency in the application of IaC remains as the standard is not well advertised.

Security development life cycle frameworks (SDLs) are considered as a reference for all IT companies to adopt a reliable security practice. These frameworks determine the best security practices and guidelines in every stage along the software development life cycle (SDLC). Due to the rapid evolution of development tools and techniques as well as the long duration of time required to establish security frameworks, these security frameworks may miss much about the new techniques in the software life cycle. They require tailoring the new techniques in the frameworks body. At the time of writing security frameworks don't cover the security practices and tests that should follow when using IaC. The solution for IaC is to shift security aspects and practices to the early stages of the SDLC. Applying DevSecOps methodology to the IaC SDLC will make it more secure. DevSecOps emphasizes



**Fig. 1** Deployment process using traditional (unsecure) IaC process

the focus on integration of security process and practices into DevOps pipeline and environment [8]. DevSecOps left shifts the security's practices and actions to the beginning of the software development life cycle, in a way that security become the responsibility of all the team members. DevSecOps employs, for example, Security as Code to inject automated security practices in the pipeline by finding places to add security checks and tests without introducing unnecessary costs or delay [9]. Figure 1 shows the traditional process to deploy IaC. The lack of consideration for security during the process of IaC is evident.

## 2 Research Aims and Objectives

IaC's risks begin from the inside the organization or development environment, represented as a poor security practice including, policies violations, to security attacks through environment boundaries. Left shifting security principles to IaC's code and practices will make deployment not only more secure but more stable and safer. This requires understanding the nature and logic of IaC to identify and remediate security risks and vulnerabilities.

IT organizations use security frameworks to combine the valuable security practices and guidelines that the organization should follow in order to secure the product. Security frameworks aren't completely perfect, and sometimes failing to take into account risks associated with new technologies. Some frameworks need two years or more to be established, so a lot of new techniques may be missed in their specification. As mentioned in [10]: "some of the practices that are good today, are not necessarily the best option for tomorrow". Many security frameworks only advice regarding IaC surrounds how it can be used to achieve automation. They fail to provide information or recommendations surrounding threats and vulnerabilities that could arise during IaC process. Similarly, they often fail to consider vulnerabilities that could appear as a result of the use of IaC. In addition, there isn't an obvious list for the potential threats or risks when the organization configure their infrastructure by IaC.

A further challenge is the lack of consideration of identifying the security factors and metrics which related to IaC's code and processes. These risks are compounded by the poor use of IaC resulting in ghost resources, snowflake servers and more. IaC requires a security pipeline to make the process of infrastructure creation more secure.

The following immediate questions should be addressed in order to examine potential solutions to this problem:

- How can the Security as Code practices for IaC be used to fill the gap in security frameworks?
- What security threats and vulnerabilities exist in IaC life cycle?

### 3 Related Work

Research has been completed in relation to IaC focused on IaC including analysis by Rahman [11], of 33,887 publications where he concludes that IaC's tools and frameworks are well researched but there is a lack of studies of the defects and security flaws related to IaC as a technique. It is noted that empirical studies that focus on test coverage, test practices, and testing techniques are needed for IaC [11]. Guerriero et al. conducted 44 interviews to discover the level of IaC adoption. He noted that over 64.3% of industry practitioners have less than 5 years of experience developing IaC, which reflects that IaC is a relatively new trend. In addition, he mentioned that available tools are still limited with developers feeling the need for new techniques for testing and maintaining IaC [6]. To improve the quality of IaC via embedded AI and machine learning inside IaC pipeline, Palma et al. [12] proposed a framework to compare five machine learning methods (Decision Tree, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine) to build a model to identify the defects in Ansible-based IaC scripts. The model focused on product and process metrics. Gasiba et al. [13] used the Tec tool to provide documentation of Terraform's code, the motivation behind that is the lack of standardized secure guidelines for the terraform programming language.

Research studies have also looked specifically to the IaC's code quality. Hasan et al. [14] categorize six practices to improve the quality of IaC scripts. The practices are avoiding coding anti-patterns, remote testing, sandbox testing, testing each IaC change, and the use of automation. These practices look to improve IaC quality, but they focus only on the implementation side of IaC pipeline. There is a need to generalize these practices and to add other practices to cover IaC testing pipelines from the outset. Rahman et al. [15] have presented a defect taxonomy for IaC scripts. He analysed 1448 defect-related commits collected from open-source repositories. The defects were mapped to eight categories [15]: conditional, configuration data, dependency, idempotency, security defects, documentation, service, and syntax. Palma et al. [16] also proposed a catalogue composed of 46 measures that identify quality IaC properties for Ansible where 8 metrics are related to language-agnostic code characteristics, 14 metrics have been adapted from a previous study [17], and 4 metrics

concern some inherent characteristics of Ansible and are founded in the other IaC languages.

Most studies related to the security of IaC have focused on code smells and static analysis of IaC. Seven security smells in puppet scripts are identified in [18]. They collected 1726 puppet scripts and implement a static analysis tool called Security Linter for Infrastructure as Code scripts (SLIC) to identify the occurrence of each of the smells [18]. The smells identified include admin by default, Empty password, suspicious comment, hard-coded secret, use of HTTP without TLS, invalid IP address binding, and use of weak cryptography algorithms. Schwarz et al. studied Chef scripts [19] categorizing code smells into three kinds as Technology Agnostic Smells, Technology Dependent Smells, and Technology Specific Smells. Bhuiyan et al. analysed 7222 puppet scripts to quantify the correlation between source code metrics and IaC scripts that contain Insecure coding patterns [4]. The result was that at least 21.06% of the scripts have one instance of ICPs and the most frequently occurring co-located ICPs are hard-coded secret, suspicious comment, and using HTTP without TLS/SSL. The security risk is evident from these brief examples.

## 4 Infrastructure as Code Security

IaC reinforces the benefits of the cloud services making infrastructure's status or resources easier to access and more available. The benefits come from the nature of IaC that allows the organizations to develop, test, and deploy this code using the same engineering practices for application development, such as version control systems and automated testing, so practices such as test-driven development and continuous integration should apply for IaC [12]. IaC automation scripts put the system into specific state [20] making it easier to test for compliance and conformance to regulations.

### 4.1 *Infrastructure Configuration Problems and Threats*

One of the largest IaC challenges arises from use of IaC comes from low level mistakes such as hard-coded credentials or secret exfiltration through logs. It is not enough to focus on the quality of IaC. As IaC is a technique cover a wide range of concepts and components such as VMs, containers, and infrastructure provision, problems related to those concepts and components should discussed and modelled when using IaC. Besides that, currently there are no standards or technology supported DevOps processes that explicitly and frequently reference the need to consider the right shift direction in the development culture (e.g. Ops to Dev) [21].

Whilst this paper could not list all relevant security threats common threats that could arise as a result of misuse of IaC include

- Intra-update sniping vulnerability: this can happen in two scenarios [22]. First, during updating of the back-end or infrastructure of the software when the private data can become public. The second one happens when components are added or removed from an infrastructure without ordering constraints.
- Configuration Drift: configuration drift is where the system changes from the desired state [21] often due to ad-hoc changes. Configuration drift makes it harder to maintain consistent systems [5]. Changes or updates outside of IaC, including opening ports, upgrading or downgrading software packages, failing to deploy structural changes, or modifying the scaling policy of virtual resources results in configuration drift. It is important to understand the reason for configuration drift to adapt cybersecurity practices to tackle the problem. Using an IaC tool like Terraform does not make environments immune to configuration drift by itself [23]. Configuration drift related to deployment strategy can happen when engineers perform small changes to experimental or temporary environments, sometimes to blue environments in Blue/Green deployment model, without transmit these changes to the IaC. In IaC, configuration drift can be the result of system faults or threats, from isolation of a server from the network, as zombie server or infrastructure breakdown.
- Shadow admins: from Microsoft server perspective, a shadow admin is a user, who is not a member of the active directory administrative groups and doesn't belong to any admins group including Domain Admins, Enterprise Admins, Schema Admins, and administrators, but still have the rights to perform some of administrative capabilities. If a shadow admin account is compromised, it can be very risky for the organization [24]. In IaC context, shadow admins could make change, which is not transparent, which may lead to a gap between the domain and controller specification and the current infrastructure state. Shadow administrators can be created through the misconfiguration of policies for users, mismanagement of user rights, temporarily created administrator accounts which are not removed, and attackers create them if they attain sufficient privilege. Attackers and malicious insiders can exploit these shadow accounts to reach critical infrastructure, alter sensitive data, or interrupt cloud-hosted service [25].
- Infrastructure sprawl and consolidation: system sprawl happens when a system becomes so complex that administrators lose track of its architecture or documentation. Two types of infrastructure sprawl include virtual sprawl and server sprawl. Virtual sprawl happens when the number of virtual machines exceeds the administrator's capacity to adequately manage them. Security issues remain consistent with those from physical machines [26]. Server sprawl happens when idle servers consume more resources than is necessary for their task [27]. Consolidation, one of the solutions to server sprawl, involves running several workloads on a shared collection of nodes [28]. Consolidation is not always an appropriate solution due to requirements of server-based functionality, time needed to consolidate existing infrastructure, etc. [29]. Server consolidation also leads to physical,



logical changes, and architecture changes which can lead to further drift. The authors have not found any studies on the impact, or the threats consolidation could bring with relation to Infrastructure as Code. Additionally, the migration of VMs should be managed by IaC as migration problems such as hyperjacking need to be considered.

- **Visibility and observability:** To decrease infrastructure upgrade times, deployment tools typically will run many operations in parallel. Poorly managed parallelism in infrastructure can lead to a violation of the intended security policy and other discrepancies. Due to complexity, the traceability may be difficult [22]. Visibility is one of the pressing open challenges in IaC, especially with the hybrid networks found with SDN with cloud. Lack of visibility is a risk. Poor visibility of the processes which run inside the container or ephemeral machines further compounds the issue. Some processes inside the containers will be like black box [30]. Containers make deploying, packaging and distribution microservices simple; however, there are security concerns about the observability into inter-service communications over the complex networks.
- **Attack surface:** containers, VMs, and microservices dramatically increase the attack surface of the infrastructure. Containers encompass the entire containerization toolchain.

There are many other IaC threats that should be considered, which IaC inherits from virtualization and containerization. The threats and security issues outlined should be considered early in the IaC life cycle. They should be mapped to the healthy code practices, so the script development team could avoid them. Additionally, tests and checks should be added to IaC creation process to check every change in the infrastructure or the code. The challenge will be the cost of these tests and the impacts on the delivery time just as it is a challenge for all software quality processes.

## ***4.2 IaC and Security Development Life Cycle (SDL)***

The security framework provides a baseline of observed activities (controls) for software security initiatives (SSIs) to build security into software and software development. In this paper, we evaluate two frameworks, Building Security in Maturity Model (BSIMM) representing a mature security framework, and the other, BSA representing the secure development life cycle frameworks.

There are additional security frameworks that fall into the main categories such as the NIST Cybersecurity Framework under the heading of a mature security framework, the NIST Secure Software Development Framework (SSDF), and the Microsoft SDL under the heading of SDLs, and finally security frameworks for risk management such as NIST Risk Management Framework (RMF) and HITRUST Common Security Framework. Here we show how BSIMM and BSA fail to consider IaC security implications. The other frameworks or standards such as ISO/IEC 27001

will be considered in a future paper, which will focus on each process in the proposed approach and tailored it to these frameworks.

**BSIMM framework alignment with IaC.** BSIMM framework is a mature security framework aiming to cover all security aspects related to the software process and the Secure Software Development Life Cycle (SSDL). In IaC context [31], BSIMM recommends the use of IaC but fails to go deeply into the impact of using this technique. In Configuration Management and Vulnerabilities Management (CMVM) practices, IaC is considered solely as a solution without placing it in the security life cycle. This goes against the DevOps principles which prize visibility, sharing of knowledge, transparency across tools and techniques throughout the software life cycle. Table 1 shows BSIMM Governance practices and their alignment with IaC. These are classified under strategies and metrics (SM), compliance and policy (CP), and training (T). The number of examples presented here is limited for this paper but extends greatly across BSIMM.

When considering BSIMM intelligence, we view practices of attack models (AM), security features and design (SFD), and standards and requirements (SR). They are given in Table 2. In the third BSIMM example SSDL, we consider architecture and analysis (AA), code review (CR), and security and testing (ST). Refer to Table 3.

In the next examples from BSIMM on deployment, we consider software environment (SE) in Table 4, configuration management and vulnerability management (CMVM) in Table 5, and penetration testing (PT) in Table 6.

**BSA Framework alignment with IaC.** The BSA Framework identifies best practices relating to both organizational processes and product capabilities across the entire

**Table 1** BSIMM Governance evaluation from IaC security perspective

Practice	Activities	Alignment with IaC
(SM)	[SM2.3] Create or grow a satellite (security champions)	[SM2.3] The selection of security champions should also consider IaC knowledge. Coding knowledge alone is insufficient. Securing the whole pipeline of IaC is needed. Research shows that there is currently a lack of the experience of IaC in the organizations [11]
(CP)	[CP3.2] Ensure compatible vendor policies	[CP3.2] Tests built directly into automation or infrastructure to ensure that vendor software security policies and SSDL processes are compatible with internal policies. Current IaC activities often fail to consider this
(T)	[T2.8] Create and use material specific to company history	[T2.8] IaC developers orchestrating containers aren't always aware of existing virtualization problems. Training of the IaC team should cover defects that could generate from IaC tools in addition to the traditional security issues and how to mitigate them in IaC environment

**Table 2** BSIMM intelligence evaluation form IaC security perspective

Practice	Activities	Alignment with IaC
(AM)	[AM1.3] Identify potential attackers	[AM1.3] Identifies potential attackers, attack surface, theoretical internal attackers, and contract staff. As containerization and virtualization add complexity to the system, the attack surface will be impacted by them and attackers on IaC could be different from the well-known attackers
	[AM2.1] Build attack patterns and abuse cases tied to Potential attackers	[AM2.1] There is lack of discussion surrounding modelling security issues related to IaC. There is no clear mapping study or list that maps traditional security threats with the use of IaC. IaC still relatively modern technique. IaC inherit the security issues from different sources as virtualization, containers, cloud, edge computing. The security of IaC process is not sufficiently researched
(SR)	[SR3.3] Use secure coding standards	The logic and the effects of IaC are different from many other languages as it works directly with the infrastructure. IaC need further standards and techniques focusing on areas such as refactoring and model-driven tests. Therefore, we should build a standard for IaC secure coding

**Table 3** BSIMM SSDL evaluation from IaC security perspective

Practice	Activities	Alignment with IaC
(AA)	[AA1.2] Perform design review for high-risk applications	[AA1.2] Perform a design review to determine whether the configuration is resistant to attack. Configuration drift could happen via misuse or attack on the IaC. The impact could be temporarily hidden, i.e. zombie servers. Configuration drift detection should be included during “secure by design” of components supported by the continuous detection process
(CR)	[CR1.4] Use automated code review tools	[CR1.4] Static analysis should be incorporated into the IaC review process to improve efficiency and consistency. There is a variety of IaC tools which can be problematic when trying to achieve consistency. Whilst studies try to find security flaws in the IaC, there is no guideline that standardization of the tools, how they work, or their efficacy
(ST)	[ST1.3] Drive tests with security requirements and security features	It may be considered that many companies don’t consider IaC security strongly in their security requirements. This results in gaps due to incomplete security requirements
	[ST1.4] Integrate security tools into the QA process	The quality, efficiency, and efficacy of IaC testing tools have not been studied in depth at this point

**Table 4** BSIMM deployment evaluation form IaC security perspective

Practice	Activities	Alignment with IaC
(SE)	[SE1.1] Use application input monitoring	[SE1.1] Mention that cloud deployments and platform-as-a-service usage can add another level of difficulty to the monitoring, collection, and aggregation approach. However, it does not mention specify how to raise the visibility of the system
	[SE2.2] Define secure deployment parameters and configurations	[SE2.2] Create deployment automation or installation guides to help teams and customers install and configure software securely. The failure to update D&C logs may lead to configuration drift [21]. The IaC guidelines should specify how to ensure consistency
	[SE2.5] Use application containers to support security goals	[SE2.5] Recommend the use of containers due to inherent security benefits. However, this brings challenges such as dependency problems, often referred to as dependency hell. Given the prevalence of containerization in IaC, this needs further consideration
	[SE2.7] Use orchestration for containers and virtualized environments	[SE2.7] Orchestration processes take advantage of built-in and add-on security features at the end of activity. Orchestration tools, such as Kubernetes, become a part of development environment, which in turn requires hardening and security patching and configuration

**Table 5** BSIMM configuration management evaluation form IaC security perspective

Practice	Activities	Alignment with IaC
CMVM	[CMVM2.1] Have an emergency response plan	[CMVM2.1] Organizations often make fast code and configuration changes when software is under attack which can prove dangerous for IaC. Rollback to a known-good state could be a good solution but pushing new code or deploying a new container may lead to other defects. Lack of visibility in decision making can be caused by urgent changes. It could also result in configuration drift [23]. Fast code and configuration changes could solve current problem but could lead to others
	[CMVM3.6] Publish risk data for deployable artefacts	[CMVM3.6] could be extended to use telemetry automation to publish risk information about the applications, services, APIs, containers, and other software it deploys. IaC tools are relatively new such that observability and telemetry of these tools is still not sufficiently studied particularly with regard to performance [11]
	[CMVM3.8] Do attack surface management for deployed applications	[CMVM3.8] In applications consisting of microservices, containers, VMs, SDN architecture and IaC, there is no clear framework to identify the attack surface. A combination of tools may be deployed to cover each technique each with some flaws making the attack surface wider

**Table 6** BSIMM penetration testing evaluation form IaC security perspective

Practice	Activities	Alignment with IaC
(PT)	[PT3.1] Use external penetration testers, i.e. red team to perform deep-dive analysis	[PT3.1] The domain of attacks is an outstanding problem as there is no standard or guidelines for attacks directly resulting from IaC. Thus, the red team should try to simulate attacks such as linking the misconfiguration exploration problem and the resulting vulnerabilities. Attacks can then be based on these vulnerabilities

software life cycle. It is organized into six columns: Functions, Categories, Subcategories, Diagnostic Statements, Implementation Notes, and Informative References. Table 7 provides examples on the impact of access management on the IaC perspective. Table 8 shows a selection of the key alignment points between IaC and the BSA Framework.

Table 9 provides examples of how BSA can be applied to areas such as testing and verification.

**Table 7** BSA development environment and identity access management evaluation from an IaC security perspective

Sub-category or diagnostic statement	Implementation	Weakness to IaC
DE.2-6. Containers and other virtualization technologies used in deploying the software use secure configurations	NIST SSDF/PO.3.2: Follow recommended security practices to deploy, operate, and maintain tools and toolchains	Although the use of IaC is recommended, there is no discussion of security issues surrounding IaC usage. SDLs consider the use of IaC sufficient for the Dev & Ops without further consideration of the risk. This issue is indicated in many sections of BSA and BSIMM
IA.2. Policies to control access to data and processes for all users and operators are developed, documented, and applied throughout the development environment	IA.2-2. Access controls are set for individual users and operators that provide only the necessary privileges required to perform an assigned task and only for the necessary time required to perform it	Lowest privileges should be applied to IaC particularly during provision. Policy as Code must combine with IaC to ensure that only authorized people reach critical infrastructure. Compliance with many policies requires input from IaC and IaC generated logs

**Table 8** BSA secure coding evaluation from an IaC security perspective

Sub-category or diagnostic statement	Implementation	Weakness to IaC
SC.1. Threat modelling and risk analysis are employed during software design to identify threats and potential mitigations	Threat modelling even with zero trust model	Risk management framework such as NIST (RMF) that is used in threat modelling process is needed to tailor with IaC
SC1.2. Threats are rated and prioritized according to risk		Any failure or unauthorized access to the infrastructure is a high risk. Mistakes in configuration for a server instance will lead to hard consequences. IaC and infrastructure risk should be considered as high risks
SC.3-1. Software avoids or includes documented mitigations for known security vulnerabilities	Software should avoid known vulnerabilities in included functions and libraries	IaC is considered as a relatively new technique with no obvious list for the threats in these scripts
SC.4-1. The software employs segmentation through sandboxing, containerization, or similar methodologies		Containers which share the same host operating system, a centralized point of failure, can make security more complicated. Risk management and threat modelling should be carried out when adopting new techniques including IaC. Visibility inside the containers and the dependencies may increase the attack surface. This became painfully evident with the Log4J attack of 2021

### 4.3 Security Approach for IaC

The security of the IaC process requires building a secure pipeline for IaC. The pipeline should include left-shift of security concepts in line with DevOps pipelines. Whilst there are many terms to consider, including DevOps, DevSecOps, ComplianceOps, research has shown [32] that this is simply a misunderstanding of the term DevOps. It is the preference of people to emphasize their aspect of project under development.

The proposed pipeline comprises the necessary security practices including clear justification of the choice of tools depending on each team and project as well as directly related concepts that both feed into and feed from this pipeline. The proposed security approach for IaC is shown in Fig. 2.

The general security practices for each phase in the approach include

- **Risk management and Threat modelling:** In DevSecOps, risk management procedures and activities should be used consistently throughout the life cycle [33]. In IaC, risk management is not well highlighted. Threat modelling is a process for identifying, communicating, and comprehending threats and mitigations in the

**Table 9** BSA testing and verification from an IaC security perspective

Sub-category or diagnostic statement	Weakness to IaC
TV.1. Analysis and validation of the software attack surface is conducted	Cloud services attack surface can be shared with other (vendors), which means sometimes that external paths could be out of our control. These paths can reach your infrastructure. The tools to identify IaC attack surface work on specific sectors, some for VMs and others for containers. The different tools do not integrate well if at all, to define the combined attack surface of a system. Further issues include lack of knowledge sharing. Compliance and polices management rely heavily on both this information sharing and the use of IaC tools
TV.2. Code review using manual and/or automated tools is conducted	The logic of IaC is different than other code. The code may look clean but can take the system to a drift state or isolate some infrastructure instances through faulty configuration settings. Code review is necessary, but it is not enough without clear structure as to common issues found in IaC
TV.3-2. The software is tested in a least privilege environment	Minimum privileges should be used in production. However, with IaC elevated permissions are needed on an intermittent basis. Careful review of elevation requirements should be carried out
TV.5-2. Software is subjected to adversarial testing techniques, including penetration testing	Fuzz tests on production requires a copy of the environment. This is time-consuming. A/B testing could help in this context

- context of defending something of value which includes infrastructure. Microsoft SDL threat modelling process [34] is not sufficient for cloud, hybrid, or multi-cloud infrastructure. Identifying the weakest components in the architecture and the causes of threats, attacks, and the flaws should be carried out. This includes modelling of security issues related IaC technique, IaC tools limitations, IaC creation risks, misuse of IaC, and poor programing practices. Assessment of vulnerabilities should be carried out continuously. System Logs files could include in this process to scan them about any security flaws.
- Another challenge in IaC is identifying the entry point for the attack. A hidden infrastructure or a zombie server could be the start point for attack and the IaC process could be blind about this component. This type of vulnerability should be considered in the threat modelling process of IaC. Increasing IaC process visibility helps to address hidden vulnerabilities. Dependency flows between system components should be created along with other risk management practices for IaC are shown in Figs. 3 and 4.



**Fig. 2** Proposed security approach for IaC creation

**Fig. 3** Security planning for IaC

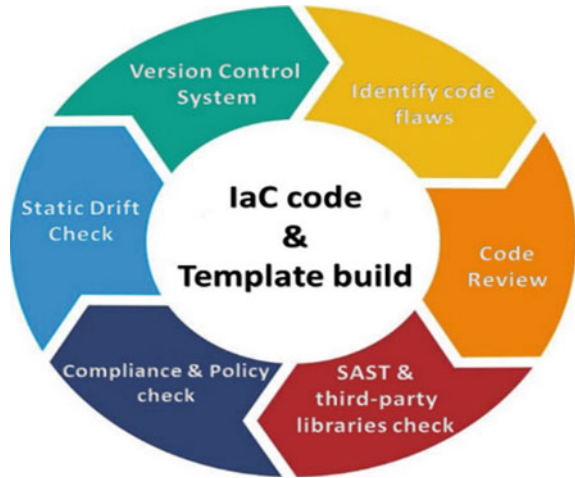




**Fig. 4** Secure design for IaC

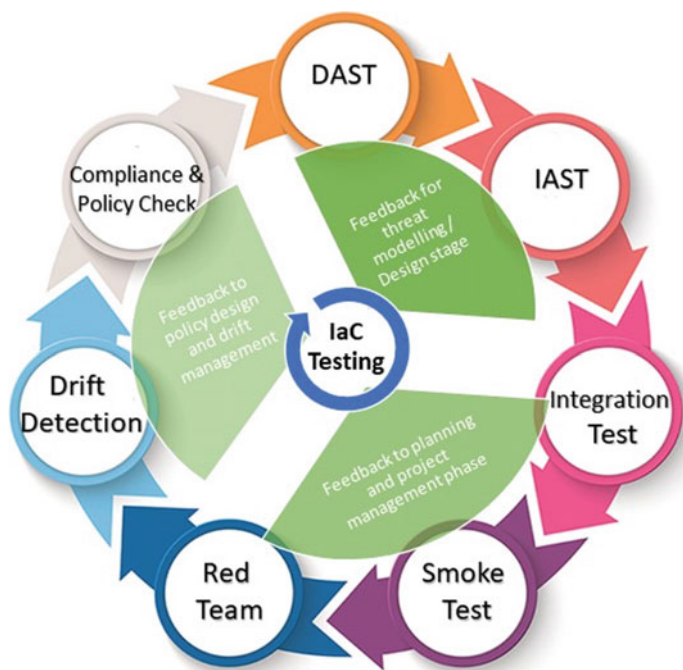
- Static application security testing (SAST) and IaC: SAST, a method for analysing programmes without running them, is useful for secure programming standards [35]. The challenge in using SAST with IaC lies in the lack of test cases for IaC. A further challenge is the lack of agreed rules for writing clean, SOLID, and secure code. Where many studies related to IaC try to discover related code smells, the variety of IaC languages and tools makes it necessary to first establish core principles for IaC to enable SAST. Applying the principles of traditional programming languages is useful, but lacks aspects such as issues related server templating or infrastructure provisioning. SAST is still necessary for IaC, but the type of the threats in IaC or infrastructure created by IaC is different than the concerns of traditional programming language issues. The importance of dependency checks increases as it relates to components from which the configuration file is built and dependencies of the components under construction. SAST should check all internal and external libraries, especially for containers and dependencies between network components. Figure 5 shows some coding and building security practices for IaC.
- Dynamic application security testing (DAST) and IaC: DAST implements black box testing of the runtime behaviour while running the application from the outside in. It can detect conditions that lead to a security flaw while the application is running. The cost and time of running DAST for infrastructure should be estimated in a different way from typical usage, as it tests critical components in the systems. Security test cases for IaC should be modelled and identified carefully to cover all vulnerabilities in IaC. It should be noted that the impact of DAST for IaC is not adequately considered in literature yet. Interactions test between containers and

**Fig. 5** Security practices for IaC build and code phases



network components must be carried out through DAST to avoid the “Dependency hell” problem.

- **Interactive Application Security Testing (IAST) and IaC:** IAST uses a specific design context; it combines static application security testing (SAST) and dynamic application security testing (DAST); it gives the ability to monitor and analyse code as it executes [36]. IAST could be achieved by an automated test, human tester, or any activity “interacting” with the application functionality. It gives the ability to monitor the interactions between the code components to find security vulnerabilities [37]. Applying IAST on IaC will bring a lot of benefits as it can cover and test the interactions between the containers and VMs in real time. IAST gives the advantage that it discovers the path of the security flaws rather than DAST. Security practices for IaC’s testing stage are shown in Fig. 6.
- **Adding security runtime layer and IaC:** Run-time Application Security Protection (RASP) can examine incoming traffic and content similarly to firewalls and decide when to end sessions. RASP can act as firewall for IaC so any suspicious action could be prevented in the production environment, but the rules of IaC should be built first, with RASP established on these rules. Graph flow is one good tool for building the rules and transferring them as test cases in RASP.
- **Integration tests and IaC:** Integration tests should take place for every change of the infrastructure or the network. A/B test strategy is one useful testing strategy with IaC. Figure 6 shows the necessary practice in the test phase for IaC.
- **Policy as Code:** The term “Policy as Code” refers to an approach of managing policies in which code is used to develop, modify, share, and enforce policies [38]. Using Policy as Code gives DevOps teams the ability to employ a “policy by design” approach for conformance. Due to the rapid rate of change and deployment that IaC offers, employing policies for Infrastructure as Code ensure that changes or updates that violates policies are not allowed into the main codebase.



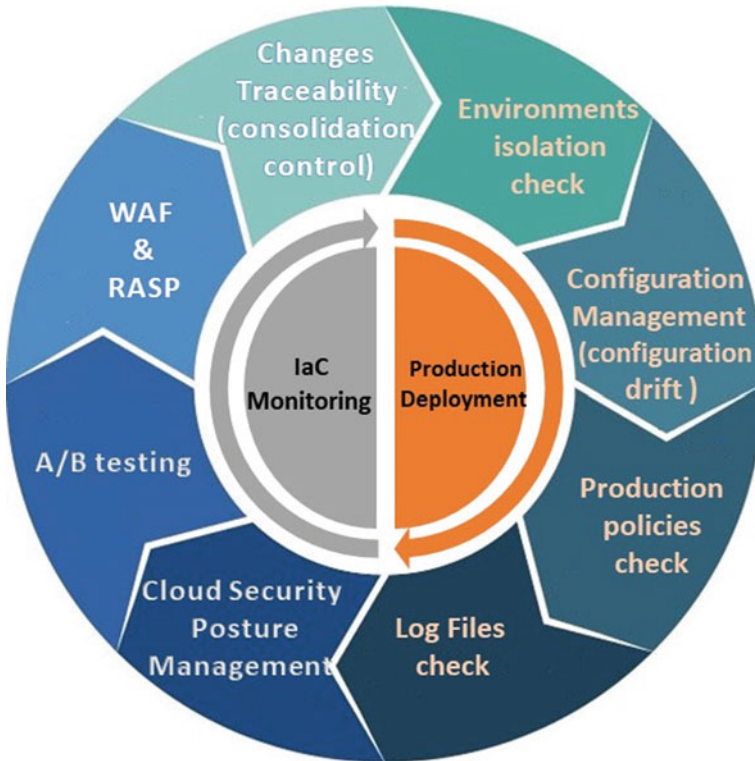
**Fig. 6** Security practices for IaC testing phase

By applying Policy as Code to the configuration in production, the ability to identify potential inconsistencies regardless of the source of the behaviour.

- Postproduction Security and IaC: IaC tools produce logs files and other telemetry when deploying or configuring the environment. Mining these logs is useful to find patterns of errors or flaws in the wider ecosystem. Code obfuscation is a good defence strategy to keep the source code of IaC safe against attacks. Secrets management practice should also apply on the IaC assets. Continuous monitoring as recommended in DevOps is useful to determine behaviour and abnormal actions. Figure 7 outlines some of useful security checks for IaC postproduction.

## 5 Limitations

More real-world testing and experiments are needed to validate the suggested framework because it is still in the conceptual stage. Surveys and interviews with stakeholders who collaborate closely with IaC should be a part of the validation process. If any of our readers would like to take part in the interviews, we encourage them to get in touch. The variety of commercial and open-source security tools used in DevOps pipeline makes it challenging to choose the best tools for the approach; therefore, an assessment of these tools' effects on the suggested approach should be



**Fig. 7** Security practices for IaC monitoring and postproduction

made to determine the most effective toolchain for IaC security. In order to deploy the strategy in a real DevOps/IaC culture, industrial data will be essential. The next analysis of the approach should consider the volume and nature of the IaC’s data, teams, and companies. This work focuses on IaC, and the domain of the approach has considered security of IaC-related technologies such as virtualization, containerization, and microservices into consideration. Therefore, it does not cover the security of other DevOps practices or stages.

## 6 Conclusion

In this research, a security framework for Infrastructure as Code process has been presented, and the aim for the proposed approach is to make the process of creating Infrastructure more reliable and secure and to achieve continuous security for IaC.

Generally, IT organizations use security frameworks to identify the required security practices in their software life cycle. Two security frameworks have been

discussed to figure out the outcomes of IaC in these frameworks. The absence of IaC security is notable in these frameworks. The two focused security frameworks recommended IaC as a technique to create and configure the infrastructure without considering that the IaC process or codes could be hacked or misused in a way that could bring a lot of risks to the organization. To raise the security of IaC, a number of threats and security risks related to IaC as a technique have been identified. It is clear that IaC, as any new technique, will bring the solution and challenges. These threats give an example that IaC has special type of threats that should be addressed carefully during threat modelling activities. IaC also inherits the threats and vulnerabilities from concepts like virtualization and containerization. An example of this inheritance was shown in the research as intra-update sniping vulnerability. To mitigate the absence, IaC should be put in a parallel pipeline and the security of it must start from the beginning of the project. The proposed framework is building a security pipeline for IaC and injecting security practices from the planning stage and continue after the production. The integration between IaC, Policy as Code, and compliance as code will enhanced IaC reliability and will be a defending line against the vulnerabilities that arise from policies and compliance violation. IaC's risk management should cover VMs, containers, SDNs, and other techniques which are related to IaC issues.

In future work, each phase of the proposed framework should be studied thoroughly to define the right security activities that could cover the IaC vulnerabilities. A list of IaC threats and vulnerabilities, similar to OWASP top ten list, should be built in a way it makes it like a security reference for all the teams that used IaC. Specification and description for all the phases and their security practices will be established in a way make this proposed approach like add-on for other SDLs frameworks.

## References

1. Senapathi M, Buchan J, Osman H (2018) DevOps capabilities, practices, and challenges: insights from a case study. In: EASE'18: Proceedings of the 22nd international conference on evaluation and assessment in software engineering 2018, Christchurch, New Zealand
2. I. C. Society (2021) IEEE Standard for DevOps: building reliable and secure systems including application build, package, and deployment. IEEE
3. Leite L, Rocha C, Kon F, Milojevic D, Meirelles P (2019) A survey of DevOps concepts and challenges. *ACM Comput Surv* 25(6), article no. 1
4. Bhuiyan FA, Rahman R (2020) Characterizing co-located insecure coding patterns in infrastructure as code scripts. In: 35th IEEE/ACM International conference on automated software engineering workshops (ASEW)
5. Morris K (2021) *Infrastructure as code, dynamic systems for the cloud age*, 2nd edn. O'Reilly Media, Inc.
6. Guerriero M, Garriga M, Tamburri D, Palomba F (2019) Adoption, support, and challenges of infrastructure-as-code: insights from industry. In: IEEE International conference on software maintenance and evolution (ICSME)
7. Mendes N (2017) DevOps maturity model report: trends and best practices in 2017 [Online]. Available: <https://www.atlassian.com/blog/devops/devops-culture-and-adoption-trends>. Accessed 6 July 2022

8. Myrbakken H, Colomo-Palacios R (2017) DevSecOps: a multivocal literature review. In: International conference on software process improvement and capability determination, Palma de Mallorca, Spain
9. Bird J (2016) DevOpsSec, 1st edn. O'Reilly Media, Inc., Sebastopol, CA, pp 29–69
10. Matthee MH (2014) Secure software development framework: principles and practices. SANS Institute
11. Rahman A, Hezaveh RM, Williams L (2019) A systematic mapping study of infrastructure as code research. *Inf Softw Technol* 108:65–77
12. Palma SD, Nucci DD, Palomba F, Tamburri DA (2022) Within-project defect prediction of infrastructure-as-code using product and process metrics. *IEEE Trans Softw Eng* 48(6):2086–2104
13. Gasiba T, Cristian L, Lechner U, Albuquerque M (2021) Raising security awareness of cloud deployments using infrastructure as code through CyberSecurity challenges. In: International conference on availability, reliability and security
14. Hasan MM, Bhuiyan FA, Rahman A (2020) Testing practices for infrastructure as code. In: LANGEI 2020: proceedings of the 1st ACM SIGSOFT international workshop on languages and tools for next-generation testing
15. Rahman A, Farhana E, Parnin C, Williams L (2020) Gang of eight: a defect taxonomy for infrastructure as code scripts. In: IEEE/ACM 42nd International conference on software engineering (ICSE), Seoul, Korea (South)
16. Palma S, Nucci DD, Palomba F, Tamburri DA (2020) Toward a catalog of software quality metrics for infrastructure code. *J Syst Softw* 170
17. Rahman A, Williams L (2019) Source code properties of defective infrastructure as code scripts. *Inf Softw Technol* 112:148–163
18. Rahman A, Parnin C, Williams W (2019) The seven sins: security smells in infrastructure as code scripts. In: IEEE/ACM 41st International conference on software engineering (ICSE)
19. Schwarz J, Steffens A, Lichter H (2018) Code smells in infrastructure as code. In: 11th International conference on the quality of information and communications technology, Coimbra, Portugal
20. Hummer W, Rosenberg F, Oliveira F, Eilam T (2013) Automated testing of chef automation scripts. In: Proceedings of the demo & poster track of ACM/IFIP/USENIX international middleware conference, Beijing, China
21. Jiménez M, Castaneda L, Villegas N, Tamura G, Müller H, Wigglesworth J (2019) DevOps round-trip engineering: traceability from dev to ops and back again. In: International workshop on software engineering aspects of continuous development and new paradigms of software production and deployment
22. Lepiller J, Piskac R, Schäfer M, Santolucito M (2021) Analyzing infrastructure as code to prevent intra-update sniping vulnerabilities. In: Tools and algorithms for the construction and analysis of systems TACAS 2021
23. Ciobanu R (2022) Infrastructure as a code—why drift management is not enough [Online]. Available: <https://www.infoq.com/articles/iac-configuration-drift/>. Accessed 29 Oct 2022
24. Keshet Y (2022) The hidden dangers of shadow admins [Online]. Available: <https://www.silverfort.com/blog/the-hidden-dangers-of-shadow-admins/>. Accessed 29 Oct 2022
25. L.CyberArk (2022) Shadow admins [Online]. Available: [https://docs.cyberark.com/Product-Doc/OnlineHelp/CEM-SharedServices/Latest/en/Content/CloudAdmin/kd\\_shadow-admin-per-platform.htm](https://docs.cyberark.com/Product-Doc/OnlineHelp/CEM-SharedServices/Latest/en/Content/CloudAdmin/kd_shadow-admin-per-platform.htm). Accessed 3 Oct 2022
26. Djenna A, Batouche M (2014) Security problems in cloud infrastructure. In: International symposium on networks, computers and communications, Hammamet, Tunisia
27. Thakur S, Kalra A, Thakur J (2013) Server consolidation algorithms for cloud computing environment: a review. *Int J Adv Res Comput Sci Softw Eng* 3(9)
28. Higginson A, Bostock C, Paton N, Embury S (2022) Placement of workloads from advanced RDBMS architectures into complex cloud infrastructure. In: 25th International conference on extending database technology (EDBT 2022), Online, UK

29. Garfias J (2022) How to eliminate server sprawl [Online]. Available: <https://www.printerlogic.com/blog/how-to-eliminate-server-sprawl/>. Accessed 10, 2022
30. Tapia F, Mora MA, Fuertes W (2020) From monolithic systems to microservices: a comparative study of performance. *Appl Sci J* 10(17)
31. B. Community (2022) BSIMM framework [Online]. Available: <https://www.bsimm.com/framework.html>. Accessed 1 Oct 2022
32. Lennon R (2021) DevOps best practices in highly regulated industry. In: Proceedings of seventh international congress on information and communication technology, Singapore
33. Schmittner C, Griessnig G, Ma Z (2018) Status of the development of ISO/SAE 21434. In: 25th European conference, EuroSPI 2018, Bilbao, Spain
34. MicrosoftSDL (2022) Threat modeling [Online]. Available: <https://www.microsoft.com/en-us/securityengineering/sdl/threatmodeling>. Accessed 8 Aug 2022
35. Brucker AD, Sodan U (2014) Deploying static application security testing on a large scale. In: Sicherheit 2014—Sicherheit, Schutz und Zuverlässigkeit
36. Pan Y (2019) Interactive application security testing. In: International conference on smart grid and electrical automation (ICSGEA), Xiangtan, China
37. Elder S, Zahan N, Kozarev V, Shu R, Menzies T, Williams L (2021) Structuring a comprehensive software security course around the OWASP application security verification standard. In: IEEE/ACM 43rd International conference on software engineering: software engineering education and training (ICSE-SEET), Madrid, ES
38. Palo Alto Networks (2022) What is policy-as-code? [Online]. Available: <https://www.paloaltonetworks.com/cyberpedia/what-is-policy-as-code>. Accessed 9, 2022



# Use of Artificial Intelligence in the Digital Marketing Strategy of Latvian Companies



Jelena Salkovska , Anda Batraga , Liene Kaibe , and Katrina Kellerte

**Abstract** As technologies advance and the interest of digital marketers and researchers in artificial intelligence increases, the ways to develop a digital marketing strategy with the tools that can improve both marketing operations and interactions with the consumers are being sought. The objective of the research is to investigate, to assess and to analyse the possibilities of using artificial intelligence in digital marketing strategy of Latvian companies, based on theoretical knowledge about the use of artificial intelligence in digital marketing strategy, the results of survey of companies' representatives and expert interviews. Research methods: In order to achieve the objective of the research, a descriptive analysis, expert interviews and a survey of representatives of Latvian companies have been carried out. Results: Artificial intelligence provides companies with several advantages that the companies can use to improve their marketing activities or to interact with the consumers; however, in Latvia, companies use artificial intelligence little, and it is difficult for a large number of companies to say whether they will implement artificial intelligence in their digital marketing strategy in the near future.

**Keywords** Artificial intelligence · Digital marketing · Digital marketing strategy

---

J. Salkovska (✉) · A. Batraga · L. Kaibe · K. Kellerte  
University of Latvia, Riga, Latvia  
e-mail: [jelena.salkovska@lu.lv](mailto:jelena.salkovska@lu.lv)

A. Batraga  
e-mail: [anda.batraga@lu.lv](mailto:anda.batraga@lu.lv)

L. Kaibe  
e-mail: [lk15078@edu.lu.lv](mailto:lk15078@edu.lu.lv)

K. Kellerte  
e-mail: [ks16058@edu.lu.lv](mailto:ks16058@edu.lu.lv)



## 1 Introduction

Companies develop digital marketing strategies and, using the capabilities and advantages of artificial intelligence, companies can get several options to develop the company's digital marketing strategy, but, knowing the shortcomings of artificial intelligence, make a decision on what solutions to implement in the company and which will help to achieve the company's digital marketing goals. Kannan and Li [1] believe that the technologies such as smartphones, the Internet of things, artificial intelligence and deep machine learning promise significant changes in the lives of consumers [1]. Improving artificial intelligence algorithms increases the capabilities of artificial intelligence, for example, supplementing digital marketing strategy with the new opportunities to reach the consumer and to improve marketing processes within the company; however, the implementation of artificial intelligence requires resources, so the potential benefits need to be assessed.

The investigated problem is the need to improve digital marketing strategy using artificial intelligence.

The objective of the research is to investigate, to assess and to analyse the possibilities of using artificial intelligence in digital marketing strategy of Latvian companies, based on theoretical knowledge about the use of artificial intelligence in digital marketing strategy, the results of survey of companies' representatives and expert interviews.

To achieve the objective of the research, the following tasks have been set:

1. Based on descriptive analysis of the literature, to analyse the theoretical and practical aspects of the use of artificial intelligence in digital marketing;
2. To conduct expert interviews on the use of artificial intelligence in digital marketing strategy of Latvian companies and the prospects of its use in future;
3. To develop a survey questionnaire for companies' representatives and to conduct a survey with the aim of finding out the experience of companies related to artificial intelligence solutions in digital marketing strategy and the factors influencing the implementation of such solutions;
4. To summarize and to analyse the obtained data, to draw conclusions and to develop recommendations for improving the digital marketing strategy of Latvian companies with artificial intelligence.

In order to achieve the objective of the research, an expert interview and an online survey of Latvian companies' representatives have been conducted. Using the *QuestionPro* online survey tool, a survey of companies' representatives (company managers, marketing managers and marketing specialists) was conducted, in which a total of 118 companies' representatives participated. The survey of companies' representatives was conducted among the companies operating in consumer (B2C) market, but companies operating in business (B2B) market were not surveyed.

The hypothesis put forward in the research: the use of artificial intelligence in digital marketing strategy of Latvian companies is hindered by a low level of awareness about the available artificial intelligence solutions in digital marketing and a lack of financial resources.

## 2 Literature Review

In the research, the articles from scientific journals and the book *Programmatic Advertising* and Eurostat statistical database have been used.

Overgoor et al. [2] have defined marketing artificial intelligence as “the development of artificial agents that, given the information they have about consumers, competitors, and the focal company, suggest and/or take marketing actions to achieve the best marketing outcome” [2]. Artificial intelligence is divided into narrow artificial intelligence (applied only to a specific task), general artificial intelligence (capable of solving problems autonomously), artificial superintelligence (can be applied in any field and surpasses a human) [3]. Artificial superintelligence is currently still being created [4]. The development of artificial intelligence is influenced by a huge amount of data, better algorithms and significantly improved computing hardware [5]. Marketing specialists need to understand how best to integrate artificial intelligence into companies and stimulate a greater individual acceptance in order to maintain a competitive advantage [6]. As artificial intelligence advances, opportunities to use artificial intelligence to improve digital marketing strategy will increase [7]. The development of artificial intelligence will not replace marketing specialists, but it will have a direct impact on consumer experience [8]. The authors highlight that different artificial intelligences can be applied to achieve different digital marketing objectives, thus achieving the best result. Artificial intelligence continues to develop and influence the everyday life of consumers and interactions with the company.

Marketing specialists can use artificial intelligence to understand consumer behaviour, actions, and metrics to reveal the appropriate approach to the consumer in a timely and efficient manner and to process large amounts of data from social media, email, and the web [9]. Artificial intelligence in digital marketing can be used for creating customized content and managing content; raising the number of conversions; voice search; programmatic media buying; propensity modelling (selecting the target audience that will make conversion); predictive analytics (predicting consumer behaviour); lead scoring (potential purchases are evaluated, using numerical scale); ad targeting (machine learning improves the result of advertisements); dynamic pricing (the prices are adjusted according to consumer data); web and app personalization; virtual assistants; retargeting; predictive customer service (the customer is being serviced, foreseeing customer behaviour in the future); marketing automation (most optimal time and topics for communication with the customer are determined); dynamic emails [10], social noise listening and analysis (artificial intelligence collects and analyses what is said and written about the brand on social media) [11]; automated decision-making; language processing [12]; image recognition and augmented reality [13]. The authors draw attention to the fact that artificial intelligence can be used to improve and supplement interaction of the brand with the

consumer by personalizing communication, communicating at the right time and offering more convenient tools, such as the ability of voice search. Higher marketing objectives can also be achieved with the help of artificial intelligence, improving data collection, data analysis and forecasting, as well as improving the delivery of advertisements to consumers.

Artificial intelligence can bring several benefits to consumers: a faster and a more convenient shopping experience; a new experience provided through personalization and a new dimension of relationship with the brand, facilitated by the possibility to view the product [28]. In a company, artificial intelligence can provide lower costs and higher revenues, faster results, accuracy, to reduce human tasks, identify and solve certain problems faster than humans, to reduce and eliminate repetitive or unimportant tasks [14]. Artificial intelligence develops marketing processes within the company—it can automate routine tasks, increase creative activity in the company, create new design innovations, improve the competencies of the marketing team and create a need for new companies that develop artificial intelligence solutions [28]. Artificial intelligence provides benefits both to the consumer, contributing to the achievement of company's objectives and to improving the company's marketing activities.

The risks and limitations of artificial intelligence in marketing are related to the fact that artificial intelligence is affected by the availability of data; there will still be tasks that will be in the competence of humans; there may be high acquisition and maintenance costs; it is necessary to evaluate the return on investment; implementation takes time, algorithms are created by humans, and therefore, there is a risk that these are incorrect; and consumer privacy may be not respected [15].

Artificial intelligence can be used not only in the field of marketing, in the process of developing a company's digital marketing strategy, in marketing research, developing marketing communication content and evaluating the effectiveness of company's marketing communication, but also in other areas, such as investigation of consumer values [16], research of the actions of commercial banks [17], in risk management process [18], in researching digital inequalities in households [19], in measuring the return and efficiency of education [20, 21]. In any field, artificial intelligence cannot feel emotions; the objectives of artificial intelligence may be not well thought out; the learning environment is limited; the data may be incomplete, which may contribute to prejudices; and artificial intelligence needs to be monitored [22]. When implementing artificial intelligence in companies, one should take into account not only the advantages, but also the risks and limitations that can interfere with the complete operation of artificial intelligence or cause risks for the company.

### 3 Research Results and Discussion

#### 3.1 *Application of Artificial Intelligence in Latvian Companies*

In order to achieve the objective of the research, a qualitative research method, expert interviews, was used, where 7 experts, whose field of activity is related to digital marketing or artificial intelligence, participated: Edgars Cerkovskis, Artis Krilovs, Rihards Zeilis, Kaspars Driks, Edgars Petersons, Edgars Koronevskis, and Aigars Armanovs. 118 companies' representatives participated in the online survey of company representatives. Of them, 91.53% or 108 companies' representatives answered all questions, but 8.47% answered in question 1 that the company does not use digital marketing channels to reach individual consumers in B2C (consumer) market, so it was not necessary to answer other questions.

In order to reveal the present situation in companies, the companies' representatives have been asked the question "Is artificial intelligence used in your company's digital marketing strategy?". 16.67% or 18 companies' representatives have answered that they use artificial intelligence in their digital marketing strategy, and 83.33% or 90 have answered that artificial intelligence is not used in their digital marketing strategy. This is a better result than the average in Europe, where in 2020, among the companies with at least 10 employees 7% of companies used artificial intelligence. In Latvia, 2% of companies did so [23].

The interviewed experts believe that companies in Latvia use artificial intelligence in digital marketing without realizing that they are doing so. Campbell et al. [24] believe that artificial intelligence is used to post marketing content on various platforms, subjecting the generated content to the algorithms present on those platforms and distributing that content further to the users of the platforms. Another tool where artificial intelligence is used partially unconsciously is digital advertising (*Facebook Ads*, *Google Ads*), where artificial intelligence can help plan campaigns, find audiences to achieve better conversion rates, generally stimulate sales and help even in the post-sale phase allowing one to continue interact with the consumer after purchasing the product (retargeting). As specific examples of use in Latvian companies, the experts named virtual assistants, e-mail services, big data processing and the tools that allow employees to focus on strategic tasks and creation of content, inspired by new activities or using the opportunities provided by communication platforms, as well as content quality control, for example, a visual solution for advertising campaigns that detects whether the created content will achieve the desired result. The authors believe that artificial intelligence is used in Latvian companies, but to some extent the companies are not aware that they are using platforms that contain artificial intelligence, as well as company representatives rarely share examples of using in public environment.

The experts have emphasized that when introducing artificial intelligence, marketing managers should assess the company's marketing objectives and how to achieve key performance indicators (KPI) with the help of artificial intelligence;

however, it should be taken into account that the Latvian market is not large, and, possibly, it is not cost-efficient to carry out highly personalized campaigns, but artificial intelligence makes it possible to improve the detailed treatment of digital marketing. The authors agree that when implementing artificial intelligence solutions, they should be coordinated with the company's existing digital marketing goals, as well as the use of artificial intelligence in Latvian companies could be affected by the size of the market, and some of the solutions are not economically sound to implement.

The experts consider the tools that help to process pictures and to create advertising texts to be the most successful applications of artificial intelligence in digital marketing in Latvian companies, because these tools improve the efficiency of employees of creative marketing communications agencies and enhance the creativity of campaigns; the second successful example of use is advertising purchase platforms that companies widely used in collecting large amounts of data, which is the basis of decision-making. The literature also highlights the influence of *Google* and *Facebook* on the development of programmatic advertising, as these companies invest financial resources and collect various data [25]. The experts mention virtual assistants as a partially successful example of the use of artificial intelligence in Latvian companies, which provide advantages in customer service; however, their use is limited by the peculiarities of Latvian language; therefore, using virtual assistants based on artificial intelligence is not cost-efficient. The experts emphasize that the fact that some consumers have bad previous experiences with virtual assistants should also be taken into account, as therefore the consumers may avoid using a virtual assistant again, and the use of a virtual assistant may harm the company's image. The authors agree that companies should only use artificial intelligence tools that help achieve the objectives and are in line with brand targets, however, considering that various tools are being developed, companies' representatives should continue to express interest in the possibilities of improving marketing activities with the help of artificial intelligence.

### ***3.2 The Benefits of Artificial Intelligence in Digital Marketing and the Barriers to the Implementation Thereof***

Among the factors that encourage a company to implement artificial intelligence in its digital marketing strategy, the experts mention: lower expenses and the opportunity to increase the efficiency of the company and to select the target audience more precisely; the wish to use interesting tools and to advance; and the growth of profit by increasing sales. Some experts have also emphasized the possibility of building the company's image by highlighting the use of artificial intelligence in marketing communication, and the opportunity for the company's employees to devote more time and attention to achieving the objectives instead of routine activities, which are

boring, as well as Latvia sometimes lacks qualified specialists, so it is beneficial for the company to use artificial intelligence.

The most often (69.44%) the companies' representatives ticked "Possibility to personalize offers to consumers", which is also one of the possibilities of use, mentioned by the interviewed experts. 58.33% of companies' representatives marked "Customer service availability 24/7", which can be implemented using a virtual assistant; however, the experts consider it a partially successful solution in Latvian market, as there is a risk that consumers may not accept it. 55.56% noted "Ability to get data analysis performed by AI". 52.78% have mentioned "Possibility to reduce the number of routine tasks". This aspect has been noted by several digital marketing experts. The other answer options have been marked in less than half of the cases, indicating that these benefits are less important for the companies. "Opportunity to build a closer relationship with the consumer"—47.22%. "Consumers get the opportunity to shop faster and in a more convenient way"—44.44%. "Opportunity to increase the productivity of a marketing specialist"—41.67%. "Getting insights about the brand/company in social networks"—36.11%. "Possibility to provide a virtual product trial"—30.56%. "Opportunity to get inspiration for design, new design solutions"—30.56%. The authors conclude that the points mentioned by companies' representatives are related to providing personalized offers, improving customer service, data analysis and improving the work of employees by reducing routine tasks. Analysing the answers to this question, the authors believe that the companies could use artificial intelligence to personalize various offers, for example, on the company's website, in e-mail messages, and introduce virtual assistants to help provide customer service 24/7, however, the company's target audience, its marketing objectives and quality of the virtual assistant should be taken into account.

Company representatives have rated "Possibility to target the audience more precisely" and "Possibility to perform marketing activities" the highest (4.17/5), which indicates that companies wish to use artificial intelligence in digital advertising, which the interviewed experts have highlighted as a successful application of artificial intelligence in digital marketing through for the big platforms—*Facebook* and *Google* and to optimize marketing activity within the company. A score of 4.14/5 has been given for the answer option "Opportunity to increase revenue", which is one of the benefits that is also highlighted in the scientific literature. Sadiku et al. [26] conclude that artificial intelligence in social media allows companies to stand out from their competitors and thereby increase profits [26]. "Possibility of reducing the number of tasks performed by a person"—4.03/5. "Lower marketing costs in the long run"—3.94/5. The advantage of artificial intelligence "Artificial intelligence does not make mistakes" has been rated the lowest (3.44/5). The authors believe that this is due to the fact that company representatives do not trust that artificial intelligence does not make mistakes. The experts have also highlighted the possibility that the artificial intelligence integrated into the virtual assistant may make mistakes, which is why consumers have a negative attitude towards this tool.

Although artificial intelligence provides a number of advantages, there are still several factors that prevent the entry of artificial intelligence and experts believe that the use of artificial intelligence in the digital marketing strategy of Latvian

companies is hindered: lack of financial resources; lack of qualified employees and knowledge on how to implement artificial intelligence solutions in the company's digital marketing activities; as well as data availability. Part of the experts believe that some marketing specialists are afraid of losing their jobs, and in general, marketing specialists used to work in a certain way and do not wish to change anything; also, marketing specialists lack information about the possibilities and benefits of using artificial intelligence in digital marketing, and artificial intelligence is necessary to monitor, including in large platforms in which artificial intelligence is embedded, such as *Facebooks Ads*. The authors believe that all these factors can significantly hinder the implementation and full use of artificial intelligence in digital marketing, as the primary need is people who have knowledge of these tools and the desire to use them, as well as a sufficient amount of resources.

When analysing the answers to next question "Factors impeding the use of artificial intelligence in digital marketing", it should be noted that none of the answer options was marked in more than 50% of the cases, which indicates that the use of artificial intelligence depends on several different factors within the company, and it is impossible to identify one or several determining factors that impede the implementation of artificial intelligence in digital marketing, based on the analysis of this question. Most often, 44.44% of the surveyed company representatives have noted "There is not enough data to create artificial intelligence solutions". This aspect has been partly mentioned by the experts, highlighting the size of the market in Latvia and data protection regulations, which impede the collection of more and variable types of data on the consumers. The authors believe that companies can address this shortcoming by using the already developed tools that require less data. 41.67% of respondents chose option "AI requires large financial investments". 33.33% marked "Artificial intelligence solutions take a long time to implement".

When evaluating various disadvantages of artificial intelligence by 5-point scale, the highest rating has "Artificial intelligence has high implementation costs" (4.03/5); however, the authors estimate that this is not a "very significant" disadvantage in the understanding of company representatives, but rather "significant", as it is 4.03. In general, the authors believe that the research hypothesis will be confirmed in this matter, since the highest rated disadvantages of artificial intelligence are related to the costs that arise during implementation and maintenance, as well as the lack of information about the return of artificial intelligence tools, which have been also noted as factors by the interviewed experts. "AI has high maintenance costs" was rated 3.89/5 and "the return of AI tools is unknown"—3.86/5. The representatives of the companies have rated as partly significant shortcomings the shortcomings related to the possible characteristics of artificial intelligence—the need to monitor (3.81/5) requires a lot of data (3.67/5); artificial intelligence does not feel emotions (3.64/5) and makes mistakes (3.58/5). To summarize, the authors believe that the companies pay more attention to measurable disadvantages and resource investments required than to potential risks that may arise when using artificial intelligence.

### ***3.3 Future Development Trends of Artificial Intelligence in Digital Marketing Strategy of Latvian Companies***

The experts were asked how they assess companies' current investments in artificial intelligence, which is used to achieve digital marketing objectives, and what risks in general the experts see in the use of artificial intelligence in Latvian market. The experts have emphasized that it is necessary to implement artificial intelligence solutions in accordance with digital marketing strategy and to define specific achievable results in order to avoid the risks associated with investing in artificial intelligence. The experts have noted that the investments of Latvian companies in digital advertising are currently increasing, which indicates that companies trust this tool; however, the experts also believe that most companies do not realize that artificial intelligence is used in digital advertising platforms, so the authors believe that companies use digital advertising platforms because they deliver results, not because they have artificial intelligence. Overall, the experts believe that employees in Latvian companies do not need to develop new tools, because many tools have already been developed, and when developing on their own they would need to invest resources (time, money, knowledge), but the return is unknown; thus, these resources may be lost.

The experts believe that in future, artificial intelligence could free marketing specialists from part of their duties, especially from routine duties, so that specialists can perform work with a higher added value; however, artificial intelligence will not replace humans. 52.78% of companies' representatives believe that the ability to reduce the number of routine tasks is an important advantage of artificial intelligence. The authors believe that it would also be valuable to carry out scientific research on this aspect in order to understand people's wish to give up some of their responsibilities or the fear that artificial intelligence could replace humans. The experts believe that artificial intelligence in digital marketing could be used to work more completely with the existing customers instead of trying to reach more new consumers and to develop the possibility to personalize messages for up to 1 person and to offer customized communication to each representative of the target audience, which the surveyed companies' representatives consider as important (4.17/5) advantages of artificial intelligence in digital marketing. The authors agree that a high personalization in marketing communication is a future perspective in Latvian companies, thus providing a personalized experience to each consumer, and this is currently a significant advantage of artificial intelligence, and as a future perspective, which is currently not openly used, is also noted in the scientific literature [27].

The experts believe that the development of artificial intelligence solutions in Latvia is hindered by the size of the market and the fact that there is not enough data on Latvian consumers, as well as the fact that companies' employees lack knowledge about artificial intelligence and skills to apply these solutions, which also coincide with the factors that currently impede the implementation of the solution of artificial intelligence. Also, the experts believe that the development of artificial intelligence depends on the development in the world, and new artificial intelligence solutions



are not programmed in Latvia, because there is a lack of financial resources to invest in development. The authors believe that attention should be paid to the aspect of insufficient allocation of financial resources for development, as this may affect the competitiveness of Latvian companies both in Latvian market and elsewhere in the world. However, part of experts are optimistic about the development of artificial intelligence and expect that globally recognized solutions will be introduced in Latvian market. Artificial intelligence is still evolving, but not all brands are ready to implement it [24]. Since artificial intelligence has potential to evolve and to develop company's digital marketing strategy, it is also necessary to analyse the willingness of companies to use this solution on a daily basis, so the following question was asked to company representatives.

Most companies find it difficult to say (49.07%) whether artificial intelligence will be used in their digital marketing strategy within the next 2 years. The authors believe that this shows that artificial intelligence in companies is not very relevant at the moment and only a small number of companies think about the development of digital marketing strategy in the medium term (2 years), and this has been also emphasized by the experts who believe that artificial intelligence is not yet relevant for a large part of companies. The authors recommend that, in future, additional research shall be conducted directly on the principles of digital marketing strategy formation for Latvian companies and the influencing factors of strategy formation. The companies in Latvia have potential to develop the use of artificial intelligence in their digital marketing strategy, because artificial intelligence is developing elsewhere in the world, so there is a potential that implementation costs will decrease and the amount of available information will grow, which are the main factors that impede the use of artificial intelligence in Latvian companies; however, companies need to be more open to new information about the available tools.

## 4 Conclusions

1. Companies in Latvia little use artificial intelligence; it is most often used in digital advertisements, virtual assistants, e-mail services and checking the content of communications, and it is difficult for company representatives to say (49.07%) whether artificial intelligence will be used within the next 2 years.
2. Companies use artificial intelligence in digital marketing because it is an opportunity to personalize offers, to provide customer service, to get data analysis and to reduce the number of routine tasks and expenses.
3. The availability of financial resources, lack of knowledge and understanding, and lack of specialists and availability of data, partly also the size of the market, hinder companies from using artificial intelligence in digital marketing.
4. It is best for Latvian companies to invest in already developed tools, because the creation of new tools requires resources, and for a large number of companies the investments may not pay off.

5. The development prospects of the use of artificial intelligence in digital marketing strategy of Latvian companies will be similar to the development elsewhere in the world, and the main ways of use could be: changing the duties of marketing specialists; building relationship with existing customers; and personalization.
6. The hypothesis put forward in the research, that the use of artificial intelligence in digital marketing strategy of Latvian companies is hindered by a low level of awareness of available artificial intelligence solutions in digital marketing and a lack of financial resources, is partially confirmed, as the experts believe that these factors impede the use of artificial intelligence, but companies' representatives highlight the lack of data, which is also noted by some experts, and the lack of financial resources.

## 5 Proposals, Recommendations

1. Company managers should assess the possibilities of using artificial intelligence solutions in digital marketing strategy, because solutions with artificial intelligence have such advantages as the ability to provide a personalized experience, customer service availability 24/7 and to perform data analysis, which are important for companies in Latvia, and there are no artificial intelligence solutions needs to be built from scratch, but already developed tools can be used.
2. Company managers should promote the interest of company's employees in latest technologies and the potential opportunities to integrate them into digital marketing in order to foster the development of company's digital marketing strategy both in interaction with consumers and in the development of the marketing team as a whole.
3. Marketing managers, when planning to implement artificial intelligence solutions in digital marketing strategy, should discuss with the company's digital marketing specialists the objectives and functions of implementation of the tools, so that the company's employees are not afraid of losing their jobs.
4. Marketing managers, when implementing artificial intelligence solutions, should pay attention to the aspects such as digital marketing objectives, available data and the environment in which artificial intelligence learns, in order to reduce the threat of artificial intelligence in digital marketing operations.
5. Digital marketing specialists should identify the possibilities to use the opportunities provided by the *Google Ads* and *Facebook Ads* platforms in company's digital marketing strategy when creating digital advertisements, as experts consider this to be the most popular and successful example of using artificial intelligence in Latvian market.
6. The Ministry of Education and Science of Latvia, together with Latvian high schools, should assess the possibilities of creating an interdisciplinary study programme with the aim of educating in the speciality of artificial intelligence, marketing and change management, because experts believe that there is a lack

of specialists in Latvia who are familiar with both marketing and artificial intelligence to be able to implement these solutions in the company.

7. Researchers and scientists should conduct additional research directly on the principles of digital marketing strategy development for Latvian companies and the influencing factors of strategy development.

## References

1. Kannan PK, Li HL (2017) Digital marketing: a framework, review and research 90 agenda. *Int J Res Market* 34(1):22–45. <https://doi.org/10.1016/j.ijresmar.2016.11.006>
2. Overgoor G, Chica M, Rand W, Weishampel A (2019) Letting the computers take over: using AI to solve marketing problems. *Calif Manag Rev* 61(4):156–185. <https://doi.org/10.1177/0008125619859318>
3. Kaplan A, Haenlein M (2021) Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Bus Horiz* 62(1):15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
4. Wirth N (2018) Hello marketing, what can artificial intelligence help you with? *Int J Mark Res* 60(5):435–438. <https://doi.org/10.1177/1470785318776841>
5. Kitsios F, Kamariotou M (2021) Artificial intelligence and business strategy towards digital transformation: a research agenda. *Sustainability* 13(4):1–14. <https://doi.org/10.3390/su13042025>
6. Kim J, Giroux M, Lee JC (2021) When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychol Market* 38(7):1140–1155. <https://doi.org/10.1002/mar.21498>
7. Rahim HA, Shafaf I, Kamaruddin SBA, Ghani NAM, Musirin I (2020) Exploration on digital marketing as business strategy model among Malaysian entrepreneurs via neurocomputing. *IAES Int J Artif Intell* 9(1):18–24
8. Dwivedi YK, Ismagilova E, Hughes DL, Carlson J, Filieri R, Jacobson J, Jain V, Karjaluoto H, Kefi H, Krishen AS, Kumar V, Rahman MM, Raman R, Rauschnabel PA, Rowley J, Salo J, Tran GJ, Wang Y (2021) Setting the future of digital and social media marketing research: perspectives and research propositions. *Int J Inform Manag* 59:102168. <https://doi.org/10.1016/j.ijinfomgt.2020.102168>
9. Haleem A, Mohd Javaid M, Qadri MA, Singh RP, Suman R (2022) Artificial intelligence (AI) applications for marketing: a literature-based study. *Int J Intell Netw* 3:119–132. <https://doi.org/10.1016/j.ijin.2022.08.005>
10. Nair KS, Gupta R (2021) Application of AI technology in modern digital marketing environment. *World J Entrep Manag Sustain Dev* 17(3):318–328. <https://doi.org/10.1108/WJEMSD-08-2020-0099>
11. Dargham M, Hachimi H (2021) Digital marketing in the era of artificial intelligence. *Int J Optim Appl* 1(3):23–28
12. Khatri M (2021) How digital marketing along with artificial intelligence is transforming consumer behaviour? *Int J Res Appl Sci Eng Technol* 9:523–527
13. Murgai A (2018) Transforming digital marketing with artificial intelligence. *Int J Latest Technol Eng Manag Appl Sci* 8:259–262
14. Ribeiro T, Reis JL (2020) Artificial intelligence applied to digital marketing. In: Rocha Á, Adeli H, Reis L, Costanzo S, Orovic I, Moreira F (eds) *Trends and innovations in information systems and technologies*. WorldCIST 2020. Advances in intelligent systems and computing, vol 1160. Springer, Cham, pp 158–169
15. Thiraviyam T (2018) Artificial intelligence marketing. In: *International journal of recent research aspects, special issue: conscientious computing technologies*, pp 449–452

16. Casno K, Sloka B, Skiltere D (2021) Valuable insights into consumer values: the case of Latvian social enterprises. *Contemp Stud Econ Finan Anal* 106:341–354. <https://doi.org/10.1108/S1569-375920210000106021>
17. Saksonova S, Koleda O (2017) Evaluating the interrelationship between actions of Latvian commercial banks and Latvian economic growth. *Proc Eng* 178:123–130. <https://doi.org/10.1016/j.proeng.2017.01.075>
18. Cekuls A (2020) The impact of remote work on team and risk management. *Int Multidiscip Sci GeoConf Surv Geol Min Ecol Manag SGEM* 20:587–594
19. Lase K, Sloka B (2021) Digital inequalities in households in Latvia: problems and challenges. *Contemp Stud Econ Finan Anal* 106:355–366. <https://doi.org/10.1108/S1569-375920210000106022>
20. Cekuls A (2019) Synthesizing the understanding of start-up from different perspectives in business learning process at university. In: *Proceedings of the 10th international multi-conference on complexity, informatics and cybernetics, proceedings, vol 2*, pp 137–140
21. Saksonova S, Vilerts K (2015) Measuring returns to education: the case of Latvia. *Analele Stiintifice ale Universitatii AI I Cuza din Iasi - Sectiunea Stiinte Economice* 62(2):251–261. <https://doi.org/10.1515/aicue-2015-0017>
22. De Bruyn A, Viswanathan V, Beh YS, Brock JK, von Wangenheim F (2020) Artificial intelligence and marketing: pitfalls and opportunities. *J Interact Market* 51:91–105. <https://doi.org/10.1016/j.intmar.2020.04.007>
23. Eurostat (2020) Artificial intelligence. [https://ec.europa.eu/eurostat/databrowser/view/ISOC\\_EB\\_AI\\_\\_custom\\_784358/bookmark/table?lang=en&bookmarkId=f34cd95d-77aa-45bd-8496-eabf49549c02](https://ec.europa.eu/eurostat/databrowser/view/ISOC_EB_AI__custom_784358/bookmark/table?lang=en&bookmarkId=f34cd95d-77aa-45bd-8496-eabf49549c02). Accessed 29 Sept 2022
24. Campbell C, Sands S, Ferraro C, Tsao HY, Mavrommatis A (2020) From data to action: how marketers can leverage AI. *Bus Horiz* 63(2):227–243. <https://doi.org/10.1016/j.bushor.2019.12.002>
25. Seitz J, Zorn S (2016) Perspectives of programmatic advertising. In: Busch O (eds) *Programmatic advertising management for professionals*. Springer, Cham, pp 37–51. [https://doi.org/10.1007/978-3-319-25023-6\\_4](https://doi.org/10.1007/978-3-319-25023-6_4)
26. Sadiku MNO, Ashaolu TJ, Ajayi-Majebi A, Musa SM (2021) Artificial intelligence in social media. *Int J Sci Adv* 2(1):15–20
27. Shah N, Engineer S, Bhagat N, Chauhan H, Shah M (2020) Research trends on the usage of machine learning and artificial intelligence in advertising. *Augment Hum Res* 5:19. <https://doi.org/10.1007/s41133-020-00038-8>
28. Jarek K, Mazurek G (2019) Marketing and artificial intelligence, *Cent Eur Bus Rev* 8(2):46–55

# SR-OIR-SSD: Super-Resolved Eyes in the Sky



Raghav Sharma and Rohit Pandey

**Abstract** In many situations, observing from a high altitude is always fruitful. Humans have been aware of this fact since ancient times, and the military watchtower that changed the result of many wars is the best example. In the current scenario, humans have the latest technology, like satellites, unmanned aerial vehicles (UAVs) that provide aerial images from high altitudes. Aerial photos can be analysed manually or through AI algorithms like object detectors and tracking. In this age of automation, manual execution is not recommended. In this work, object detection has been performed on aerial images to get inside aerial photos. Object detection is a mature technique in AI. Several state-of-the-art deep learning methods for object detection, such as R-CNN, Fast R-CNN, Faster R-CNN, SSD, and YOLO, have been published in recent years. However, these methods perform well with standard and low-resolution photos and are inappropriate for aerial images. Aerial images are very different as compared to standard images. It contains too many objects in a single image. Pixel per object is significantly less as compared to the traditional image. The conventional image has a front view of the object, while the aerial photo has a top view. Hence none of the existing algorithms works fine for aerial images. This work proposes a novel technique based on a GAN and modified SSD architecture for object detection on aerial images attaining high mAP. This architecture has improved mAP from 0.72 to 0.92 on the Stanford data set, i.e. (28% improvement), while from 0.04 to 0.586 on Visedrone2018.

**Keywords** Object detection · Aerial vision · Modified SSD · SR-GAN · Selective patching

---

R. Sharma (✉) · R. Pandey  
Hughes Systique Corporation, Gurugram, India  
e-mail: [Raghav.sharma@hsc.com](mailto:Raghav.sharma@hsc.com)

R. Pandey  
e-mail: [Rohit.pandey@hsc.com](mailto:Rohit.pandey@hsc.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_65](https://doi.org/10.1007/978-981-99-3091-3_65)

799

# 1 Introduction

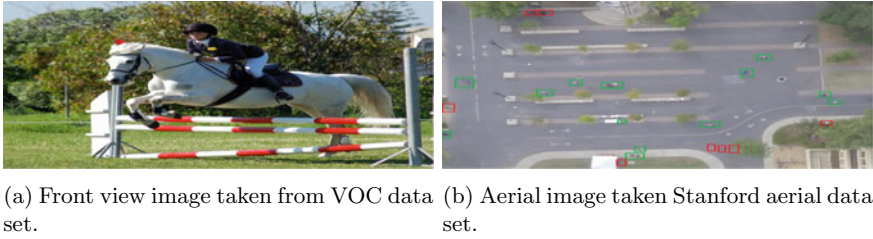
Object detection is the algorithm that can localise and classify each desired object present in an image. Each object class has its unique feature, which helps to classify and localise the object. Humans can simply extract that feature to identify the item in an image. However, in the case of machines, machines must first determine the probable object region. Then, it will extract hand-crafted features such as SIFT, HOG, and Haar-like to represent them before classifying them using classifiers such as DPM, SVM, and AdaBoost. These methods are difficult to develop due to their complexity. To reduce complexity and enhance object identification, numerous methods based on deep learning frameworks such as R-CNN [2], YOLO [9], and SSD [6] were developed. Algorithms based on deep learning can automatically learn object characteristics through training. However, the proposed work includes detecting objects in aerial photos, which is a very challenging task, even for deep learning. Aerial images captured from aircraft or other flying objects cover a large geographical area and thousands of things. As a result, aerial images and standard images are very different in many ways. A GAN-based, patching-based object detector has been proposed to handle this issue. The motivation, challenges, and problem statement behind this work have been discussed in this section.

## 1.1 Motivation

Object detection is a well-known technique, and even humans analyse a variety of circumstances using object detection on images captured by their eyes. Identical to automating a machine, it is necessary to do object recognition on images for them to make situation-dependent decisions like humans. It has various applications in the industry, such as self-driving vehicles, identifying group activities, and ensuring proper quality control of parts in the assembly, defence, automation, manufacturing, agriculture, infrastructure, media, railways, and roads. This work focused on object detection on aerial images. Aerial images have their advantage because they have taken on high altitudes and cover a vast area and thousands of objects in a single image to provide a big picture of the site in many situations. It has various applications like surveillance, disaster management, object tracking, resource management, product delivery, traffic control, destructive weapon, and many more.

## 1.2 Challenges

The ideal object detector should handle numerous types of diversity in the image, like lighting, colour, size, scale, pose, view, noise, and many more. Challenges to facing an object detector can be divided into two categories. Challenges common



**Fig. 1** Sample images of data set

between standard and aerial images and Challenges occurred for aerial images. In this section, the challenges that occurred with aerial images have been discussed.

**Scale and size** The dimensions and size of an object in an aerial photograph are substantially less than in frontal images. For instance, the object in front-view images is sufficiently large compared to aerial images. As shown in Fig. 1a, a child riding a horse is easy to identify, whereas it is difficult to identify an object in an aerial image. The green patch in Fig. 1b indicates the object to identify, whereas the red spot in the image indicates an area where the object is not present. Still, it appears identical to an object that cannot be distinguished easily.

**Number of object** The ideal object detector should be independent of the object present in the image. However, most existing algorithms detect a limited number of objects efficiently, which is only suitable for front-view images. Aerial images cover a large area in a single image ( $0.12\text{--}0.36\text{ km}^2$  per image), which can contain thousands of objects. Therefore, the existing approaches are inefficient for aerial images.

**Size and resolution of image** The existing algorithm is specially designed for low-resolution images (300–600), whereas aerial images are high-resolution images (2K–4K). Now, if the image's size is reduced, it will lose the vital information which will help to detect the desired object. On the other hand, the existing algorithm cannot process an image without making it low resolution.

### 1.3 Problem of Statement

As discussed in Sect. 1.2, the ideal object detector should be able to handle illumination, colour, scale, pose, size, viewpoint, deformation, number of objects, and any size of the image. Object detection is a mature technique. Most of the existing algorithms can handle the variation associated with the standard image. However, these algorithms cannot handle problems associated with aerial images: the size and scale of the object, the number of objects present in the image, the pixel per object, and the image's resolution. This work is dedicated to handling the problem associated with aerial images.

## 2 Literature Survey

Aerial data sets acquired from a considerable height might cover large geographical areas and contain hundreds or thousands of items per image. The existing algorithm cannot process these types of data sets. Modifying existing algorithms or new development is needed to handle these data sets. Deep learning algorithms either specially developed for aerial images or commonly used for aerial images as well as front-view images have been discussed in this section.

Deep learning object detectors can be categorised into two groups multi-stage networks and single-stage networks. The multi-stage algorithm in which object detection and classification need more than one stage is slow because of multiprocessing and unsuitable for real-time applications like R-CNN, Fast R-CNN, and Faster R-CNN. The single-stage network was introduced to improve object detection speed. It needs only one step for object detection and classification. It is far faster than the multi-stage network, like you look only once (YOLO), single shot detector (SSD), and Retina Net. All object detectors were originally intended for standard images. These object detectors can be utilised for aerial photographs. However, they are limited in their capabilities.

Object detector [8] used the existing algorithm YOLO and performed well, but it is limited to some data. Object detector [3, 7, 15] used some modifications in the existing network to detect and track the object in aerial images. Object detectors [1, 14, 16] are specially designed to handle the orientation of objects that occur in aerial images. All these object detectors used CNN and required bulky hardware. The DroNet [5] used a compact network similar to tiny-YOLO. It is a small network and requires less computation than other deep learning networks. It can work with Raspberry Pi, but the processing speed is 5–6 FPS, which is insufficient. Some other object detectors [4, 18] proposed a hardware-specific network that can perform well on specific low-power hardware in real time. All these detectors could not handle all challenges discussed in Sect. 1.2. The proposed network tried to address all the issues using the novel approach discussed in the next section.

## 3 Proposed Methodology

Object detection is primarily used to locate the desired image object. The object detector's detection accuracy depends on the object's size, image quality, and object-to-image ratio. According to Sect. 1.2, aerial picture object size and object-to-image ratio are pretty low. It is one of the primary reasons why object detection on aerial photographs has a lower mAP (mean Average Precision) than on ordinary pictures. Mean average precision (mAP) is a metric for evaluating object detection. It measures the similarity between prediction and ground truth. mAP was utilised to evaluate the performance of the proposed model. The proposed architecture aims to enhance mAP with a low object-to-image ratio and object size. It combines four modules selective patching, GAN (Generative adversarial network), an object detector, and a fusion of





**Fig. 2** SR-OIR-SSD combine four-block selective patching, GAN, object detector, and fusion of results. RGB image of size  $(1024 \times 1024 \times 3)$  feeds to patching block, which converts it into multiple patches of size  $(256 \times 256 \times 3)$  with an overlap of 128. Selective patching only generates patches containing objects. All these patches feed to GAN, increasing its resolution by two with high PSNR. The output of the GAN feeds to the object detector to perform detection. In the end, output of all the patches combines for the final result

results. All four modules have been discussed in this section. The overall architecture of the proposed network is shown in Fig. 2.

### 3.1 Selective Patching

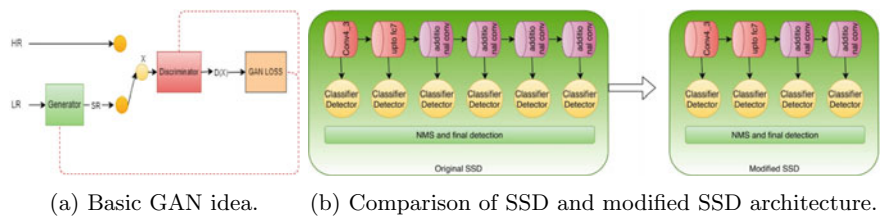
It has been discussed in Sect. 1.2 that the object-to-image ratio in aerial images is meagre, which affects the mean average precision of object detection. The object-to-image ratio can be increased by decreasing the image size while the object size should be constant. It is possible using patching, for example, the object of size  $10 \times 10$  present in the image of size  $1024 \times 1024 \times 3$ , so the object-to-image ratio is around 0.01. The image cuts into patches of  $256 \times 256 \times 3$ . Now the object-to-image ratio of each patch is around 0.04. However, it generates multiple patches to process in place of a single image, which reduces the speed of the object detector but improves the mAP. In this work,  $256 \times 256 \times 3$  patch size has been used with overlapping of  $128 \times 128 \times 3$ . Numerous patches do not contain any objects, and the processing time would increase significantly if all patches were to be processed. Selective patching has been implemented to address this issue, which will generate only patches containing objects. CNN's sliding window has been used to implement selective patching discussed below.

**Selective patching architecture** Selective patching consists of two components. The first component is CNN's sliding window, which generates the heat map of the entire image. Second, generate  $256 \times 256 \times 3$  patches According to this heat map.

CNN sliding window architecture is shown in Table 1. It is trained as a classifier with image size  $64 \times 64 \times 3$  with two classes, object and background. It only consists of convolution layer, so it is not a size-dependent network. When photos exceeding  $(64 \times 64 \times 3)$  in size are fed to network, it starts producing a heat map of that image. The produced heat map is a binary image, with the value one where the object is present and zero where it is absent. Using this heat map, only the patched areas where the object is present would be cropped. GAN and object detector would further process this cropped patch.

**Table 1** Sliding window network architecture

Type of layer	Filter size	No. of filters	Type of layer	Filter size	No. of filters
Conv2D	(3,3)	16	Maxpooling2d	(4,4)	
Conv2D	(3,3)	16	Conv2D	(4,4)	128
Maxpooling2d	(4,4)		Dropout (0.5)		
Conv2D	(3,3)	32	Conv2D	(1,1)	1
Conv2D	(3,3)	32			



**Fig. 3** Basic block diagram of GAN and modified SSD

3.2 GAN

Generative adversarial network (GAN) [13] has many applications in various fields. In this work, GAN has been used as a super-resolved network. It is used to increase the image size by two with high PSNR. High PSNR improves object detection [11]. In this work, GAN has been used to improve the PSNR, image size, and object detector performance. The GAN has two module generators and discriminators, as shown in Fig. 3a.

The generator tries to generate a high-resolution image with the help of an input or low-resolution image called a super-resolved (SR) image. The discriminator tries to differentiate between the SR and HR images (Ground truth). Using this technique, it learns to generate the high-resolution image of the input image. Patches of size  $256 \times 256 \times 3$  feed to this network. It will super-resolve it into  $512 \times 3$  as shown in Fig. 2. GAN used in this work is ESRGAN [13].

3.3 Object Detector

Object-to-image ratio and object scale has been improved using patching and GAN. The output image of GAN has a sufficient object-to-image ratio and object scale so that the object detector can perform well. In this work modified SSD object detector has been used with parameter and layer modification. The parameter and layer were altered to be more suitable for aerial images than standard images. Parameter modification and architecture modification have been discussed in this section.

**Architecture modification** SSD has six layers to handle the various scale of objects. All the layers are trained to handle specific scale or size objects. The Last two layers are used to handle large objects that can cover at least half the image. The object size in the aerial image is significantly less. Consequently, it is assumed that the remaining two layers are not required for aerial views. In this work, the last two layers have been removed, as shown in Fig. 3b. Only four layers have been used to handle the object. This modification does not affect the mAP of the detector, but it slightly increases the speed of the detector.

**Parameter modification** Most object detectors use the anchor box to perform the detection. SSD also uses the anchor box at every layer. The size and number of anchor boxes differ for each layer. The size (height and width) of the anchor boxes is decided using the scale parameter, which is different for every layer.

$$h_j^i = \frac{s_j}{\sqrt{a_i}}$$

$$w_j^i = s_j \sqrt{a_i}$$

where

$h_j^i$  height at  $j$ th scale and  $i$ th aspect ratio.

$w_j^i$  weight at  $j$ th scale and  $i$ th aspect ratio.

The minimum scale in the original SSD was 0.2. It has changed to 0.05 so SSD can detect the small-scale object. This modification improves mAP from 0.258 to 0.312 on the Visdrone2018 data set.

### 3.4 Fusion of Results

The above step generates multiple patches of the single image. The object detector will produce the bounding boxes with respect to every patch. For the final result, the bounding boxes of every patch need to be merged. For this, patch-bounding boxes are converted to image-bounding boxes using the relative location of the patch then non-maximum suppression (NMS) is applied to the bounding boxes. It will produce the final result for a single image.

## 4 Data Set

**Stanford aerial pedestrian data set** Stanford aerial pedestrian data set [10] provided by computational vision and geometry lab Stanford. The Stanford aerial pedestrian data set consists of annotated videos of pedestrians, bikers, skateboarders, cars, buses, and golf carts navigating eight scenes on the Stanford University campus. The total

video is 59, and every video contains frames of size  $1424 \times 1088$ . In the collection of this data set, camera was fixed at a particular location, whereas an object is in moving condition.

**Visdrone** Visdrone2018 and Visdrone2019 [17] data set collected from 14 cities in China. It contains 6471 training images and 548 validation images for the object detection task. During the acquisition of this data set, the drone and an object were in motion throughout fourteen cities. Aside from this, drone altitude also varies, resulting in a vast degree of fluctuation in the object's background, scale, and size. The objects in the image's pixel range vary from 15 to 200 pixels. Significant scale variation is present in the data set. The Visdrone data set contains vast diversity, making it more difficult than the Stanford aerial data set. This is also reflected in the outcome, where the model got higher accuracy in Stanford aerial data set. Even though there is a substantial difference between the two data sets, the proposed model performs well with both data sets.

## 5 Result

In this work, patching, SR-GAN, and modified SSD have been used to improve the mAP of the object detector. Each module or modification affects the accuracy of the object detector. Initially, the SSD was utilised without modification, patching, or SR-GAN, and the mAP was 0.04. Later, patching was utilised, which increased it to 0.257 as the image's object-to-image ratio was enhanced. In the second stage, SSD has been modified to detect objects with a low object-to-image ratio. It has improved the mAP from 0.257 to 0.312. In the third phase, SR-GAN was applied, which enhanced the object scale with a high PSNR and increased the mAP from 0.31 to 0.58. Table 2 shows the importance of each stage and mAP improved with each modification.

**Result on different data set** To evaluate the proposed model, three data sets were utilised. The performance of the model has been shown in Table 3. The Visdrone data set has a significant variation in the background, scale, and size of the object, so the performance of that data set is low compared to Stanford aerial data set. However, the performance of the proposed model is far better than the other model on the Visdrone data set discussed in the next section.

**Comparison with the other algorithms** The result of applying the proposed approach to the Stanford aerial data set is shown in Table 4b. It compares the proposed approach with other methods [12]. As per our best knowledge, proposed method is getting the highest accuracy on this data set. Table 4a shows the mAP comparison of the proposed method with other methods [17] on the Visdrone2018 data set.

**Table 2** Comparison of the mAP on Visdrone2018 data set

Object	mAP with patching	mAP with SSD modification and patching	mAP with modification, patching, and SR-GAN	Object	mAP with patching	mAP with SSD modification and patching	mAP with modification, patching, and SR-GAN
Ignored-regions	0.021	0.025	0.095	Truck	0.278	0.35	0.653
Pedestrian	0.282	0.336	0.635	Tricycle	0.228	0.264	0.641
Awning-tricycle	0.170	0.205	0.552	People	0.230	0.274	0.650
Bicycle	0.131	0.179	0.605	Bus	0.426	0.515	0.804
Car	0.628	0.675	0.822	Motor	0.307	0.366	0.728
Van	0.384	0.437	0.726	Others	0.25	0.120	0.120
				mAP	0.257	0.312	0.586

**Table 3** mAP of SR-OIR-SSD on a different data set

Data set	mAP without SR-GAN	mAP with SR-GAN	Data set	mAP without SR-GAN	mAP with SR-GAN
Stanford	0.9175	0.946	Visdrone2019	0.343	0.413
Visdrone2018	0.312	0.586			

**Table 4** Comparison with other approaches on different data sets

(a) Comparison on Visdrone2018 data				(b) Comparison on Stanford data			
Algorithm	mAP	Algorithm	mAP	Algorithm	mAP	Algorithm	mAP
Faster R-CNN2	0.4018	L-H R-CNN+	0.4028	SSD (RN-50)	0.804	FR-CNN (RN-101)	0.853
RD4MS	0.4485	CFE-SSDv2	0.4730	FR-CNN (RN-50)	0.836	SSD (RN-101)	0.819
DE-FPN	0.4872	HAL-Retina-Net	0.4618	RetinaNet (RN-50)	0.852	RetinaNet (RN-101)	0.866
DPNet	0.5462						
SR-OIR-SSD	0.586			SR-OIR-SSD	0.946		

## 6 Conclusion

It has been demonstrated in Sect. 5 that the proposed adjustment improves image object detection. It demonstrates the positive effect of patching and GAN on object detection, which boosted the model's performance on Visedrone2018 from 0.04 to 0.586. Moreover, it outperformed other proposed algorithms. As patching has been used, which generates many patches to execute, GAN was also utilised, which is computationally expensive. These two modules greatly enhanced performance; however, they also make it computationally expensive.

## References

1. Ding J, Xue N, Long Y, Xia GS, Lu Q (2018) Learning RoI transformer for detecting oriented objects in aerial images. [arXiv:1812.00155](#)
2. Girshick RB, Donahue J, Darrell T, Malik J (2013) Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR. [arXiv:1311.2524](#)
3. Hu Y, Xiao M, Zhang K, Wang X (2019) Aerial infrared target tracking in complex background based on combined tracking and detecting. Math Probl Eng 2019
4. Jose G, Kumar A, Kruthiventi S, Saha S, Muralidhara H (2019) Real-time object detection on low power embedded platforms. In: Proceedings of the IEEE international conference on computer vision workshops, pp 0–0
5. Kyrkou C, Plastiras G, Theodoridis T, Venieris SI, Bouganis CS (2018) Dronet: efficient convolutional neural network detector for real-time UAV applications. In: 2018 Design, automation & test in Europe conference & exhibition (DATE). IEEE, pp 967–972
6. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C, Berg AC (2015) SSD: single shot multibox detector. CoRR. [arXiv:1512.02325](#)
7. Malagi V, Ramesh B, Rangarajan K (2016) Multi-object tracking in aerial image sequences using aerial tracking learning and detection algorithm. Defence Sci J 66(2):122–129
8. Radovic M, Adarkwa O, Wang Q (2017) Object recognition in aerial images using convolutional neural networks. J Imaging 3(2):21
9. Redmon J, Divvala SK, Girshick RB, Farhadi A (2015) You only look once: unified, real-time object detection. CoRR. [arXiv:1506.02640](#)
10. Robicquet A, Sadeghian A, Alahi A, Savarese S (2016) Learning social etiquette: human trajectory understanding in crowded scenes. In: European conference on computer vision. Springer, Berlin, pp 549–565
11. Shermeyer J, Van Etten A (2019) The effects of super-resolution on object detection performance in satellite imagery. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 0–0
12. Wang X, Cheng P, Liu X, Uzochukwu B (2018) Fast and accurate, convolutional neural network based approach for object detection from UAV. CoRR. [arXiv:1808.05756](#)
13. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Loy CC, Qiao Y, Tang X (2018) ESRGAN: enhanced super-resolution generative adversarial networks. CoRR. [arXiv:1809.00219](#)
14. Xu Y, Fu M, Wang Q, Wang Y, Chen K, Xia GS, Bai X (2019) Gliding vertex on the horizontal bounding box for multi-oriented object detection. [arXiv:1911.09358](#)
15. Yang F, Fan H, Chu P, Blasch E, Ling H (2019) Clustered object detection in aerial images. In: Proceedings of the IEEE international conference on computer vision, pp 8311–8320
16. Yang X, Yang J, Yan J, Zhang Y, Zhang T, Guo Z, Sun X, Fu K (2019) SCRDet: towards more robust detection for small, cluttered and rotated objects. In: Proceedings of the IEEE international conference on computer vision, pp 8232–8241

17. Zhu P, Wen L, Du D, Bian X, Ling H, Hu Q, Nie Q, Cheng H, Liu C, Liu X et al (2018) VisDrone-DET2018: the vision meets drone object detection in image challenge results. In: Proceedings of the European conference on computer vision (ECCV), pp 0–0
18. Zhu Y, Samajdar A, Mattina M, Whatmough P (2018) Euphrates: algorithm-SoC Co-design for low-power mobile continuous vision. [arXiv:1803.11232](https://arxiv.org/abs/1803.11232)

# Participatory Design as an Audiovisual Strategy in Brand Manuals



Carlos Borja-Galeas  and Hugo Arias-Flores 

**Abstract** The brand manuals serve so that the corporate image of a company is not distorted, when it is used in activities such as the design of advertising material. Designing a brand through participatory design allows accelerating and improving its development with new technologies. This study exposes a brand building model in which 96 companies registered at the National Institute of Popular and Solidarity Economy of the city of Quito participated. As a result of this research, the participants obtained the design of their logo and brand, based on their needs and the business in which they specialize.

**Keywords** Brands · Logos · Participatory design · Technological tools

## 1 Introduction

The brand manuals serve so that the corporate image of a company is well used in the fields of design. Ecuador is recognized as one of the countries that generates the largest number of companies in the world, which has become more important in the last decade, both as an economic and academic activity, as well as at the center of public policies [1]. These business activities, in their business construction phase, require for their commercialization to have a brand design that represents them and is recognized by their consumers. A logo has a significant influence on consumer evaluations of the brand [2]. The symmetry of the brand logo generates product design inferences [3].

---

C. Borja-Galeas (✉)

Facultad de Administracion de Empresas, Universidad Tecnologica Indoamerica, Av. Machala Y Sabanilla, Quito, Ecuador

e-mail: [carlosborja@uti.edu.ec](mailto:carlosborja@uti.edu.ec)

H. Arias-Flores

Centro de Investigación en Mecatrónica Y Sistemas Interactivos (MIST), Universidad Tecnológica Indoamérica, Av. Machala Y Sabanilla, Quito, Ecuador

e-mail: [hugoarias@uti.edu.ec](mailto:hugoarias@uti.edu.ec)



Thus, the construction of a logo that expresses the essence of a brand requires a whole process of research, study and analysis. Knowing the business model, being clear about the company's value proposition, being clear about the public to which the product or service is directed [4], generates better results in brand design.

The use of technological tools as a transformative element in digitalization, printing and prototyping speeds up the process and significant savings are achieved in the use of materials before its construction [5]. Although participatory design may seem similar to joint design or design user-centered at the level of design activities, PD differs in terms of the mechanisms used and how the effects are conceived and maintained [6].

In this context, we worked with 96 companies registered with the National Institute of Popular and Solidarity Economy, which were in the process of establishing the brand design to market their products and services. Several companies had graphic proposals prepared by their own owners, without knowledge of design and composition, and according to the initial diagnosis, the need to improve the image of their products and services was confirmed.

By having a large number of companies participating in this process, in relation to the number of collaborators, we proceeded to develop a methodology that allows an accelerated understanding of the essence of each of the businesses and solving their requirements. The creation of new visual proposals and the refreshment of the brands in a period of 60 days, was the result of this process.

This article presents a brand development model and analyzes the scenario in which it is developed. The second section presents the methods used to develop the proposal. The third section presents the model developed for the research. In the fourth section, the proposed model is compared with those existing in the market, exposing the advantages and deficiencies of these at the time of achieving a result. And finally, conclusions and future work related to the theme of brand design are presented.

## 2 Methods and Materials

To design a logo you can start from several approaches. Using participatory design in this process made it possible to speed up the obtaining of results and the approval of business designs, since they contributed to the research process.

### A. Participatory design

The participatory design allowed all members to be involved in the theory, understanding and production of the process. In this method, it is a priority to take into account the mechanisms used, the effects produced and the way in which they are sustained [6].

## B. Virtual meeting platform

There are several virtual platforms for holding meetings. The Zoom platform was used in this research, as it can be downloaded for free from the Google Play Store and was used during the COVID-19 pandemic [7]. By allowing meetings with more than 100 people, it was possible to work in general meetings and meetings in small rooms from the main room.

## C. Jamboard

One of the main challenges of teaching during the pandemic was creating active learning experiences in small groups in a synchronous online learning environment [8]. One of the main challenges of teaching during the pandemic was creating active learning experiences in small groups in a synchronous online learning environment.

## D. Alex Osterwalder's business model canvas

The application of the Canvas with its 9 boxes to explain a business model, allows in a very short time to understand the essence of the business model in a time that does not exceed 5 min [4]. This canvas model allowed participants to understand or define elements of their business model that were not clear to them, such as the public their business was focused on, the value proposition that the company has, and its use in business discourse.

## E. Ideation matrix with visual metaphors

The matrix [9] gathers the relevant materials, links them to the brand, the promise and the personality of the client. In this matrix the most outstanding attributes of the brand are identified and in the other axis there are words that are simple enough for a child to understand. From there arise multiple combinations that generate unsuspected and very creative metaphors.

## F. Typography

The typography divided into 3 large groups: without serif, with serif and decorative allowed to capture the essence of the brand in a size, thickness, shapes, also allowed the application of legibility and legibility [10].

## G. Color

Defining the color that a brand will have is based on a previous study of its meanings. The preferential use of 1–3 colors allowed the viewer's attention to focus on the shape and its harmony with the selected color [11].

## H. Adobe illustrator

This tool allowed to digitally redraw different contents [12]. These, in turn, were later exported and taken to the construction of labels, packaging, manuals, audiovisual material, etc. This tool allowed a harmonious presentation of the designs, their retouching and improvement and their final export to the different scenarios of use of the brand.

### 3 Proposed Model

For this research project, the total number of entrepreneurs was distributed in work teams. For each group, two collaborators were assigned who were the guides. In this way, with an initial presentation, specific activities were prepared that were developed by each working group with the support of the assigned guides. This work methodology was applied due to the COVID-19 pandemic. Due to issues of public health, mobility and distance from the university center, it was proposed to work in 2-h meetings virtually for 2 weeks in a row.

Using technological tools such as the Google Jamboard, shared work boards were created, where the work of the entrepreneurs was exhibited. The first group work was the elaboration of the Canvas Canvas by Alex Osterwalder, as can be seen in Fig. 1. The existence of people with limited technology or knowledge or in the use of these tools led some participants to deliver their contributions with virtual communication tools such as WhatsApp, sending photographs of their written contributions.

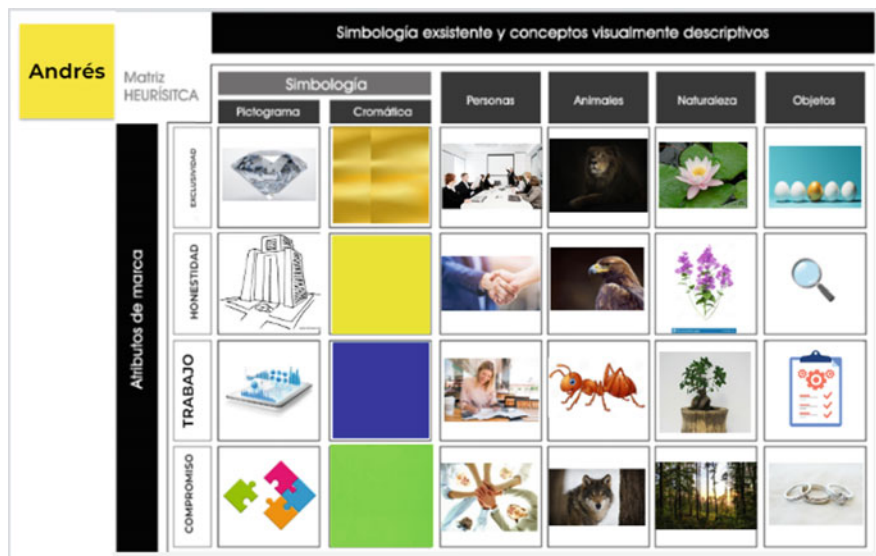
In a second moment, we worked with the graphic ideation matrix with visual metaphors, as can be seen in Fig. 2, in which the owners of the companies participated.

With the images that were located in this graphic matrix, various combinations were generated that resulted in an average of 3 sketches per company. Some brands did not require this processing, but only a refresh, which consisted of improving the strokes, changing colors, improving the typography used in the brand name.

With the renewed sketches and designs, the brand owners were given the go-ahead for their subsequent preparation of brand manuals and design of labels and packaging



**Fig. 1** Business model canvas by Alex Osterwalder adapted to the Jamboard tool developed by Carlos Borja



**Fig. 2** Heuristic matrix of ideation with visual metaphors elaborated by Carlos Borja

for their different products and services. Figure 3 shows some of the results obtained in this process.

4 Discussion

This methodology was applied virtually due to the context of the COVID-19 pandemic, but its application in a face-to-face environment would allow obtaining results with different development times and with feedback from participants with specific characteristics. The limitation in this new scenario would be the mobilization of all the participants. In the participating human group, a distance survey of the commitments with the university center was carried out and the mobilization times would be between 30 and 120 min.

5 Conclusions

The study allowed the development of various trademarks, products and services virtually in a specific and programmed time, but this process could reduce production time if the dynamic becomes face-to-face in a suitable space for that purpose. By working with digital technological tools that require the use of the Internet, it limits the participation of those who do not have computers or telephones that have



**Fig. 3** Presentation of final designs of a group of renewed brands and others created from scratch by the group of students and companies that participated in the brand building process, prepared by Carlos Borja

connectivity plans. In the case of companies that had connection problems, they were asked to send their contributions via WhatsApp and then these were digitized by the students and socialized in the work group. For this human group, face-to-face meetings would allow a better understanding and contribution to their brands.

The motivation generated by working as a team in building a brand with the participation of graphic design students allows accelerating the expected results with the confirmation and approval of the brand owners.

Lack of knowledge of digital tools slows down the achievement of the objectives expected per meeting, requiring additional training to learn how to use them, which was contemplated in this process.

There was a commitment from all the ventures, but during the development sessions some members started late due to connectivity problems or personal problems that had to be justified to continue with the brand building process.

## References

1. Wong SA (2022) Entrepreneurship in Ecuador. *Entrepreneurship in South America: context, diversity, constraints, opportunities and prospects*. Springer, Cham, pp 99–115
2. Septianto F, Paramita W (2021) Cute brand logo enhances favorable brand attitude: the moderating role of hope. *J Retail Consum Serv* 63:102734. <https://doi.org/10.1016/j.jretconser.2021.102734>
3. Bettels J, Wiedmann KP (2019) Brand logo symmetry and product design: the spillover effects on consumer inferences. *J Bus Res* 97:1–9. <https://doi.org/10.1016/j.jbusres.2018.12.039>
4. Osterwalder Y, Pigneur A (2010) *Business model generation: a handbook for visionaries, game changers, and challengers*. Wiley, Hoboken
5. Borja-Galeas C, Arias-Flores H, Jadan-Guerrero J (2022) Creative packaging design for products, pp 907–911. [https://doi.org/10.1007/978-3-030-85540-6\\_115](https://doi.org/10.1007/978-3-030-85540-6_115)
6. Hansen NB, et al (2019) How participatory design works. In: *Proceedings of the 31st Australian conference on human-computer-interaction*, pp 30–41. <https://doi.org/10.1145/3369457.3369460>
7. Wijaya F, Solikhatin SA, Tahyudin C (2021) Analysis of end-user satisfaction of zoom application for online lectures. In: *Proceedings of the 2021 3rd East Indonesia conference on computer and information technology (EIConCIT)*, pp 348–353. <https://doi.org/10.1109/EIConCIT50028.2021.9431903>
8. Sullivan P (2022) Leveraging the power of Google Apps to support active learning in a synchronous online environment. *Int J Math Educ Sci Technol* 53(3):610–618. <https://doi.org/10.1080/0020739X.2021.1994159>
9. Capsule (2007) *Claves del diseño LOGOS*. Barcelona
10. Rolo E (2019) The typographic grid in the editorial project: an essential resource to the graphic consistency and perception. In: *Proceedings of the 20th congress of the international ergonomics association (IEA 2018) Volume X: auditory and vocal ergonomics, visual ergonomics, psychophysiology in ergonomics, ergonomics in advanced imaging*. Springer, New York, pp 61–72. [https://doi.org/10.1007/978-3-319-96059-3\\_7](https://doi.org/10.1007/978-3-319-96059-3_7)
11. Grzybowski A, Kupidura-Majewski K (2019) What is color and how it is perceived? *Clin Dermatol* 37(5):392–401. <https://doi.org/10.1016/j.clindermatol.2019.07.008>
12. Adobe (2020) *Illustrator está por todas partes*. <https://www.adobe.com/la/products/illustrator.html>

# Citizen Engagement on Government Social Media: Validation of Measurement Items



Ari Wedhasmara, Samsuryadi, and Ab Razak Che Hussin

**Abstract** The emergence of social media with web 2.0 technology and User Generated Content (UGC), has brought change, this can be seen by the increasing number of social media users around the world, in line with the increasing Internet penetration. This is what makes governments in various parts of the world use social media primarily to gain community involvement. However, until now the use of social media has not been fully optimal, especially in Indonesia, where based on several research results it is still only for conveying information, and has not been optimal in responding to and monitoring feedback from citizens, in fact it is still far from collaboration between the community and the government. Based on these problems, this study aims to find factors that will become a model for citizen involvement in government social media, using a systematic literature review (SLR). The factors that have been obtained will first be assessed in terms of content validity [content validity index (CVI) and scale-content validity index (S-CVI)] to obtain question instruments that are in line with citizens' expectations, which will then be make a questionnaire for residents, as supporting data in a quantitative model assessment. There are 14 constructs that are part of the citizen engagement model on government social media, with two items out of 56 question items that were rejected based on CVI calculation results, 54 of these question items which will then be distributed as questionnaires to citizens.

**Keywords** Citizen engagement model on government social media • CVI • S-CVI

---

A. Wedhasmara (✉) • Samsuryadi  
Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia  
e-mail: [a\\_wedhasmara@unsri.ac.id](mailto:a_wedhasmara@unsri.ac.id)

Samsuryadi  
e-mail: [samsuryadi@unsri.ac.id](mailto:samsuryadi@unsri.ac.id)

A. R. C. Hussin  
Azman Hashim International Business School (AHIBS), Universiti Teknologi Malaysia (UTM),  
Johor Bahru, Malaysia  
e-mail: [abrazak@utm.my](mailto:abrazak@utm.my)

# 1 Introduction

Social media as we know it today, started in 1998 with the emergence of “Open Diary” [1], where the notion of social media is a collection of devices that use the Internet with web 2.0 technology and User Generated Content (UGC) [2]. There is an increase in active social media users in 2022, as many as 58.4% of the world’s population, this is considered reasonable because it is accompanied by an increase in Internet penetration, as many as 62.5% of the world’s population and global cellular users, as much as 67.1%, the current trend of social media usage more widely used in work-related activities, advertising, and marketing [3]. Seeing the enormous opportunities from using social media, it’s no wonder that so many interested parties use it, especially the benefits and challenges for the government [4].

The use of social media by the government is known as government 2.0, this is a continuation of government 1.0, where government 1.0 focuses more on the use of Information and Communication Technology (ICT) with web 1.0 technology, this initiative improves government performance in serving various kinds of stakeholders (citizens, businesses, and the government itself) [5]. However, this is felt to be lacking in citizen engagement with the government [6], to further ensure linkages with openness, transparency, and citizen engagement with the government, social media is considered suitable as a mediation for web 2.0 and UGC-based technologies [7]. Various kinds of labels for the use of social media by the government have been coined by many researchers such as the social government, the we government, and others [8–15], social media has enabled a new type of governance known as “Government 2.0”, which emphasizes openness, collaboration, and transparency [7]. This is what changes the government from e-government services, to social government, by providing services through social media [16].

The use of social media around the world, especially in the public sector, has different goals, such as disseminating information, developing mass collaboration, and strengthening laws and regulations [17]. The use of social media in the public sector, especially government at first and some are still continuing, such as in the United States, politicians and government agencies have used social media extensively to inform and interact with citizens, in the UK, local governments and British councils have involved citizens and local communities using social media, in Canada politicians are actively present on social media and some government departments combine social media with government websites, in India government agencies have also used social media to engage their citizens in two-way communication [18]. On the other hand the use of social media by citizens in assisting the government is also carried out through comments and feedback, one good example of cooperation initiated by citizens against the government, is reporting by citizens to the police department regarding illegal parking through the online media “MyBikeLane” or “Caughtya”, as well as other examples of collaboration between the government and citizens in a many-to-many context in achieving certain common goals, such as “Adopt a Fire Hydrant” and “Adopting a Tsunami Siren”, which are online crowdsourcing platforms [17].



## 2 Background of Study

Government social media research began in 2009 and has continued to grow to date. It was revealed that previous government social media research was generally oriented towards the government side rather than the user side, more on how the government is than how the users are, while content created by the government is still the image of the government and a political tool, instead of increasing government openness, this is what causes the failure to capture the capabilities offered by web 2.0 technology, namely users as the main actors in triggering social media interactions [19]. There is very little research on user engagement research on government social media [20, 21].

Based on a survey conducted by UN E-Government in 2020, most government portals in the world have social media networking tools [22]. This is also the case in Indonesia, where in general Indonesian government agencies have used social media as an alternative service to the public [23], this is also strengthened by the existence of a legal umbrella, namely guidance. President no. 3 of 2003, concerning the National Policy and Strategy for the Development of E-Government [24], and Regulation of the Minister of Empowerment of State Apparatuses and Bureaucratic Reform No. 83 of 2012 concerning Guidelines for the Utilization of Social Media for Government Agencies [25]. However, based on the E-Participation Index (E-PI), the level of citizen electronic participation through the Indonesian government's ICT application is 0.75, Indonesia is ranked 57th out of 193 countries in the world along with the Philippines, and is ranked 4th out of 11 countries in Southeast Asia, even though Indonesia is already above the world E-PI standard (0.5677), Asian regional E-PI (0.6294), and Asia Sub-Regional E-PI (0.6126), it is still below other countries. Other Southeast Asian countries, such as Thailand, Malaysia, and Singapore each rank third, second, and first in Southeast Asia [22].

The use of social media other than official sites by the Indonesian government has started since 2010, where this use is carried out by ministries and non-ministries, while the most commonly used social media are Instagram and Twitter which represent millennials, while Facebook and YouTube generally used by older citizens, as for the purpose of its use, conveying information related to the government, both in infographics and video-graphics, monitoring feedback from citizens regarding government information that has been submitted previously on social media, but very little feedback is received. The response is rarely even taken into consideration in making decisions [26–28]. From these problems, this study aims to find factors and indicators of citizen involvement in government social media, by evaluating the respondent's question instrument which will become a questionnaire form to capture expectations from the citizen's side in their involvement in government social media, by using content validity. So that research motivation in optimizing the use of government social media can be achieved, not just as government imagery [26–29].

### 3 Methodology

This research is a continuation of the previously conducted systematic literature review (SLR) and produced factors and indicators of citizen involvement in government social media, in a comprehensive research procedure, SLR was used to search and summarize all published papers. The SLR is processed through various stages so that it can be carried out. First, the review protocol is described and developed. Then, the inclusion and exclusion criteria are explained. After that, the search strategy is carried out, the search strategy is a scientific approach from the procedure that shows how the article sources of information about this research domain are organized and carried out. Next, the study selection process was determined. After that, the quality assessment is clarified. Then, data extraction and synthesis is described. Then proceed with the following steps, namely identifying constructs, measuring item development, and content validity.

### 4 Identified Constructs

This section presents the process of defining each related construct for this study. Referring to the SLR process for this study, 14 constructs were selected as models of citizen engagement on government social media. Table 1 presents the identified constructs and their source references.

### 5 Measurement Items Development

This study used a research instrument, namely the validation sheet instrument for expert judgment and a questionnaire instrument for respondents as buyers. In this study, 11 experts were asked to validate the instrument as presented in Table 2. However, only 5 experts were willing to validate the instrument items.

Trust in sincere intention is the first construct to be a measurement item. Where this construct looks at the extent to which citizens believe that the government has a sincere intention and is committed to listening and, if possible, implementing the ideas generated through the participatory process. Table 2 presents the process of refining the trust in sincere intention construct.

The second construct is expected personal gratification which refers to the extent to which citizens expect that they will participate bring them satisfaction and enjoyment. Table 3 presents the refinement process for the expected personal gratification construct.

The third construct is the dialogic loop, which means the extent to which governments posting content to social media can stimulate public dialogue, provide dialogue

**Table 1** Constructs identification

No	Constructs	Identification	Source
1	Trust in sincere intentions	The degree to which citizen believe the government is committed to listening to and, where possible, implementing ideas generated through the participation process	[30]
2	Expected personal gratification	The degree to which citizens believe that engagement will provide them with pleasure and fulfillment	
3	Dialogic loop	The organization which posts content to social can stimulate public dialogue, providing the dialogue channel for the public and responding to public feedback in a timely manner	[31, 32]
4	Situational motivation	In problem solving is “a state of situation-specific cognitive and epistemic readiness to make problem solving efforts” or the drive to stop and think about a problematic situation	[33]
5	Crisis efficacy	Reflects people’s beliefs about whether they can successfully perform recommended behaviors during crisis situations that are largely out of their control	
6	Community commitment	A contribution to the community (e.g., participating in community events, contributing to solving community problems)	[34]
7	Community ownership	A sense of possessions and feelings of ownership	
8	Trust in government	The degree to which citizens believe that the government will work	
9	Strength of social ties	The extent to which individuals regularly interact with others	
10	Attitude towards the behavior	An individual’s positive or negative feelings (evaluative affect) about engaging in government social media engagement behavior	[30, 34]
11	Subjective norm	The person’s perception that most people who are important to him or her think he or she should or should not perform the behavior in question	
12	Perceived behavioral control	The perceived ease or difficulty of performing the behavior	
13	Intention of citizen engagement (CE) on Government social media (GSM)	An individual’s willingness to engage in government social media engagement	
14	CE on GSM	Individual or collective behavior aimed at solving social problems in society by utilizing government social media	

**Table 2** Refining trust in sincere intention items

Section	No	Item code	Refined items
Trust in sincere intentions	1	TIS1	I think that citizen's opinions will be used by the government on the GSM
	2	TIS2	I think the government will take seriously the opinion of citizens on GSM
	3	TIS3	I think the government will seriously implement the programs with citizens through GSM
	4	TIS4	I think the government will seriously have a dialogue with citizens through live streaming on GSM

**Table 3** Refining expected personal gratification items

Section	No	Item code	Refined items
Expected personal gratification	1	EPG1	I feel happy to participate on GSM
	2	EPG2	I will feel satisfied participating on GSM
	3	EPG3	I think other citizen members will participate on GSM
	4	EPG4	I will feel proud to participate on GSM

channels for the public, and respond to public feedback in a timely manner. Table 4 presents the refinement process for the dialogic loop construct.

The fourth construct is situational motivation, which is a situation-specific cognitive state and epistemic readiness to make “problem solving efforts” or the urge to stop and think about problematic situations. Table 5 presents the refinement process for the situational motivation construct.

**Table 4** Refining dialogic loop items

Section	No	Item code	Refined items
Dialogic loop	1	DL1	I think it is important that GSM post request opinion to citizens in public dialogue
	2	DL2	I think it is important that GSM reply on every citizen's comments in public dialogue
	3	DL3	I think it is important to use direct messaging application on GSM in public dialogue with citizens
	4	DL4	I think it is important to use live streaming application on GSM in public dialogue with citizens

**Table 5** Refining situational motivation items

Section	No	Item code	Refined items
Situational motivation	1	SM1	I'm curious about the problems related to the emergency alerts posted on GSM
	2	SM2	I often think about the problems of emergency alerts posted on GSM
	3	SM3	I would like to better understand the problem of emergency alerts posted on GSM
	4	SM4	I want to help in solving the problem that occurred in the emergency alert on GSM

The fifth construct is crisis efficacy, which “reflects people’s beliefs about whether they are successful at carrying out the recommended behavior during crisis situations that are largely out of their control”. Table 6 presents the refinement process for the crisis efficacy construct.

The sixth construct is community commitment, which represents a contribution to the community (e.g., participates in community events, contributes to solving community problems). Table 7 presents the refinement process for the community commitment construct.

The seventh construct is community ownership, which is a sense of belonging and a feeling of belonging. Table 8 presents the refinement process for the community ownership construct.

The eighth construct is trust in government, which is the extent to which citizens believe that the government will work for their best interests. Table 9 presents the refinement process for the trust in government construct.

**Table 6** Refining crisis efficacy items

Section	No	Item code	Refined items
Crisis efficacy	1	CE1	I feel prepared when I know that early warnings for emergencies are posted on GSM (such as tsunami, flood, COVID-19)
	2	CE2	I have sufficient access to information on GSM in the event of an emergency (such as hospitals, police stations, SAR agencies, fire departments, disaster management agencies, ambulances, red cross, and the COVID-19 task force)
	3	CE3	I have sufficient information regarding emergency evacuation instructions on GSM in case of an emergency (such as fire, earthquake)
	4	CE4	I have sufficient information regarding post-emergency management on GSM (such as disaster recovery center)

**Table 7** Refining community commitment items

Section	No	Item code	Refined items
Community commitment	1	CC1	I actively participate in community events whose information on these activities is available on GSM
	2	CC2	I always follow the government's advice on GSM (such as staying at home during the COVID-19 pandemic)
	3	CC3	I always respond to the planning of local government regulations that are distributed on GSM
	4	CC4	I feel that I am contributing to solving problems that are communicated through GSM

**Table 8** Refining community ownership items

Section	No	Item code	Refined items
Community ownership	1	CO1	I am interested in my area on GSM
	2	CO2	I don't think I should ignore what's happening in my area through GSM
	3	CO3	I have a responsibility to care for the community on GSM by eliminating discrimination, hoaxes, pornography, hate speech, etc
	4	CO4	I take part in execution the running of government work programs published on GSM

**Table 9** Refining trust in government items

Section	No	Item code	Refined items
Trust in government	1	TIG1	The GSM contributes to the public good and social integration
	2	TIG2	The GSM contributes to bridging the gap between social classes
	3	TIG3	The GSM provides professional services
	4	TIG4	The GSM contributes to the promotion of democracy

The ninth construct is strength of social ties, which is the extent to which individuals regularly interact with other people. Table 10 presents the refinement process for the construct of strength of social ties.

The tenth construct is attitude towards the behavior, which is a person's positive or negative feelings (evaluative influence) about being involved in government social

**Table 10** Refining strength of social ties items

Section	No	Item code	Refined items
Strength of social ties	1	SOST1	I use GSM because of family encouragement
	2	SOST2	I use GSM because of neighbor's encouragement
	3	SOST3	I use GSM because of a friend's encouragement
	4	SOST4	I use GSM because of the encouragement of social group members

media engagement behavior. Table 11 presents the refinement process for the attitude towards the behavior construct.

The eleventh construct is subjective norms, which is a person's perception that most people who are important to him think that he should or shouldn't do that behavior. Table 12 presents the refinement process for the subjective norms construct.

The twelfth construct is perceived behavioral control, which is the perceived ease or difficulty in carrying out the behavior. Table 13 presents the refinement process for the perceived behavioral control construct.

**Table 11** Refining attitude towards the behavior items

Section	No	Item code	Refined items
Attitude towards the behavior	1	ATTB1	It will be a good idea for me to engage with the government through GSM
	2	ATTB2	It [will be] a wise idea for me to engage with the government through GSM
	3	ATTB3	I think it is very beneficial to interact with the government through GSM
	4	ATTB4	It is great for me to be involved with the government through GSM

**Table 12** Refining subjective norms items

Section	No	Item code	Refined items
Subjective norms	1	SN1	My family thought that I should use GSM to engage with the government
	2	SN2	My friends thought that I should use GSM to engage with the government
	3	SN3	My relatives think that I have to government use social media to engage with the government
	4	SN4	My community thinks that I should use social media by the government to connect with the government

**Table 13** Refining perceived behavioral control items

Section	No	Item code	Refined items
Perceived behavior control	1	PBC1	I use GSM to build engagement between citizens and the government
	2	PBC2	Using GSM is completely under my control
	3	PBC3	I have the necessary knowledge to use GSM
	4	PBC4	I have the skills necessary to use GSM

The thirteenth construct is intention CE on GSM, referring to the individual's willingness to engage in government social media engagement. Table 14 presents the refinement process for the intention CE on GSM construct.

The fourteenth construct is CE on GSM, individual or collective behavior aimed at solving social problems in society by utilizing government social media. Table 15 presents the refinement process for the CE on GSM construct.

The next section content validity is for validating the survey items for this study which already been refined in this section.

**Table 14** Refining intention CE on GSM items

Section	No	Item code	Refined items
Intention CE on GSM	1	ICEGSM1	I am thinking of getting engaged on GSM
	2	ICEGSM2	I would certainly think about engaging on GSM
	3	ICEGSM3	I am thinking of continuing to use GSM in the future
	4	ICEGSM4	I am thinking of recommending the use of GSM to others

**Table 15** Refining CE on GSM items

Section	No	Item code	Refined items
CE on GSM	1	CEGSM1	I often ask about social issues through GSM
	2	CEGSM2	I often have an opinion on government posts on GSM
	3	CEGSM3	I often use direct messaging applications on GSM in public dialogue
	4	CEGSM4	I often use live streaming applications on GSM in public dialogue



6 Content Validity

Content validity indicates the level of acceptance of an item in measurement based on the operational definition of the reflective constructs [35, 36]. Content validity index (CVI) is used to validate this research instrument [37]. The CVI approach is flexible and requires a minimum of 3 experts as interrater panel members [38]. Therefore, the items are validated in terms of their relevance to the content. A 4-point rating scale is used for each item to measure relevance in content. In addition, there are provisions for experts to provide comments on each item if needed. Four scales for measuring relevance were used because they are the labels most frequently used by researchers for content validity. The labels are: “Not Relevant = 1; Relevant but not important = 2; Relevant but worth reviewing = 3; Highly Relevant = 4”. The proportion agreement is 3 and 4.

In this study, eleven experts were asked to validate the instrument as presented in Table 16. However, only five experts were willing to validate the instrument. Therefore, as many as five valid ratters and used for ratings that are above the minimum score such as scale-content validity index (S-CVI) 0.8 and above can be accepted and adopted in this study [35, 36, 39]. Scale validity index is liberally interpreted as S-CVI (average), and it is calculated using average item-content validity index (I-CVI), I-CVI was computed as the number of experts giving a rating of 3 or 4, divided by the total number of experts [40]. Therefore, 54 out of 56 items were accepted and two items, namely CC4 (0.6) and CO1 (0.6) were rejected because the item scores were below the threshold.

In addition, the responses and suggestions of experts are also adjusted. Therefore, several revisions regarding word improvement, using certain terms in the government domain, such as democracy or transparency, adding to the indicator question items based on the features available on social media platforms, making sentences clear by providing examples based on features Government social media features, as well as improvements to several indicator question items that are less relevant to one of the factors or constructs proposed as models. Based on the comments of experts, the necessary improvements were made. In other words, all recommendations were implemented, and a revised questionnaire was used for the pilot study.

Table 16 List of IS experts’ invited for content validity

Serial	Status of expert	No of expert invited	Numbers of experts that respondent
1	Professors	4	1
2	Associate prof	3	2
3	Doctors	4	2
Total		11	5

## 7 Conclusion

To test the success of the citizen involvement model on government social media based on cluster of beliefs, dialogic communication theory (DCT), situational theory of problem solving (STOPS), social-mediated crisis communication (SMCC), individual social capital, and theory of planned behavior (TPB), this study establishes reliable measurement items for constructs taken from SLR. These items were developed from existing material and adapted to learning objectives. For validation, the content of the questionnaire items was assessed. The elements are verified in terms of relevance and simplicity using CVI. As a result, item counts are adjusted, and item content is edited and changed in terms of phrases. The reliability and validity of the assessment items will be tested using pilot and real surveys in this ongoing study, which may lead to suggestions for future exploration. Referring to the findings of this paper, the measurement items are significantly useful for the research area of citizen engagement on government social media and can be used by students or researchers in the same field of study.

## References

1. Seminerio M (2022) The open diary takes off. <https://www.zdnet.com/article/the-open-diary-takes>. Accessed 19 Oct 2022
2. Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. *Bus Horiz* 53:59–68. <https://doi.org/10.1016/j.bushor.2009.09.003>
3. Hootsuite W (2022) Digital 2022. Springer, New York
4. Beveridge C, Tran T (2001) Government: benefits, challenges, and tactics
5. Khasawneh S, Jalghoum Y, Harfoushi O, Obiedat R (2011) E-government program in Jordan: from inception to future plans. *Int J Comput Sci Issues* 8:568–582
6. Cegarra-Navarro JG, Pachón JRC, Cegarra JLM (2012) E-government and citizen's engagement with local affairs through e-websites: the case of Spanish municipalities. *Int J Inform Manag* 32:469–478. <https://doi.org/10.1016/j.ijinfomgt.2012.02.008>
7. Khan GF (2015) The Government 2.0 utilization model and implementation scenarios. *Inform Dev* 31:135–149. <https://doi.org/10.1177/0266666913502061>
8. Eggers WD (2007) Government 2.0: using technology to improve education, cut red tape, reduce gridlock, and enhance democracy. Rowman & Littlefield Publishers, Lanham
9. McGuire M (2006) Collaborative public management: assessing what we know and how we know it. *Public Adm Rev* 66:33–43. <https://doi.org/10.1111/j.1540-6210.2006.00664.x>
10. Chun SA, Luna Reyes LF (2012) Social media in government. *Gov Inform Quart* 29:441–445. <https://doi.org/10.1016/j.giq.2012.07.003>
11. Margetts H, Dunleavy P (2013) The second wave of digital-era governance: a quasi-paradigm for government on the Web. *Philos Trans R Soc A Math Phys Eng Sci* 371:382. <https://doi.org/10.1098/rsta.2012.0382>
12. O'Reilly T (2011) Government as a platform. *Innov Technol Gov Glob* 6:13–40. [https://doi.org/10.1162/inov\\_a\\_00056](https://doi.org/10.1162/inov_a_00056)
13. McDermott P (2010) Building open government. *Gov Inform Quart* 27:401–413. <https://doi.org/10.1016/j.giq.2010.07.002>
14. Khan GF, Yoon HY, Park HW (2012) Social media use in public sector: a comparative study of the Korean and US government. In: ATHS panel duration of the 8th international conference webometrics, informatics science 13th COLLNET Meeting, pp 23–26

15. Linders D (2019) From e-government to we-government: defining a typology for citizen coproduction in the age of social media. <https://www.sciencedirect.com/science/article/pii/S0740624X12000883>. Accessed 02 Oct 2019. <https://doi.org/10.1016/J.GIQ.2012.06.003>
16. Khasawneh RT, Abu-Shanab EA (2013) E-government and social media sites: the role and impact. *World J Comput Appl Technol* 1:10–17
17. Khan GF (2015) Models for social media-based governments. *Asia Pac J Inf Syst* 25:356–369
18. Srivastava M (2013) Social media and its use by the government. *J Public Adm Gov* 3:161. <https://doi.org/10.5296/jpag.v3i2.3978>
19. Medaglia R, Zheng L (2017) Mapping government social media research and moving it forward: a framework and a research agenda. *Gov Inform Quart* 34:496–510. <https://doi.org/10.1016/j.giq.2017.06.001>
20. Lev-On A, Steinfeld N (2015) Local engagement online: municipal Facebook pages as hubs of interaction. *Gov Inform Quart* 32:299–307. <https://doi.org/10.1016/j.giq.2015.05.007>
21. Reddick CG, Chatfield AT, Ojo A (2016) A social media text analytics framework for double-loop learning for citizen-centric public services: a case study of a local government Facebook use. *Gov Inform Quart* 34:110–125. <https://doi.org/10.1016/j.giq.2016.11.001>
22. UN (2020) UN E-government survey 2020 digital government in the decade of action for sustainable development
23. Budi NFA, Fitriani WR, Hidayanto AN, Kurnia S, Inan DI (2020) A study of government 2.0 implementation in Indonesia. *Socioecon Plann Sci* 16:100920. <https://doi.org/10.1016/j.seps.2020.100920>
24. Instruction P (2004) UU No 3 Tahun 2003, p 55
25. KEMENTERIAN PANDRBRIT (2012) Peraturan menteri pendayagunaan aparatur negara dan reformasi birokrasi republik indonesia nomor 83 tahun 2012 tentang pedoman pemanfaatan media sosial instansi pemerintah dengan. Undang-Undang
26. Sari WP, Soegiarto A (2021) Indonesian government public relations in using social media. International conference on humanities, education, law, and social sciences (ICHESS). Faculty of Social Science, State University of Jakarta, Jakarta, pp 495–508
27. Idris IK (2018) Government social media in Indonesia: just another information dissemination tool. *J Komun Malay J Commun* 34:337–356
28. Roengtam S, Nurmandi A, Almarez DN, Kholid A (2017) Does social media transform city government? A case study of three ASEAN cities: Bandung, Indonesia, Iligan, Philippines and Pukhet, Thailand. *Transf Gov People Process Policy* 11(3):343–376. <https://doi.org/10.1108/TG-10-2016-0071>
29. Azmi AF, Budi I (2018) Exploring practices and engagement of Instagram by Indonesia government ministries. In: Proceedings of the 2018 10th international conference on information technology electronic engineering smart technology better society ICITEE 2018, pp 18–21. <https://doi.org/10.1109/ICITEED.2018.8534799>
30. De Jong MDT, Neulen S, Jansma SR (2019) Citizens' intentions to participate in governmental co-creation initiatives: comparing three co-creation configurations. *Gov Inf Q* 36:490–500. <https://doi.org/10.1016/j.giq.2019.04.003>
31. Chen Q, Min C, Zhang W, Ma X, Evans R (2021) Factors driving citizen engagement with government TikTok accounts during the COVID-19 pandemic: model development and analysis corresponding author. *J Med Internet Res* 23:1–13. <https://doi.org/10.2196/21463>
32. Chen Q, Min C, Zhang W, Wang G, Ma X, Evans R (2020) Unpacking the black box: how to promote citizen engagement through government social media during the COVID-19 crisis. *Comput Hum Behav* 110:106380. <https://doi.org/10.1016/j.chb.2020.106380>
33. Liu BF, Xu S, Rhys Lim JK, Egnoto M (2019) How publics' active and passive communicative behaviors affect their tornado responses: an integration of STOPS and SMCC. *Public Relat Rev* 45:101831. <https://doi.org/10.1016/j.pubrev.2019.101831>
34. Choi JC, Song C (2020) Factors explaining why some citizens engage in E-participation, while others do not. *Gov Inf Q* 37:101524. <https://doi.org/10.1016/j.giq.2020.101524>
35. Grant JS, Davis LL (1997) Selection and use of content experts for instrument development. *Res Nurs Health* 20:269–274. [https://doi.org/10.1002/\(sici\)1098-240x\(199706\)20:3%3c269::aid-nur9%3e3.3.co;2-3](https://doi.org/10.1002/(sici)1098-240x(199706)20:3%3c269::aid-nur9%3e3.3.co;2-3)

36. Polit DF, Beck CT (2006) The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res Nurs Health* 2:488–495. <https://doi.org/10.1002/nur>
37. Lynn MR (1986) Determination and quantification of content validity. <http://ijoh.tums.ac.ir/index.php/ijoh/article/view/26>
38. Tojib DR, Sugianto L-F (2006) Content validity of instruments in is research. *J Inf Technol Theory Appl* 8:31–56
39. Robinson MA (2018) Using multi-item psychometric scales for research and practice in human resource management. *Hum Resour Manag* 57:739–750. <https://doi.org/10.1002/hrm.21852>
40. Shrotryia VK, Dhanda U (2019) Content validity of assessment instrument for employee engagement. *SAGE Open*, New York, p 9. <https://doi.org/10.1177/2158244018821751>

# Shared Parking Concept in the Smart City Environment



Zuzana Špitálová<sup>✉</sup>, Lucia Mandová, and Martin Opatovský

**Abstract** With the modern city infrastructure and increasing number of subjects in the traffic environment, there is a need for increased number of parking places. As the parking place is usually not used for a whole day, it could be shared with the others in the certain community. As it eliminates the useless driving due to the finding of free parking place, the traffic becomes more effective and produces less emissions. We designed the application for smart city environment, which helps to achieve the described situation. It consists of two parts, the frontend and backend. The data processed by the backend part can be further used for the city infrastructure improvements. The front-end part, designed in React Native, represents the graphical interface. The back-end part, designed in PostgreSQL, represents the database.

**Keywords** Smart city · PostgreSQL · React Native · City cloud · Infrastructure · API

## 1 Introduction

In the last years, the popularity of smart city development is increased dramatically. The idea of modern and intelligent cities is present more than a decade. The concepts related to smart cities are mentioned in [1]. According to [2], smart city can be defined as the place which uses the digital solutions for traditional networks and services for the benefit of the business and the people which live there. It connects the information and communication technologies (ICTs) with various aspects needed for the modern urban development. There are various technologies belonged to the ICTs, like artificial intelligence, machine learning for face recognition, autonomous vehicles, sensorical systems used in the cities or algorithms for service improvements including the high level of security [3, 4]. The possibility of effective smart phones

---

Z. Špitálová (✉) · L. Mandová · M. Opatovský  
Faculty of Informatics and Information Technology, Institute of Computer Engineering  
and Applied Informatics, Slovak University of Technology, Bratislava, Slovakia  
e-mail: [zuzana.spitalova@stuba.sk](mailto:zuzana.spitalova@stuba.sk)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information  
and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_68](https://doi.org/10.1007/978-981-99-3091-3_68)

833

**Table 1** Examples of ICTs [6]

Smart city technologies		
4G	Data visualization	IoT
5G	Data centres	IoS
Advanced analytics	Digital modelling	LoRa
Artificial intelligence	Drones	Network slicing
Big data	Geoinformation	Open data access
Cameras	Distributed architecture	Robotics
City platform	GPS	Service integration
Cloud	Hardware development	Smart card
Blockchain databases	Information technology	Virtual reality
Computer networking	Infrastructure development	Wifi access points

usage is also welcomed in this environment. From the ICTs point of view, the most important item of smart city is Internet of Things (IoT) [5]. This means, the network of sensorical and non-sensorical devices, which are connected, is able to be identified and communicated between each other (Table 1).

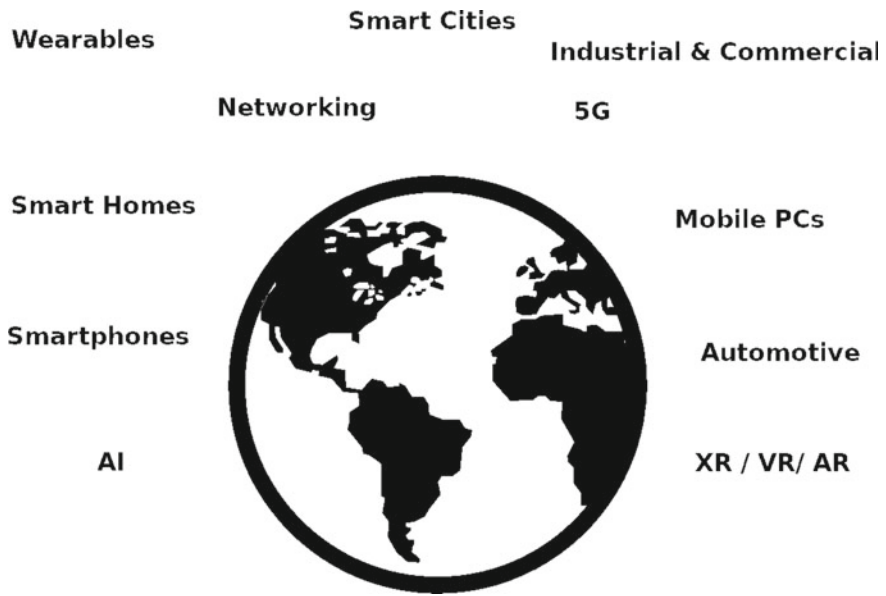
The big cities improve their infrastructure and services for achieving the certain level of smart city. According to the Juniper Research [6] from the year 2016, the World's 5 Smartest Cities were defined, including the following ones:

- Singapore
- Barcelona
- London
- San Francisco
- Oslo.

This conclusion was arrived by ranking the factors like the adoption of smart grid technologies, the intelligent lighting usage, the Wi-Fi access points availability, and the others. The other studies are mentioned in [6], too. There are various standards for ranking the cities. Some of them consider the availability of applications, including the traffic apps or city information apps.

Integration of the technologies, sensors, and any personal device in smart city can help to gain and share data which are further used for improving the services or finding the information. The good example is the intelligent transport system (ITS) which consists of all the traffic information gained from the connected devices and from the wireless devices situated in the traffic. This means, every subject in the environment can contribute to the data gaining and sharing. The data are not only shared between the certain group, like ITS but also between multiple infrastructures.

Due to the multiple infrastructures and sharing the information between them, there was a study [8, 9] regarded to the citizen cloud. In this study, the city of Shanghai was ranked as a city with the best implementation of smart city elements. Shanghai created the intelligent data platform called Citizen Cloud which provides



**Fig. 1** Smart city vision [7]

the public services, like education, culture, health care, public transportation, etc. Basically, the information regarded to the city services can be found there.

Regarding to the apps, which can be used in the smart city environment, the choice of right technology has a significant importance. The main priority is the possibility for multiplatform usage. To achieve this, React Native [10] is a good choice. It is the open-source multiplatform framework based on JavaScript. Thanks to the big community participating in the development, React Native offers the huge amount of APIs. So, it is not necessary to develop everything by yourself. As it promises the code sharing between various development platforms, the time needed for development is shortened (Fig. 1).

From the database point of view, there is also the importance to use the optimal one. The good choice is the open-source alternative. The database systems can be relational and non-relational. The main difference between them is the data storage. In the relational ones, there are the data stored in the tables and the strict structure is required. So, this means, the connections are known. In the case of non-relational ones, the data are not stored in the tables and the structure can change. As we plan to work with the huge amount of data in our case, we chose PostgreSQL [11] database, which is object-relational database system. It offers more complex data types and allows object inheritance.

## 2 Design

The idea of our work was to design an application for shared parking places between the people in the certain community, consisted of front-end and back-end part. The users can offer their own parking places to the others during the hours when they are not used. The user can be everyone, the private person or the subject which owns the parking places in the parking lot. The advantage of our app is the data gained by the back-end which can be processed and used for other purposes.

### 2.1 *Front-End*

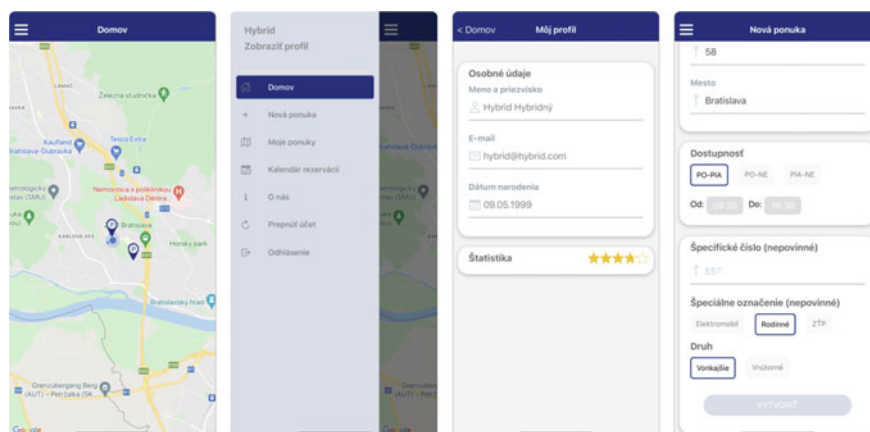
Our proposal consists of two parts, web application and front-end mobile application. Basically, both are the same but the mobile one can handle also the actual GPS position data and taken photos of parking place. It is programmed in the open-source multiplatform framework, React Native. So, it can be used by multiple mobile platforms like Android or iOS. Web is basically designed for big companies which own the parking lots and can share the places outside the working hours.

The application has two roles, the provider one and the user one. It is not necessary to use the both roles, means users can be only the providers or vice versa. Registration is strictly necessary, with respecting a good manners. So, only an owner is able to provide the parking place. After registration, the data are stored in two places. At first, in the local trusted storage, secondly at the backend where was sent in secure way.

After login, the user has to choose which role wants to use, provide or search the place. Application requires the coordinate data, so user has to provide an access to them by enabling the GPS on the mobile or on the web. Once this is done, data are processed by the Google API and sent to the back-end. We chose the free variant, because it was enough for our usage.

In the case of searching mode, it is necessary to enter the target address, the distance to selected target, date and the hour as the mandatory items. There is also the possibility to choose the type of place, like inside, outside, charger availability, place for disabled person, etc. After filling the request, available places are displayed including the map. There are the multiple options to sort the data by available time, distance, popularity, and type. After the selection, it is possible to check the availability for the next day. We disabled the option to book the place for longer than one week to prevent from blocking the place. Then, there is an option to use the Google API to be navigated to the selected place. After the parking, the mark is required to inform the others that the place is occupied. After leaving the place, the check out from parking is required. If something was wrong with the place, it is possible to leave the message or the photo of the current status. At the end, the service can be favorited.





**Fig. 2** Navigation part of our app

As we have mentioned already, in the case of providing mode, it is necessary to be the owner of the place which you want to share. Then, the information about the owner, location, type, photo and any additional data are required. After that, the place is registered for sharing in the available time which is entered. In the case of parking lot, there is also the possibility to put the availability according to the sensor below the parking place which indicates the usage. The availability is processed by the smart building and shared with the app. In the case of left message by the user, there is an app notification for the provider of the parking place. The reservation could be cancelled by the provider only one day before the usage, not during the reserved sharing.

The front-end part of application was evaluated just in the virtual environment. We created virtual customers which tried to create/share the parking places. There were also the tests for places reservation. We faked the response from backend and tried to sort them by selected available items. At the end, we tested the requests from multiple users in the same time. It means, firstly, we tested the app without backend server interaction. Secondly, we did the same with backend interaction (Fig. 2).

## 2.2 Back-End

Our proposal of back-end part is designed in Python with Django modules. The real server is Ubuntu Container. Inside there is running the PostgreSQL database and Nginx server which provides https connection. We required the Relational Database Management System (RDBMS) to be provided the relational database management functionality. The data are stored in the tables while RDBMS ensures the integrity and rules between databases. That was the reason why we chose the PostgreSQL

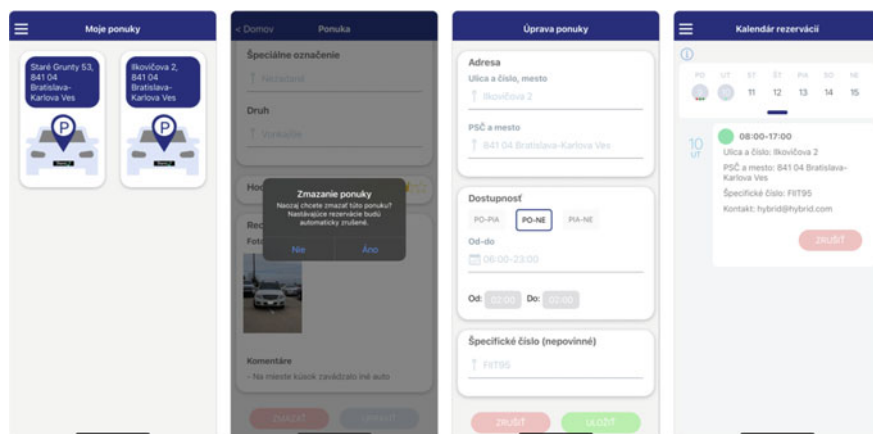


Fig. 3 Parking place reservation

database. By design, this runs in the asynchronous way to be able to process multiple requests in the same time. There was implemented the locking mechanism which prevents from ordering the same place in the same time by different users. As it was mentioned in the front-end part, the user credentials are stored like the hashes to prevent from misusing the data. The back-end processes the front-end requests like user registration, login, place administration, etc. Basically, it stores and processes all the data which has to be shared between users (Fig. 3).

In Fig. 5, there is seen an example of parking place creation query. It creates the parking place reservation only if it is not already created. The data are filled according to the ones which are stored in other tables, like Days. This table contains the data about the parking place availability. At the end, all tables are filled according to the specified values from user.

The server processes requests for the multiple APIs. Basically, they can be divided into two groups, usage and statistical group. The usage group provides the APIs for registration, login and user services. It also contains the calculation logic for parking slot availability. Basically, it handles all the parking management but without processing the statistics. That is the job of the second group, which can be used by external platforms belonged to the smart city. The statistical data can be exported in multiple sorted way, like by time, type of the parking place, location, etc.

The first testing was targeted for user management, like registration, login, and token generating. The following tests were targeted to parking place handling. We filled the database with faked data for simulating the requests from frontend. At the end, we simulated the multiple requests for getting the same items.

This is an example of the real API used in application:

1	Folder	Contents
2	/BE\_parking	
3	/User	API for user management
4	login	API for login
5	signup	API for registration
6	check user	API for verification
7	provider rating	API for avarage rating
8	/Parkings	API for parking place management
9	create parking	API for new parking place creation
10	parkings	API to get all parking places
11	provider parking	API to get a selected parking place
12	provider parkings	API to get provider's parking places
13	parking	API to get the parking slot
14	delete parking	API to delete parking places
15	update parking	API to change the parking place
16	parking exhaustion	API to get the load of parking place
17	parking filter	API to get the filtered place
18	/Reservation	
19	create reservation	API to create the reservation
20	get res	API to get the reservation
21	get res provider	API to get the reservation at own parking places
22	delete res	API to delete the reservation
23	/Review	
24	create review	API to put the review
25	get reviews	API to get the review
26	/Googleapis	
27	get formal adress	API to get the location
28	/Favourites	
29	create fav	API to create the favourite parking places
30	get fav	API to get the favourite parking places
31	delete fav	API to delete the favourite parking places

After finishing the both parts, the application was evaluated by testing in the real area using the Android mobile phones. As the app is the prototype, the testing was made just by the developers. It was not tested by more people due to the lack of resources (Figs. 4 and 5).

### 3 Future Work

As our app is in the prototype phase, we plan to do more complex testing involving more users. So, it can be moved from close source to open source for the quality improving. If we move the app for example to Github, there are the options to track the bugs, feature requests, or the possibility for everyone to be the part of the team.

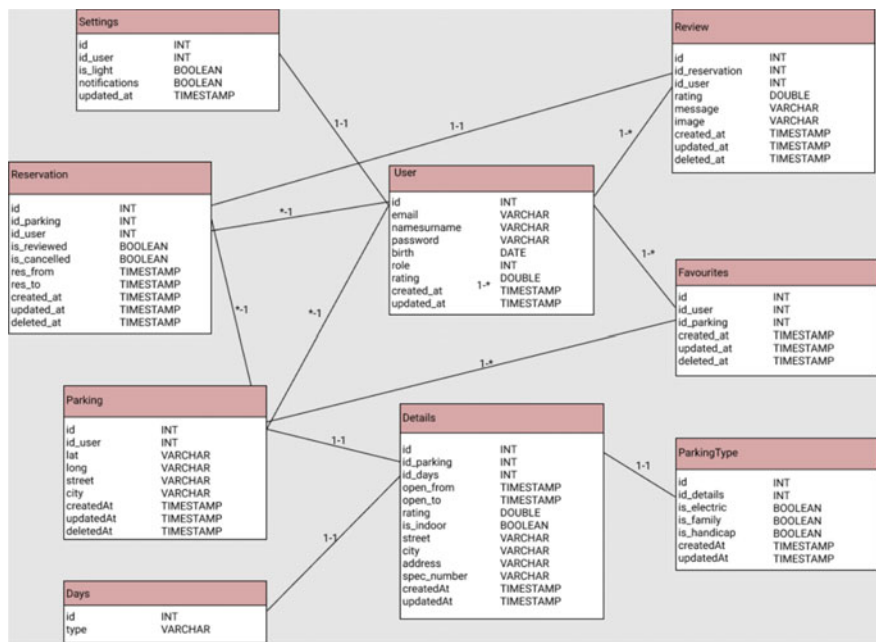


Fig. 4 Back-end database example

```
Executing (default): SELECT "id", "spec_number", "id_days", "id_parking", "open_from", "open_to", "rating", "is_indoor", "street", "city", "address", "createdAt", "updatedAt" FROM "parking"."Details" AS "Details" WHERE "Details"."spec_number" = 'E68' AND "Details"."address" = 'Dubová 10, 818 07 Ľilina-Solinky, Slovakia' LIMIT 1;
Executing (8f186a63-6b04-488b-93c2-a63941c34e19): START TRANSACTION;
AS "User" WHERE "User"."id" = 3;
Executing (8f186a63-6b04-488b-93c2-a63941c34e19): INSERT INTO "parking"."Parkings" ("id", "id_user", "lat", "long", "street", "city", "createdAt", "updatedAt") VALUES (DEFAULT, $1, $2, $3, $4, $5, $6, $7) RETURNING "id", "id_user", "lat", "long", "street", "city", "deletedAt", "createdAt", "updatedAt";
Executing (default): SELECT "id", "type" FROM "parking"."Days" AS "Days" WHERE "Days"."type" = 'PO-NE' LIMIT 1;
Executing (8f186a63-6b04-488b-93c2-a63941c34e19): INSERT INTO "parking"."Details" ("id", "spec_number", "id_days", "id_parking", "open_from", "open_to", "rating", "is_indoor", "street", "city", "address", "createdAt", "updatedAt") VALUES (DEFAULT, $1, $2, $3, $4, $5, $6, $7, $8, $9, $10, $11, $12) RETURNING "id", "spec_number", "id_days", "id_parking", "open_from", "open_to", "rating", "is_indoor", "street", "city", "address", "createdAt", "updatedAt";
Executing (8f186a63-6b04-488b-93c2-a63941c34e19): INSERT INTO "parking"."ParkingTypes" ("id", "id_details", "is_electric", "is_handicap", "is_family", "createdAt", "updatedAt") VALUES (DEFAULT, $1, $2, $3, $4, $5, $6) RETURNING "id", "id_details", "is_electric", "is_handicap", "is_family", "createdAt", "updatedAt";
Executing (8f186a63-6b04-488b-93c2-a63941c34e19): COMMIT;
```

Fig. 5 Parking place creation query

We want to make the connection more secure and improve algorithms. We think that CI/CD pipeline can be also the improvement, so the code quality can be automatically checked. The other aim is the statistical part of the app. Actually, we collect the data which can be used for further processing. We plan to implement the data export mechanism which will be able to support multiple formats. This could be used like an import for other applications (maybe a big companies) which will be able to use the data for the improvements of smart city.

## 4 Conclusion

In this paper, we describe the shared parking application which is able to find and offer the free parking place for everyone who uses it. All the concept is designed to be for free. In future, the possibility of donating can be implemented due to the infrastructure growing. The servers could be splitted into multiple decentralized clusters. Implementing this, the data will be stored at more places, which improves the stability of the network. Actually, the collected data untill now represents the actual state in the certain area. It is possible to identify the busiest places, times, and type of the car, which drives into the certain zone. According to this data, we have the various useful information. The infrastructure can be improved according to this. Let us imagine that the parking place via the sensors can check the availability and share its status through this app. This process could be fully automated without any user interaction. The other advanatge is the possibiliity to predict the energy consumption in the parking lots using the chargers. It can also lead to the increasing or decreasing of parking places in the zones.

**Acknowledgements** This work was partially supported by the project SANETII.

## References

1. Camero A, Alba E (2019) Smart city and information technology: a review. *Cities* 93:84–94
2. E. Commission (2022) Smart cities. [https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities\\_en](https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en). Accessed 2022-12-10
3. Noori N, de Jong M, Janssen M, Schraven D, Hoppe T (2020) Input-output modeling for smart city development. *J Urban Technol* 28:71–92
4. S. Engineering (2022) Cities strive for more smarts, security. <https://semiengineering.com/cities-strive-for-more-smarts-security/>. Accessed 2022-12-10
5. Atzori L, Iera A, Morabito G (2018) The internet of things: a survey. *Comput Netw* 54:2787–2805
6. Yavuz MC, Cavusoglu M, Corbaci A (2018) Reinventing tourism cities: examining technologies, applications, and city branding in leading smart cities. *Int Interdisc Bus-Econ Adv J* 3:57–70
7. Qualcomm (2022) Qualcomm company webpage. <https://www.qualcomm.com>. Accessed 2022-12-10
8. Yu J, Wen Y, Jin J, Zhang Y (2019) Towards a service-dominant platform for public value co-creation in a smart city: evidence from two metropolitan cities in China. *Technol Forecast Soc Change* 142:168–182
9. Templeman V-B (2022) Juniper research names Shanghai the world's no. 1 smart city. <https://www.digitalnationaus.com.au/news/juniper-research-names-shanghai-the-worlds-no-1-smart-city-575086>. Accessed 2022-12-10
10. R. Native (2022) React native webpage. <https://reactnative.dev/>. Accessed 2022-12-10
11. PostgreSQL (2022) Postgresql webpage. <https://www.postgresql.org/>. Accessed 2022-12-10

# Applying Machine Learning Techniques to the Analysis and Prediction of Financial Data



Pablo Flores-Siguenza , Darío Espinoza-Saquicela ,  
Marlon Moscoso-Martínez , and Lorena Siguenza-Guzman 

**Abstract** Data analysis and processing allow for acquiring competitive advantages both in the business and academic and research worlds. One of the sciences that carries out this analysis is machine learning, which has evolved with greater emphasis in recent years due to its advantages and applicability in different areas. Aware of the importance and current relevance of data management for industries, especially in the banking sector, this study applies supervised learning techniques to generate classification and prediction models by treating a set of data from an Ecuadorian financial institution. Different algorithms are compared, and each of the steps to follow in constructing the models is explained in detail. This allows the financial entity to classify its clients as VIPs or not with greater certainty, as well as to predict the investment amounts of the potential clients based on variables such as age, occupation, and among others. The main results show that the  $K$ -nearest neighbor algorithm with  $k = 5$  is optimal for classification, while for prediction, the multilayer perceptron algorithm is the most favorable.

**Keywords** Machine learning · Data analysis · Classification model · Prediction model · Financial industry

---

P. Flores-Siguenza

Department of Applied Chemistry and Systems of Production, Faculty of Chemical Sciences, Universidad de Cuenca, Cuenca, Ecuador

D. Espinoza-Saquicela

Institute of Sectional Regime Studies of Ecuador, Universidad del Azuay, Cuenca, Ecuador

M. Moscoso-Martínez

Faculty of Sciences, Escuela Superior Politécnica de Chimborazo (ESPOCH), Panamericana Sur Km 1 1/2, 060106 Riobamba, Ecuador

L. Siguenza-Guzman (✉)

Department of Computer Sciences, Faculty of Engineering, Universidad de Cuenca, Cuenca, Ecuador

e-mail: [lorena.siguenza@ucuenca.edu.ec](mailto:lorena.siguenza@ucuenca.edu.ec)

Research Centre Accountancy, Faculty of Economics and Business, KU Leuven, Leuven, Belgium

# 1 Introduction

The amount of data currently generated in companies is increasing exponentially. Therefore, analyzing and extracting valuable information from them is a competitive advantage that cannot be underestimated [1]. This has aroused the interest of the scientific world in generating models and techniques of a different nature that best fulfill this purpose. Among the most used methods in data analysis are genetic algorithms, time series, data mining, and machine learning (ML) [2], the latter being the study basis due to its development and relevance in recent years.

ML is a scientific discipline in the field of Artificial Intelligence (AI). It is a data analysis method that automates the construction of analytical models [3]. It is based on the idea that systems can learn from data, identify complex patterns, predict future behavior and make decisions [4]. This learning method is broadly classified into supervised learning [5] and unsupervised learning [6]. The general learning procedure is as follows. First, the data are divided into two subsets known as training and evaluation data. Then, the model is trained using the first subset. And the prediction performance of the model is evaluated using the second subset [7].

ML has been successfully adopted in many fields, such as image recognition [8], time series analysis [9], and language processing [10]. Recent surveys also show that credit institutions are increasingly adopting ML tools due to the opportunities offered [11, 12]. For example, areas of interest include analysis of consumer interactions [13], risk management as well as credit underwriting [14] and regulatory compliance [15]. Similarly, globally, banks have been using AI for many years, although initially only for a few applications. However, they have just realized the full potential of these advances, which has resulted in a considerable increase in their investments [16]. This world reality is also affecting the financial institutions of Ecuador, which have taken their first steps in the application of ML techniques. Unfortunately, no studies or case studies are still related to its applicability, which gives this study originality.

This study aims to develop step-by-step supervised learning techniques in classification and prediction models in an Ecuadorian financial institution, evidencing their importance and usefulness by comparing several algorithms. To this end, a unique and anonymized database of individual credit portfolios provided by a Credit Union in the province of Azuay is used. The generation of these models serves as a support tool for the financial institution to classify its VIP clients and predict the investment amounts that their clients will have. The rest of this document is structured as follows. Section 2 discusses the methodology. Section 3 explains the steps in building the model. Section 4 sets out the main results. Finally, Sect. 5 highlights the main conclusions.

## 2 Methodology

The study was developed based on a three-phase methodology described below.

**Dataset extraction:** It contains two macro activities: the search and the selection. Regarding the former, an extensive investigation was carried out in open-access digital databases. As a result, financial institutions in the city were contacted, reaching an agreement with the Cooperative under study. Regarding the latter, i.e., the selection of the dataset to be used, although the open-access databases on the web presented more fields and information, these did not reflect the national reality. For this reason, the knowledge of the Banking Cooperative of Azuay-Ecuador province was selected.

**Dataset pre-processing:** Phase dedicated to data pre-processing to clean junk information, complete missing information, and generate useful information through a rigorous statistical process to learn and validate models.

**Model development:** in this phase, the respective classification and prediction models were first developed using Weka software and some algorithms such as ZeroR, K-nearest neighbors, and multilayer perceptron. Finally, the results were analyzed, and the configuration parameters were modified to optimize the models.

## 3 Model Generation

This section details the activities carried out in stages 2 and 3 of the methodology.

### 3.1 Dataset Pre-processing

The selected dataset contains information on the credit portfolio of a financial institution in Ecuador for the period 2019–2022. It consists of 4118 observations and 15 variables, of which 5 are quantitative (age, investment, checkbook balance, card balance, sum of campaigns) and 10 are qualitative (occupation, marital status, education, personal loan, campaign 1–2–3,4, VIP client, gift received).

With the help of Excel, using statistics (mean, median, mode, minimum, maximum) and descriptive graphs (bars, histogram), some anomalies were identified and summarized in Table 1.

Regarding correcting these anomalies found on the age variable, it was decided to select the records between 18 and 100 years of age. On the other hand, for the variables Marital Status and Occupation, considering that their missing records do not represent even 1% of the data, it was decided to eliminate them as they were deemed non-representative. Finally, for the Education variable, since it has several missing entries that represent 4.05% of the total data, it was decided re-label through an analysis of contingency tables, related variables (Occupation and Marital Status), and



**Table 1** Summary of anomalies found in the dataset

Variable	Anomaly type	# Data	% Representation
Age	Negative values and >100 years	3	0.07
Marital status	Missing data	11	0.26
Occupation	Missing data	39	0.94
Education	Missing data	167	4.05

imputation of the missing entries based on the weights that each category contributes to the analyzed segment. Thus, obtaining a clean database, free of anomalies with 4065 observations.

The next step of the data pre-treatment consisted of normalizing some quantitative variables (Investment, Checkbook Balance, and Card Balance). Since they have values on different scales, normalization is used based on each variable's maximum and minimum, thus placing them on a scale between 0 and 1 [17].

As a final activity of the data pre-processing stage, an initial selection of the variables of interest was made. This was developed according to its potential contribution to the classification and prediction models to be developed, thus reducing the 15–10 variables, eliminating: Personal Loans, Campaigns 1–4.

### 3.2 Classification Model

VIP Client is chosen as the output variable of the developed classification model; the variable is dichotomous, with a YES and NO response, its relative frequency (Yes = 3.91%, No = 96.09%) shows imbalance between its classes. Regarding the input variables, an analysis of contingency tables was carried out to select those related to the VIP Client, configuring the variables Education, Occupation, Marital Status, Investment, Checkbook Balance, Card Balance.

Performing a classification with  $K$ -nearest neighbors ( $K = 6$ ) and ignoring the imbalance of the “VIP Client” class, an accuracy of 96.02% is obtained. However, when analyzing the confusion matrix, it is noted that the model has specialized in classifying individuals from the predominant class, which does not come close to the generalization capacity that a model should have.

Once the need to balance the data is demonstrated, a new dataset is created in which the abundant class is reduced to “Non-VIP Client” by randomly taking an equal number of individuals to the scarce class, a process known as under-sampling [18]. The result of the subsampling process gives a balanced sample with 318 individuals, where the result of the VIP Client variable is 50% YES and 50% NO.

The balanced dataset was loaded into the Weka software, then, a pre-treatment was done to assign the correct category to each variable type, whether numerical or

categorical. Regarding the algorithms used in the execution of the model, ZeroR, and K-nearest neighbors were used. In addition, the test modes cross-validation (tenfolds) and division into training and test percentages [19] were chosen for each. Despite not being a very complex implementation algorithm, ZeroR provides a beneficial first approach.

### 3.3 Prediction Model

For the prediction model, the investment variable was chosen as the output. On the other hand, the input variables were selected through a statistical analysis based on the correlation of quantitative variables. The result of this analysis showed that the investment variable is strongly related to the age variable, so age was one of the inputs of the prediction model together with the variables Education, Occupation, and Marital Status. The model then allows the financial institution to have a tool to predict the investment amounts of a client based on the input variables analyzed.

The algorithms used to develop this prediction model were ZeroR and a multi-layer perceptron neuronal network [19]. For each one, cross-validation test modes (tenfolds) divided into training and test percentages are chosen. For the generation of the models, the total number of records of the dataset has been used.

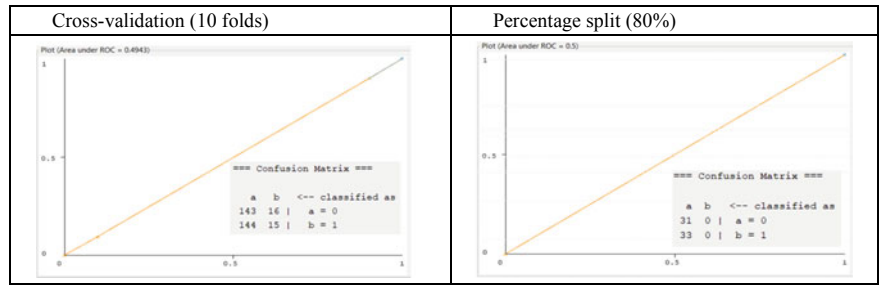
## 4 Results

### 4.1 Classification Model

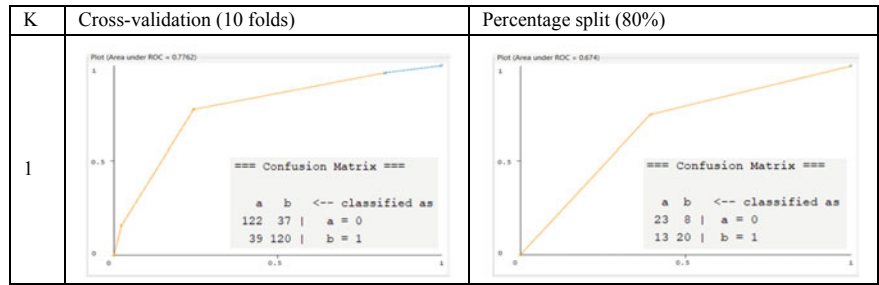
In the first instance, the ZeroR algorithm, with a “Cross-validation (tenfold)” test option, obtains an accuracy of 49.69% and the receiver operating characteristic (ROC) shows that the model is not functional. Its ability to generalize is omitted since it has specialized in classifying individuals of the class “Client not VIP (0)” despite managing to classify some individuals as “Client VIP”. While for the test option “Percent division (80%)”, an accuracy of 48.44% is obtained. Like the previous model, it has specialized in classifying class 0 clients. These results are shown in Fig. 1.

The ZeroR algorithm has given the first approximation. The K-nearest neighbors algorithm is then used with the “Cross-validation (tenfold)” and “Percent Split (80%)” test options. Figures 2, 3, 4 show the results obtained by varying the value of K.

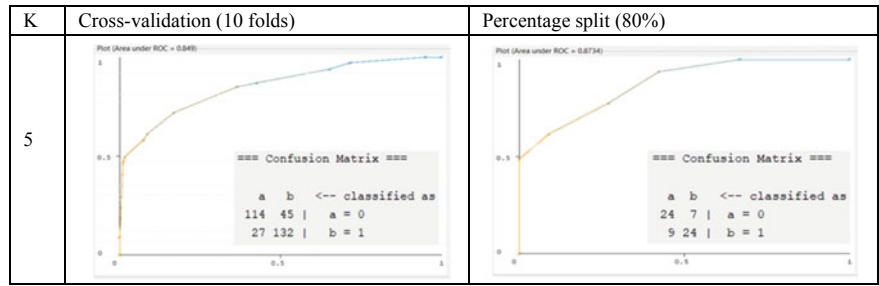
The metrics considered to evaluate the generated models are the true positive rate (TPR), the false positive rate (FPR), the precision, the F-measure, and the ROC area. These are summarized in Table 2.



**Fig. 1** Results classification model–ZeroR algorithm

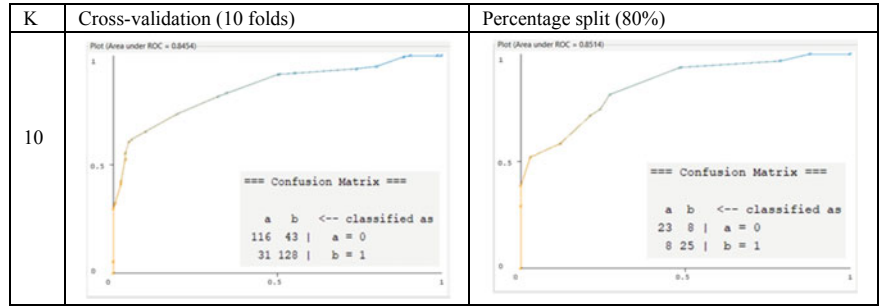


**Fig. 2** Results classification model– $K$ -nearest neighbors algorithm,  $K = 1$



**Fig. 3** Results classification model– $K$ -nearest neighbors algorithm,  $K = 5$

From the analysis of Figs. 2, 3, 4 and Table 2, the best results are obtained using the  $K$ -nearest neighbor classifier with  $K = 5$ . Its metrics have close values, achieving between the two test options the highest levels of measurement  $F$  and accuracy in the model generated with the “Cross-validation (tenfolds)” option. Despite obtaining a lower ROC versus “Split Percentage (80)”, its threshold curve behaves closer to the ideal. The confusion matrices for both cases concentrate the most significant amount of data on the main diagonal, unlike their variants with  $K$  values equal to 1 and 10.



**Fig. 4** Results classification model– $K$ -nearest neighbors algorithm,  $K = 10$

**Table 2** Metrics of classification model– $K$ -nearest neighbors algorithm

<i>K</i> value	Metrics	Cross-validation (tenfolds)	Percentage split (80%)
1	TPR	0.761	0.672
	FPR	0.239	0.324
	Precision	0.761	0.678
	F-measure	0.761	0.671
	ROC area	0.776	0.674
5	TPR	0.774	0.750
	FPR	0.226	0.249
	Precision	0.777	0.751
	F-measure	0.773	0.750
	ROC area	0.849	0.873
10	TPR	0.767	0.750
	FPR	0.233	0.250
	Precision	0.769	0.750
	F-measure	0.767	0.750
	ROC area	0.845	0.851

4.2 Prediction Model

First, the results obtained using the ZeroR algorithm are analyzed. The “Cross-validation (tenfold)” and “Percent Split (80%)” test options yield low correlation coefficients, with  $-0.04$  and  $0$ , respectively. As a result, its mean absolute errors (MAE) and mean square errors (MSE) are low. However, the relative absolute error (RAE) and the root relative square error (RRSE) were  $100\%$ , which shows the deficiency of the model generated with this algorithm.

Next, a multilayer perceptron neural network was used, varying the learning rates by 0.3, 0.5, and 0.7. For the first approach, the software could generate the hidden layers in the default mode. The metrics obtained are shown in Table 3.

The best of the six models generated by the neural network was obtained with the default learning rate of 0.3 and the cross-validation test mode (tenfolds). As a result, its correlation coefficient expressed a better approximation between the input and output variables than the other five models. Similarly, its RAE and RRSE errors were the lowest of the models, followed by the metrics obtained by the model generated with a learning rate of 0.5 and “Cross-validation (tenfolds)”.

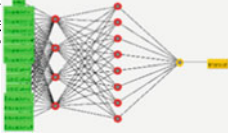
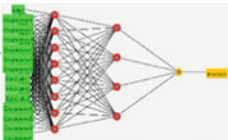
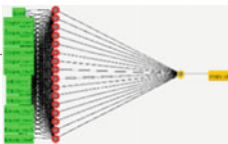
So far, the best test option, “Cross-validation (tenfolds)”, was used in the following approach to generate models with variations in the hidden layer. For the next process, the number of neurons in the hidden layers was varied, and the learning rate of 0.3 was kept since it yielded the most encouraging results. The results and metrics of these variations are seen in Table 4.

Table 4 shows that the results are very similar in this group of three models. The model that uses a hidden layer with 8 and 4 neurons is the one that provides a better correlation coefficient between the input and output variables, minimizing the RAE and RRSE metrics.

**Table 3** Metrics of prediction model—neuronal network

Learning rate	Metrics	Cross-validation (tenfolds)	Percentage split (80%)
0.3	Correlation coef	0.763	0.615
	MAE	0.021	0.024
	MSE	0.032	0.047
	RAE	74%	85%
	RRSE	66%	83%
0.5	Correlation coef	0.752	0.581
	MAE	0.021	0.032
	MSE	0.033	0.052
	RAE	74%	112%
	RRSE	66%	92%
0.7	Correlation coef	0.717	0.528
	MAE	0.023	0.033
	MSE	0.035	0.053
	RAE	82%	117%
	RRSE	71%	94%

**Table 4** Metrics of prediction model—Neuronal network—Cross-validation (tenfolds)—Variations in the hidden laver

<div><div><div>Hidden layer</div><div>4 y 8</div></div></div>	Metrics		Cross-validation (tenfolds)	
	Correlation Coef		0.763	
	MAE		0.021	
	MSE		0.032	
	RAE		74%	
	RRSE		66%	
<div><div><div>Hidden layer</div><div>8 y 4</div></div></div>	Correlation Coef		0.752	
	MAE		0.021	
	MSE		0.033	
	RAE		74%	
	RRSE		66%	
<div><div><div>Hidden layer</div><div>16</div></div></div>	Correlation Coef		0.717	
	MAE		0.023	
	MSE		0.035	
	RAE		82%	
	RRSE		71%	

5 Conclusions

The databases stored in companies or institutions are subject to errors of various kinds or lack of information. For this reason, it is of the utmost importance to review all the data in general in the first instance. These errors and anomalies can be found with the help of different statistics and descriptive graphs.

The decisions to carry out an adequate pre-treatment of the data must be based on statistical criteria. For this study, frequency and contingency tables were used, which made it possible to complete the missing data and eliminate the anomalies presented adequately.

In the case of the classification model, data balancing was a fundamental step that allowed improving the metrics of the generated model significantly. However, using an unbalanced dataset in classification algorithms increases the possibility of poor results due to the abundant presence of individuals for a particular class. In the present study, this problem was overcome using the subsampling technique, which generated a balanced sample with 318 individuals.

The ZeroR and *K*-nearest neighbors algorithms were used in the classification model, the latter being the best with a value of *K* = 5. For the test phase, two options were proposed, “Cross-validation (tenfolds)” and “Percent Split (80%)”.

The ZeroR and multilayer perceptron neural network algorithms are used in the prediction model, the latter being the best with a learning rate of 0.3. In the prediction model, similar results are obtained between the three generated options by varying

the number of neurons in the hidden layer. However, it can be said that the variation with 8 and 4 neurons in the hidden layer is the one that provides a better correlation coefficient between the input and output variables.

Thanks to the classification and prediction models elaborated in this work, the importance of supervised learning techniques, the strength of this type of application in data management, and its applicability in the national context are demonstrated. In addition, given that a tool has been generated that allows the financial institution under study to more accurately classify its clients into VIP and non-VIP, it will also be able to predict with greater certainty the investment amounts of its clients, thus generating information base for the planning of its operational management and opening the range for future applications of machine learning in the Ecuadorian banking system.

**Acknowledgements** The authors would like to thank to “Vicerrectorado de Investigación” of the University of Cuenca, Ecuador, for the financial support given to the present research, development, and innovation work.

## References

1. Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. *J Bus Res* 70:263–286. <https://doi.org/10.1016/j.jbusres.2016.08.001>
2. Agarwal R, Dhar V (2014) Editorial—big data, data science, and analytics: the opportunity and challenge for IS research. *Inform Syst Res* 25:443–448. <https://doi.org/10.1287/isre.2014.0546>
3. Kubat M (2021) Ambitions and goals of machine learning. In: Kubat M (ed) *An introduction to machine learning*. Springer International Publishing, Cham, pp 1–15
4. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci* 116:22071–22080. <https://doi.org/10.1073/pnas.1900654116>
5. Crisci C, Ghattas B, Perera G (2012) A review of supervised machine learning algorithms and their applications to ecological data. *Ecol Model* 240:113–122. <https://doi.org/10.1016/j.ecoimodel.2012.03.001>
6. Ghahramani Z (2004) Unsupervised learning. In: Bousquet O, von Luxburg U, y Rätsch G (eds) *Advanced lectures on machine learning: ML summer schools 2003*, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised lectures. Springer, Berlin, Heidelberg, pp 72–112
7. Ota R, Yamashita F (2022) Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics. *J Control Release* 352:961–969. <https://doi.org/10.1016/j.jconrel.2022.11.014>
8. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: a 10 million image database for scene recognition. *IEEE Trans Pattern Anal Mach Intell* 40:1452–1464. <https://doi.org/10.1109/TPAMI.2017.2723009>
9. Jaquart P, Dann D, Weinhardt C (2021) Short-term bitcoin market prediction via machine learning. *J Finance Data Sci* 7:45–66. <https://doi.org/10.1016/j.jfds.2021.03.001>
10. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Copenhagen, pp 670–680

11. Alonso-Robisco A, Carbó JM (2022) Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio. *Int Rev Financial Anal* 84:102372. <https://doi.org/10.1016/j.irfa.2022.102372>
12. Lagasio V, Pampurini F, Pezzola A, Quaranta AG (2022) Assessing bank default determinants via machine learning. *Inform Sci* 618:87–97. <https://doi.org/10.1016/j.ins.2022.10.128>
13. Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. *J Bank Finance* 34:2767–2787. <https://doi.org/10.1016/j.jbankfin.2010.06.001>
14. Wang Y, Wang S, Lai KK (2005) A new fuzzy support vector machine to evaluate credit risk. *IEEE Trans Fuzzy Syst* 13:820–831. <https://doi.org/10.1109/TFUZZ.2005.859320>
15. Mainelli M, Yeandle M (2006) Best execution compliance: new techniques for managing compliance risk. *J Risk Finance* 7:301–312. <https://doi.org/10.1108/15265940610664979>
16. European Banking Federation (2020) AI in the banking industry: EBF position paper. <https://www.ebf.eu/innovation-cybersecurity/ai-in-the-banking-industry-ebf-position-paper/>
17. Borkin D, Nemethova A, Michalconok G, Maiorov K (2019) Impact of data normalization on classification model accuracy. *Res Papers Faculty Mater Sci Technol Slovak Univ Technol* 27:79–84. <https://doi.org/10.2478/rput-2019-0029>
18. Mohammed R, Rawashdeh J, Abdullah M (2020) Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: *Proceedings of the 2020 11th international conference on information and communication systems (ICICS)*, pp 243–248
19. Tharwat A (2020) Classification assessment methods. *Appl Comput Inform* 17:168–192. <https://doi.org/10.1016/j.aci.2018.08.003>



# Time Series Analysis of Public Opinion on Work from Home During and After COVID-19 Pandemic



Gabriela G. Mendoza-Leal, Jorge A. Mendez-Vargas,  
Francisco J. Cantú-Ortiz, and Héctor G. Ceballos-Cancino

**Abstract** After the COVID-19 pandemic years, and with government restrictions being lowered, many workers are still preferring to Work From Home. Companies, schools, and institutions across the world have switched to this remote mode because of its various advantages. This paper aims to analyze the impact of the pandemic on the % of increase relative to the number of people working from home before the pandemic. The data will be recollected from Aman Kumar's Public Domain dataset [1] found in Kaggle. Some questions that this paper intends to answer include: Is the increase in percentage related to the pandemic or just something that increases constantly per year? What countries had the greatest increase and what could be some factors that affected these numbers? How will this percentage continue to behave specifically in Mexico? We try to answer these questions by performing both a cluster analysis on the time series and a forecasting using ARIMA. The importance of studying this information lies in the fact that employers must be aware of the population's preference and adapt accordingly, to be able to stay competitive in this supply and demand market.

**Keywords** Work from home · COVID-19 · Time series analysis

---

G. G. Mendoza-Leal (✉) · J. A. Mendez-Vargas · F. J. Cantú-Ortiz · H. G. Ceballos-Cancino  
Instituto Tecnológico y de Estudios Superiores de Monterrey, Ave. Eugenio Garza Sada 2501,  
64849 Monterrey, NL, Mexico  
e-mail: [A00819743@exatec.tec.mx](mailto:A00819743@exatec.tec.mx)

J. A. Mendez-Vargas  
e-mail: [A01176369@exatec.tec.mx](mailto:A01176369@exatec.tec.mx)

F. J. Cantú-Ortiz  
e-mail: [fcantu@tec.mx](mailto:fcantu@tec.mx)

H. G. Ceballos-Cancino  
e-mail: [ceballos@tec.mx](mailto:ceballos@tec.mx)

## 1 Introduction

This paper focuses on data in which time is an important factor. And because of that reason, the proposed method of approach is a time series analysis (TSA). Time series analysis is a method focused in analyzing data points that were collected over a period of time. This way, it's possible to visualize how data adjusts over the course of the data points, as well as predictions.

This method usually needs a large amount of data to better guarantee reliability in the analysis and predictions. When the models are fed a large amount of data, it is possible to better visualize it and realize if any patterns or trends are not outliers and can account for seasonal variance. It can also be used to create prediction models that are able to forecast data based on the input historical data. In this paper we will make use of the autoregressive integrated moving average (ARIMA) model, which is fitted to the time series to be able to better understand the data and predict future data points.

The ARIMA analysis has evolved through time, and it is now possible to visualize the data far beyond line graphs [2]. This project aims to study data collected by Aman Kumar, in which the “Increase in Residential Stay” was recorded by date in several countries starting from February 17, 2020, up to January 30, 2022.

## 2 Background

This project and all the experimentation will be realized on Python. Thankfully, Python has many useful libraries to experiment with TSA. TSA has several components that will prove relevant in the upcoming sections. Listed below are the definitions for these elements [3].

1. **Trend:** A trend lets us see how the data varies over time in a simplified manner. It could remain stable, increase, or decrease.
2. **Seasonality:** The data will have variations, but seasonality lets us visualize if these variations occur regularly over time. They could be due to seasons, specific monthly or yearly events, festivals, etc.
3. **Stationary:** This aspect relates to the way the properties of the time series change over time. If these properties remain the same on all the series, it is stationary. And the time series must be stationary for us to be able to perform an analysis on it. If the series is stationary, it will have a constant variance, covariance, and mean.

### 3 Related Work

Other works on time series analysis such as the ones done by Verma [4], Brownlee [5], and Fernandez [6] show similar methodologies in the use of autoregression models, moving averages, testing stationarity, and using the fit and forecast Python functions. Nevertheless, most TSA works show how to find the best ARIMA parameters, while only a few of them take the next step and show how to predict future values, which is one of the main reasons for doing this type of analysis.

### 4 Methodology and Data

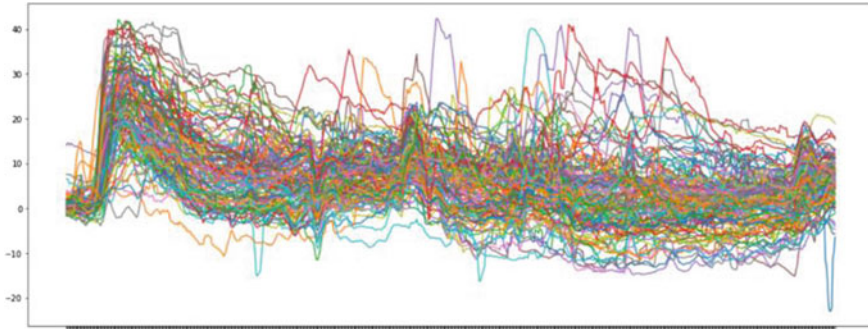
As stated in the previous sections, the data was downloaded from Kaggle. This section begins with the exploration of these data values and understanding what can be obtained from it. This will be followed by the experimentation step, in which a cluster analysis will be performed, and then an ARIMA analysis to predict future values. The clustering will involve all countries in the data set, but the forecasting will be focused only on Mexico.

#### 4.1 Data Understanding and Preparation

One of the first things to note were the three main data columns: Entity, Day, and Increase in Residential Stay. The first column represents the country from which the data comes from, the second column is what makes this a time series because it contains the days in which the data was obtained, and the third column contains the actual data, which is the increase or decrease in residential stay. There are a total of 91,933 rows.

When analyzing the data, it was also noticed that almost all countries had a total of 714 rows, except for a few that had less. It was decided to drop these data values to obtain a more comparable data set. The countries that were deleted were: Afghanistan, Antigua and Barbuda, Cape Verde, Georgia, Papua New Guinea and Serbia. Also, to be able to do a clustering analysis, it was necessary to modify the data frame to a way in which each column represented a different country and had its corresponding time series.

Continuing with the data exploration, a plot comparing all time series was created. It is shown in Fig. 1. It can be noted from the plot of all time series that most countries show a similar tendency in the Increase in Residential Stay variable with a few exceptions which would be the lines that deviate from the denser parts of the graph. This can be further compared with government restrictions regarding WFH and a pattern could be found.



**Fig. 1** Plot of all time series

Later, the top 5 time series with the greatest single increase were obtained. This means that they reached the highest single point in the data. These time series corresponded to Rwanda, Zimbabwe, Botswana, Panama, and Singapore. These are the five countries with the highest increase. Most of their maximum peaks are at the beginning, but there are a few instances in which some of them reach a very close maximum peak again, especially Rwanda and Panama. Note that future work can be done to define whether the reason of these peaks was government mandatory lock downs or if personal preference was also a factor. With this, the exploration of the data was concluded, and the experimentation phase began. On the next section, ARIMA and clustering was performed.

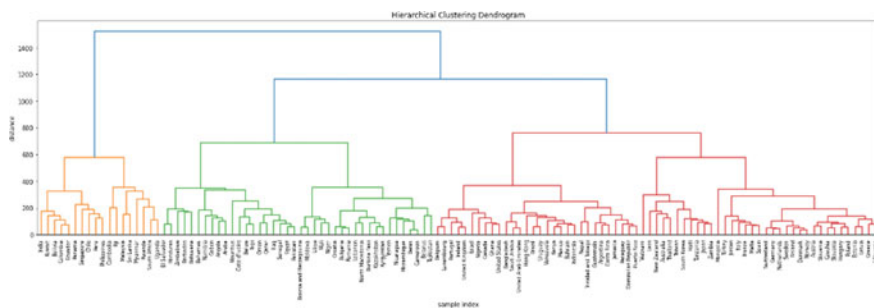
## 5 Experimentation

Moving onto the experimentation phase of the project, this section will be divided into two different procedures. First, a cluster analysis on all data including all countries, and then focusing specifically on Mexico and trying to forecast using the ARIMA model.

### 5.1 Clustering

Moving forward with the clustering analysis, a hierarchical clustering was constructed with several variations to the methods. First, using the Ward method. It is an alternative to single-link clustering. This method is computationally intensive but has significantly fewer computations than other methods.

The Ward method minimizes the sum of squared differences within all clusters. It is similar to the  $k$ -means objective function but tackled with an agglomerative hierarchical approach. Note that agglomerative cluster has a behavior that leads to



**Fig. 2** Clusters using Ward method

uneven cluster sizes, and in this regard, single linkage is the worst strategy, with the Ward method giving the most regular sizes. One downside is that affinity (distance used in clustering) cannot be varied with Ward.

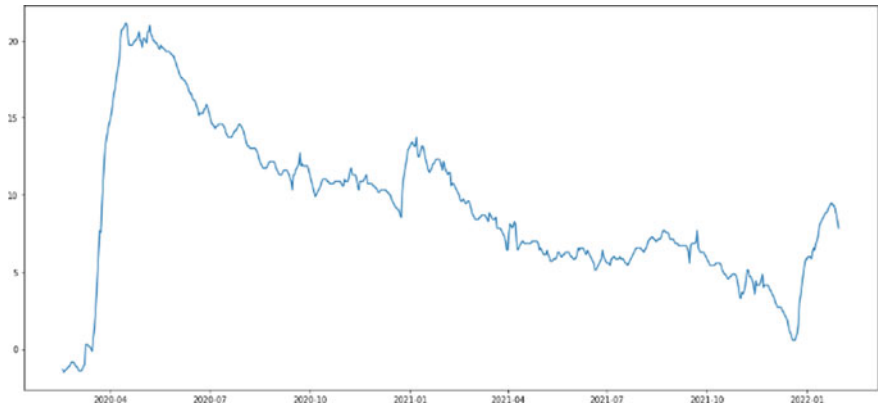
The visualizations displayed how both the Spearman and Pearson correlation clusters did not show clear separations, especially compared to the clusters obtained by using the Ward method which are very identifiable as seen in Fig. 2. In general, the Spearman and Pearson correlation clusters were much more different compared to the cluster obtained from the Ward method. This could be due to the different groupings caused by the different distance matrices since the two methods calculate it as a 1-correlation coefficient. Also, the dendrogram drawing method is the Single method (the distance between clusters in the tree is calculated using the shortest distance between them) instead of the Linkage method used by Ward.

## 5.2 Forecasting

This project will be centered on the replication of Mexico's data using ARIMA method. The autoregressive integrated moving average (ARIMA) model works specifically for time series and is fitted to the existing data. It can be used to either better understand the data or to try and predict new values. Up until this point it has only been used to see how well it keeps up with the data, but later it will also be used to predict new values.

There are many time series, one for each country, but Mexico was the main focus on this project, so the procedures will only be performed on Mexico's time series. The relevant data was extracted from the original dataset. Then, a function was defined to obtain the rolling mean and rolling standard deviation with a window of 30 data.

As mentioned beforehand, a **stationary** time series is necessary to be able to analyze and predict it. To check if it is stationary or not, a Dickey-Fuller Test was conducted. It is a test of statistical significance, and as such it yields results based on a null and alternative hypothesis. That means the results come in the form of a  $p$ -value. The  $p$ -value must be  $< 0.05$  to reject the null hypothesis, and that would mean the



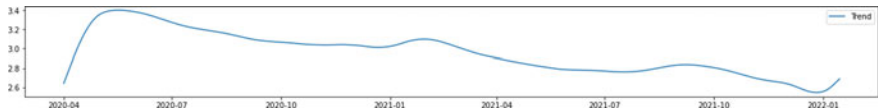
**Fig. 3** Visualization of the time series for Mexico

time series is stationary. A Dickey-Fuller Test was conducted, and the resulting test statistic was  $-3.365811$  and the  $p$ -value was  $0.012181$ . This result supposedly states that the time series is stationary, but after graphing the time series (Fig. 3), the data did not show a convincing stationarity and therefore, we wished to improve it.

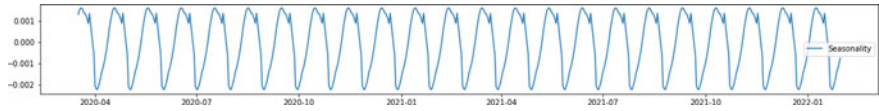
Trying to improve the stationarity, a logarithm version of the data was calculated with a positive offset of 10. Another rolling mean was obtained from this log-converted data and it was subtracted from the original data to obtain its stationarity. Another Dickey-Fuller Test was done with a resulting test statistic of  $-5.600812$  and a  $p$ -value of  $0.000001$ . Finally, after this, stationarity was obtained in the data.

Then a seasonal decomposition function was used to obtain the trend and seasonality. Note that the period used was 30 which corresponds to the average number of days in a month. The following Figs. 4, 5, show the trend and seasonality.

After obtaining these results, the ARIMA method was used. The first step was to run autocorrelation functions to analyze the data and obtain the best AR and MA parameter combinations to find the best predictions. It is possible to visualize an



**Fig. 4** Visualization of the trend in the Mexico time series

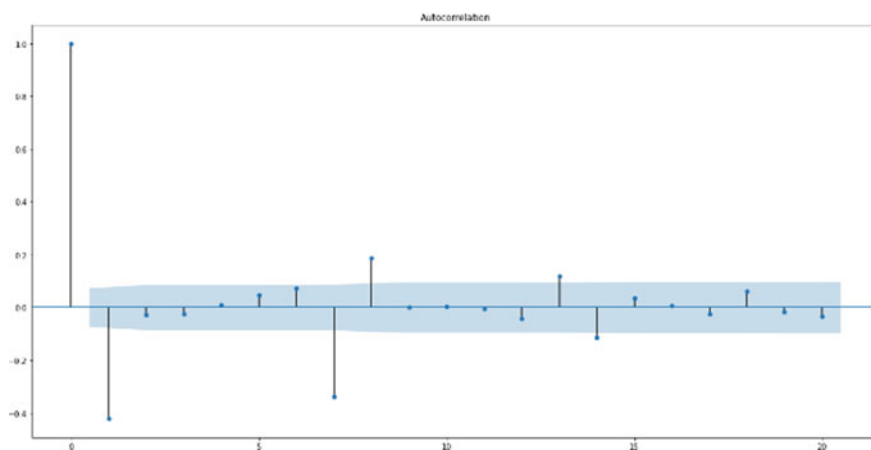


**Fig. 5** Visualization of the seasonality in the Mexico time series

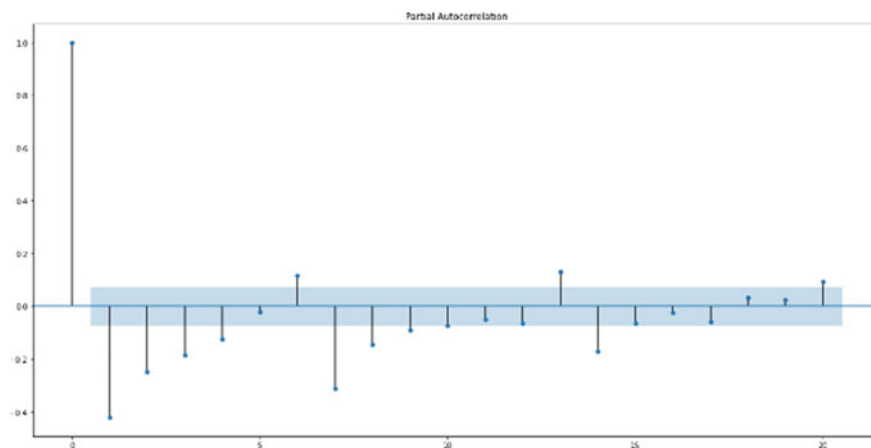
example of the resulting correlograms on Figs. 6, 7. The hyper parameter values chosen were therefore 1, 2, 1.

Figure 8 shows how the model adjusts with the chosen parameters. It can be noted from this graph that the obtained parameters were successful in describing the model.

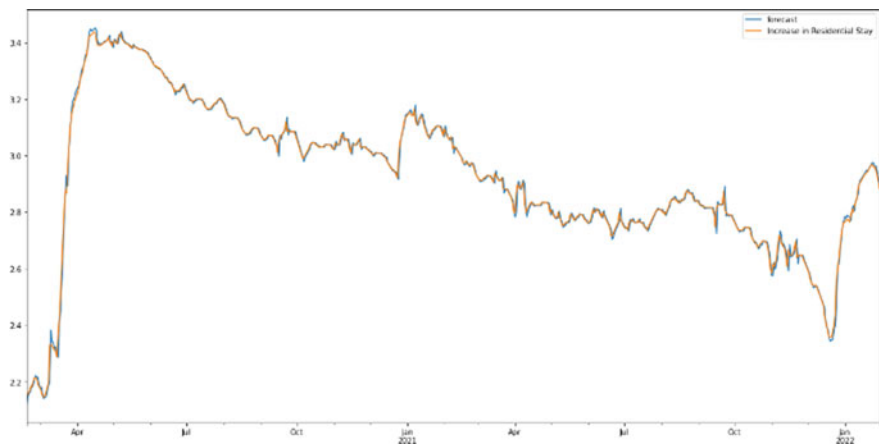
The next step was to predict values that hadn't been given to the model in training. For this part, it was necessary to also select a small number of data to be retained from the model to be able to compare the predicted results from the known results. In this case, it was observed that 35 days was a good value, since retaining more data would prevent the model from having necessary information to predict with better accuracy in the short term.



**Fig. 6** Visualization of the autocorrelation function with log of data



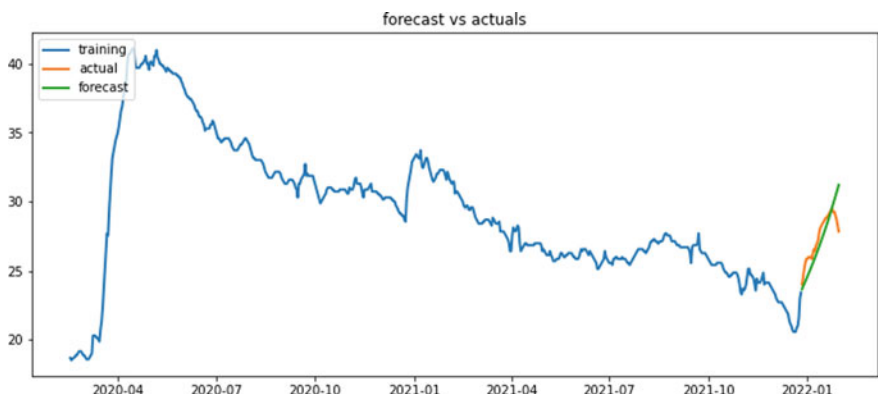
**Fig. 7** Visualization of the partial autocorrelation function with log of data



**Fig. 8** Visualization of the ARIMA procedure with parameters 1, 2, 1

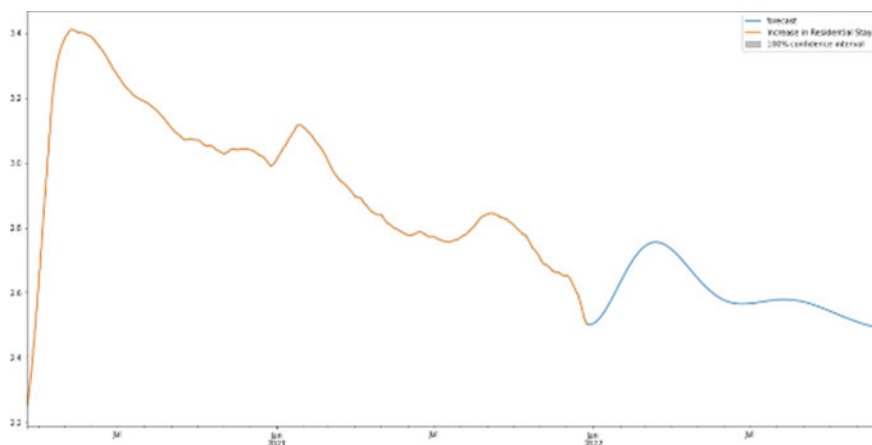
However, the obtained forecast graph did not have the expected behavior. A straight line (marked in green) was the prediction result as shown in Fig. 9. This behavior remained the same even when predicting more and less data values as well as when trying different ARIMA hyper parameters. Therefore, it was deduced that the data being used was the cause of this pattern.

To solve this, it was decided that instead of selecting as input the data as it originally was, the rolling mean of the data would be used. With this, the data was explored in a different way using other parameters and at the same time this was able to improve the ARIMA predictions. Therefore, the procedure had to be repeated to obtain the AR and MA parameters. The best parameters according to these functions were 1, 3, 1.



**Fig. 9** Predicted values versus original data with parameters 1, 2, 1





**Fig. 10** Predicted values and original data with parameters 3, 1, 3

Nevertheless, the ARIMA model did not accept a value higher than 2 for its “ $d$ ” parameter (number of nonseasonal differences needed for stationarity). Therefore, values 1, 2, 1 were tested. These parameters, however, also presented a linear trend in the long term. It was therefore decided to vary the other parameters (AR and MA) to see which values showed a better prediction. One example is shown on Fig. 10.

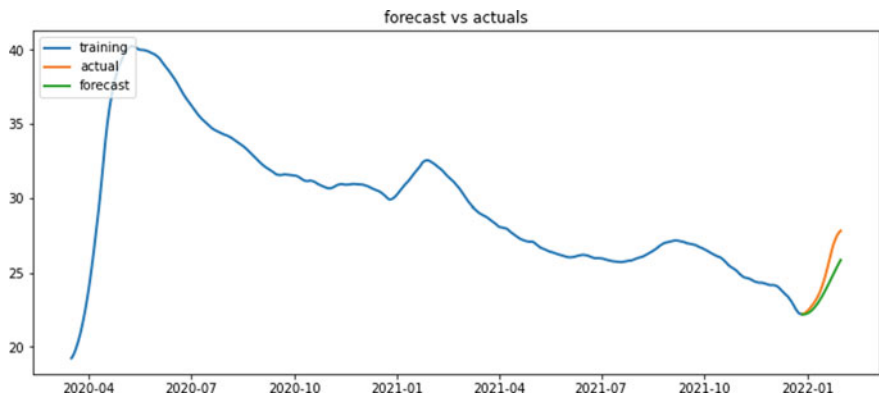
## 6 Results

Finally, the following two Figs. 11, 12 were obtained using parameters 4, 1, 4. These parameters also presented the lowest AIC and BIC values ( $-7584.370$  and  $-7539.616$ , respectively), which is why they were selected as the best result.

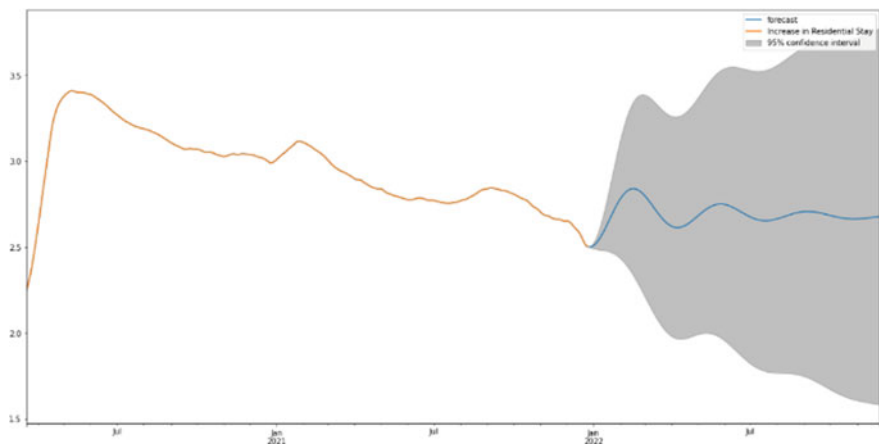
Table 1 shows some of the parameters that were tested and their respective AIC and BIC values. Note that the parameters selected for having the lowest values were marked in bold.

## 7 Discussion

After observing the results, it is possible to assume from Fig. 12 that the tendency over time of the predicted values will stay constant. This means there will be (in average) a constant increase of people who will desire to work from home. Nevertheless, these results are not definitive, considering the tendency shown on Fig. 10, which, even if those parameters did not have the lowest AIC and BIC values, the graph shows a more expected tendency to decline over time even with a few oscillations.



**Fig. 11** Visualization of the predicted values and original time series comparing 35 days using final parameters 4, 1, 4



**Fig. 12** Visualization of the predicted values and original time series comparing 35 + 300 days using final parameters 4, 1, 4

**Table 1** AIC and BIC values obtained for different ARIMA parameter tests

Data used	ARIMA parameters ( $p, d, q$ )	AIC	BIC
Log of OD	(1, 2, 1)	−3628.860	−3610.789
Log of OD	(4, 1, 4)	−3666.879	−3621.688
RM of log of OD	(1, 2, 1)	−7528.634	−7510.738
RM of log of OD	(3, 1, 3)	−7515.075	−7479.272
RM of log of OD	(4, 2, 4)	−7563.009	−7518.270
<b>RM of log of OD</b>	<b>(4, 1, 4)</b>	<b>−7584.370</b>	<b>−7539.616</b>

OD stands for original data. RM stands for rolling mean

One observation that was noted on a few of the graphs when trying different ARIMA parameters (and excluding the graphs which outputted a straight line), was that a small “bump” was seen around the month of June meaning an increase in the number of people who desired to work from home. This could be partly attributed to people who, for example, have children and wish to stay home for their summer school vacations.

## 8 Conclusions

This project was useful to realize how time series analysis is done, why it’s useful, and most importantly, how each parameter affects the results. The most notable ones, in this case, included the data given as input, as well as the data excluded for training and finally the  $p$ ,  $d$ , and  $q$  parameters. It can be said that when the ARIMA shows an incongruous prediction, one of the first things to test is the data input. Exploring the data to extract representative parameters without losing information (such as using the rolling mean as in this case) might prove to be useful to predict more suitable results.

On the other hand, correlograms do tend to show the best parameters to be used in ARIMA, but it is important to still test other combinations in case a better solution is obtained. In this case, the most significant parameter change was the  $p$  value, which was the number of autoregressive terms included in the model and which also changed the prediction from a straight line to showing change over time.

A final recommendation to have the best result predictions is to keep feeding the model with updated data. Figure 12 shows a gray area which represents a 95% confidence interval, but considering it is exponential, it will grow more over time unless new values are given to the model.

As mentioned beforehand, the contribution includes the description on how to predict values, as opposed to what most ARIMA publications cover. This work could be useful for recruiters and businesses to observe and predict future trends of employees’ preferences and be able to adapt the roles with these characteristics in mind.

## 9 Future Work

An interesting question that could be covered in the future work is to observe whether the increase and decrease in percentage observed was related to the pandemic or if it was something that happened constantly each year. This could not be done with the data currently on hand, and one or two more years of data recording is recommended to be able to conclude on this question.

With the information currently on hand, it would be interesting to see the predictions of other countries and observe similarities and differences. For example,




Mexico's predictions could be compared to those of the United States, or with another Latin America country.

## References

1. Kumar A (2019) Increase in residential stay across globe since outbreak of COVID-19 pandemic. Kaggle, San Francisco. <https://www.kaggle.com/datasets/aestheteaman01/people-staying-in-home-during-covid19>
2. Tableau (2020) Time series analysis: definition, types, techniques, and when it's used. Tableau, Mountain View. <https://www.tableau.com/learn/articles/time-series-analysis>
3. Simpli Learn (2021) Understanding time series analysis in Python. <https://www.simplilearn.com/tutorials/python-tutorial/time-series-analysis-in-python>
4. Verma Y (2009) Quick way to find p, d and q values for ARIMA. AnalyticsIn Diamag, Bengaluru. <https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/>
5. Brownlee J (2017) Multistep time series forecasting with LSTMs in Python. Machine Learning Mastery, San Juan. <https://machinelearningmastery.com/multi-step-time-series-forecasting-long-short-term-memory-networks-python/>
6. Fernandez J (1983) Creating an ARIMA model for time series forecasting. Towards Data Science, Toronto. <https://towardsdatascience.com/creating-an-arima-model-for-time-series-forecasting-ff3b619b848d>

# Determination of Air Quality with Unmanned Vehicles in Cement Plants



Diego Verdugo-Ormaza , Jean P. Mata-Quevedo ,  
Ricardo Romero Gonzalez , and Luis Serpa-Andrade 

**Abstract** The World Health Organization considers environmental pollution to be one of its priorities for human health; their studies indicate that there is a risk at the concentrations seen in many developed countries today. To do this, the regulatory bodies in charge of environmental surveillance, protection and prevention use air quality measurement systems in the form of fixed stations located in areas of interest, generating high investment costs due to the need to replicate instrumentation systems at strategic points for data collection. Taking into account the above, this project presents an alternative solution to the problem posed by instrumenting an unmanned aerial vehicle as a displacement device and implementing a measurement system for environmental variables anchored to it. As a result, the air quality measurement system was obtained using low-cost sensors mounted on a drone. From the data obtained from the real-time system, it was possible to establish that the parameters evaluated are within the norm established for our country. It was also determined that in the area near the cement factory, the average particulate matter is  $8.01 \mu\text{g}/\text{m}^3$  and carbon dioxide of 373.26 ppm, in relation to the center of Guapán with values of  $6.34 \mu\text{g}/\text{m}^3$  and 158.45 ppm, respectively, therefore, the levels of contamination near the company are higher than those measured in the parish center.

**Keywords** System · Monitoring · Environmental · Pollutants · Air quality

---

D. Verdugo-Ormaza · J. P. Mata-Quevedo · R. R. Gonzalez  
Universidad Católica de Cuenca, 030101 Azogues, Ecuador

L. Serpa-Andrade (✉)  
Universidad Politécnica Salesiana GIHEA, 010102 Cuenca, Ecuador  
e-mail: [deverdugoo@ucacue.edu.ec](mailto:deverdugoo@ucacue.edu.ec)

# 1 Introduction

Studies carried out by the World Health Organization (WHO) determine that nine out of ten people in the world breathe polluted air, this caused the premature death of about 4.2 million people in 2016 [1]. Air pollution represents a significant environmental risk to health. According to the WHO, particulate matter (PM) is considered one of the most dangerous air pollutants, industries and vehicular traffic cause great changes in relation to the environment [2]. In Ecuador, Official Register No. 464 establishes that the arithmetic average of the concentration of particulate matter PM10 (particles of  $<10\text{ }\mu\text{m}$  aerodynamic diameter) of all the samples in a year, should not exceed  $50\text{ }\mu\text{g}/\text{m}^3$ ; the continuous arithmetic average of 24 h, should not exceed  $100\text{ }\mu\text{g}/\text{m}^3$ . Likewise, it establishes that the arithmetic average of the concentration of particulate matter PM2.5 (particles with an aerodynamic diameter  $<2.5\text{ }\mu\text{m}$ ) of all the samples in a year, should not exceed  $15\text{ }\mu\text{g}/\text{m}^3$ ; the continuous arithmetic average of 24 h, shall not exceed  $50\text{ }\mu\text{g}/\text{m}^3$  [3].

In the industrial district of Garmsar located southeast of Tehran in Iran, samples were taken for four months to measure air quality, obtaining as a result that on some sampling days PM10 is higher than the maximum allowed concentration of the ambient air standard. Therefore, continuous monitoring of emission sources is essential, as well as the generation of control policies and the creation of local and industrial guidelines in this area [4]. In this same context, it was observed that the average annual values of PM10 and PM2.5 in most of the sites measured in the Latin American and Caribbean (LAC) region were significantly higher than the WHO air quality guidelines, with  $<5\%$  of the cities (among 117 studied) that comply with the guidelines given [5].

Cement companies have a significant impact on the environment due to their production processes, which, due to the complexity of their activities, generate large amounts of pollutants that can be very toxic both for the environment and for the community where they are located. Factories are located, if they are not controlled [6]. In a case study carried out in Jordan in the southeast of Amman to evaluate the air quality inside a cement factory, measurements of total suspended particles (TSP), PM10, PM2.5, lead, carbon monoxide carbon (CO), nitrogen dioxide ( $\text{NO}_2$ ), sulfur dioxide and hydrogen sulfide, as a result it was obtained that the gaseous pollutants were below the detection limits; however, the particulate matter was relatively high [7].

According to [8] the emissions of a cement industry, they exist from particulate materials, sulfur dioxide and nitrogen oxide, which can cover the vegetation, the soil and affect the entire biotic environment. It is for this reason that the cement industries are one of the largest contributors to air pollution. The authors [9], based on the observation, specify that at noon the difference in the level of particles is greater in the districts of Minasatene and Bungoro (Indonesia), because the intensity of mining activity in both locations is also greater. This activity contributes to the formation of PM2.5. Furthermore, both locations are close to the limestone mining area.

One of the environmental problems is the massive erosion of the quarry due to the continuous extraction of limestone and other materials that severely affects the soil. Opencast mining is a disturbance caused by man that affects all components of an ecosystem (vegetation, animals, soil, etc.) [10]. Another problem is the loss of surface soil, contamination of surface waters, dust emissions into the atmosphere and noise emissions [11].

As [10] described, PM is the main pollutant released into the atmosphere by the cement industry. In the municipality of Sogamoso, daily emissions are around 180  $\mu\text{g}$ , which substantially affects the health of the population due to the existence of particles smaller than PM10, which can affect vision, cause eye irritation and one of the main problems are that when these small particles are inhaled they can reach the lungs and cause serious respiratory problems. In accordance with the foregoing [12], it states that the manufacture of cement is a process that results in the emission of significant amounts of PM into the air. The investigation was carried out in Coimbatore, in the state of Tamil Nadu, in southern India, as a result it was obtained that most of the monitored places were found to be well above the limits specified in the National Air Quality Standards. Environment of said country.

In [13], states that the air quality index (ICA) is commonly used to indicate to the public the level of air quality through a number, color schemes, graphics, labels such as: good, moderate or poor; Generally based on measuring the main pollutants that are:  $\text{O}_3$ ,  $\text{NO}_x$ ,  $\text{SO}_2$ , CO, PM and oxidants.

For these reasons, an alternative to measure air quality using low-cost sensors placed on an unmanned aerial vehicle (UVA) is presented. Sensors are devices that convert a physical phenomenon to an electrical signal. Low-cost sensors (SBC), due to their price and size advantages, make them the most widely used sensors currently in various industrial sectors and recently for air quality monitoring [14]. According to the authors [15, 16], the advantages of portable sensors are their efficient implementation over time, their easy operation and their portability, which makes them ideal instruments for measuring PM concentrations. These instruments have been used to measure particle emissions outdoors, in urban or rural regions [17, 18]. On the other hand, field measurements made with portable instruments lack the 24-h average information offered by fixed stations. In any case, there are places and daily procedures where it is difficult to use conventional fixed measuring stations; therefore, alternative wearable measurement devices based on wearable sensors can be used to monitor daily events. Low-cost wireless sensors have been used for monitoring purposes and when the results obtained are validated with the stationary monitoring instruments, the correlation coefficient ( $R^2$ ) is found to be acceptable [19].

Under this context, a PM measurement system was built using a UVA as a displacement device and an environmental variable measurement system was implemented, placed on it. For this purpose, a two-phase methodology was carried out: the design of a PM measurement system focused on the characterization of air quality and the manufacture of a UVA with its own specifications for this application, that is, suitable for the assembly of the measurement system [20].

In the study carried out by [21], they state that an intelligent multisensor system was used to monitor air quality in the city of Torreón, the system uses a UVA and

the communication is carried out through LoRa, as an alternative for remote and in situ atmospheric measurements. Similarly, [22] they measured the air quality in the Toucheng exchanger and the results obtained provide a reference model to assess environmental impacts in areas prone to environmental contamination.

In the city of Azogues there is cement activity in the Guapán parish; Therefore, the Azogues Municipal Decentralized Autonomous Government (GAD), through its Environmental Management Unit, has carried out studies of sedimentable material but not of suspended materials such as PM<sub>2.5</sub> and PM<sub>10</sub> [23], it is so; that the deterioration of air quality in the city of Azogues is a problem that increases over time and has effects on people's health, causing cardiovascular and respiratory diseases [24].

The importance of the investigation is to have data on air quality in the area near the cement factory, in order to provide information that protects public health and the ecosystem. Ambient air quality in all cities should meet the national ambient air quality standards and the ambient air quality reference values established by the WHO [25].

Therefore, the Ministry of the Environment and the Decentralized Autonomous Governments have the obligation to monitor PM concentrations, compare them with legal regulations and in case of non-compliance, issue the respective environmental policies for the care of people's health and the environment. It is important that institutions such as the Ministry of Public Health have information on the concentration and characterization of PM in the local air intake, since through them potential respiratory diseases can be predicted.

Based on the previous studies carried out regarding the topics covered, it can be inferred that the problem consists of establishing if the air quality in areas close to the cement manufacturing industries affects the quality of the air that is breathed, increasing the health risks. In the nearby population, making it possible to contract respiratory diseases.

Obtaining air quality data using low-cost sensors will serve as a study basis to establish the necessary tasks to mitigate air pollution caused by the cement industry. It is expected that based on the results of the proposed investigation, it will be possible to determine the factors that affect the quality index of the air that is breathed in the areas surrounding the cement factories and based on their analysis, it can be established from legal regulations actions that mitigate the impact on the community.

## **2 Material and Methods**

### **2.1 Material**

The unmanned aerial system using low-cost sensors to measure air quality, was designed to measure the concentrations of different pollution variables in urban areas with acceptable sensitivity and precision. The measurement points determined



**Table 1** Factors for the selection of low-cost sensors

Physical factors	Environmental factors	Economic factors
Dimensions (D)	Temperature	Price (C)
Weight (P)	Humidity	Availability
Operating voltage (VO)	Atmospheric pressure	
Maximum current consumption (CM)	Carbon dioxide (CO <sub>2</sub> )	
Measuring ranges (RM)	Particulate matter	

both in the area near the UCEM SA factory and in the center of the Guapán parish, whose distance in a straight line is 1.5 km.

Firstly, the sensors to be used to measure pollutants such as CO<sub>2</sub> and PM were defined; in addition to atmospheric pressure, temperature and humidity. For this reason, a search was carried out in the market for sensors capable of measuring these contaminants, taking those specified in Table 1 as selection criteria.

The sensors are: humidity and temperature sensors DHT11, CO<sub>2</sub> measurement, the MQ-135 sensor was chosen and for the choice of the PM sensor, we proceeded with the same criteria as for choice of the sensor, without taking into account the precision because they all have 1 µg/m<sup>3</sup>. After the analysis, the DSM501A sensor was selected for its attributes and price. To measure atmospheric pressure, the GY-BMP280 sensor was chosen, due to its simple programming, stability, economic price and availability.

For the embedded system selected for the monitoring station, this being the Arduino Fio for the number of peripherals, for the size, power consumption, connectivity and costs that are adapted according to the requirements of the project, For the communication module, Xbee Pro S2C was selected due to its low cost and accessibility, which meets the characteristics required for the project's data communication.

The monitoring station consisted of 4 sensors, whose total weight including the protection box is 202.67 g. And the DJI Mavic 2 Pro drone used to carry out the measurements whose weight is 907 g, its dimensions of 91 × 214 × 84 mm and its flight autonomy of 25 min.

## 2.2 Methods

For the development of the project, the research methodology based on a quantitative application design was used.

This investigation was carried out in four stages: (i) determination of sensors, (ii) design, construction and calibration of the system, (iii) implementation of the air quality measurement system and (iv) measurements were made in situ.

In the first stage, an analysis and identification of the variables was carried out through a study of the operation criteria, where physical factors such as dimensions, weight, operating voltages and maximum current consumption will be valued; as well as the environmental factors to be measured as air quality variables and measurement ranges; in the same way, factors that allow a communication or storage of the data; also, the possible feeding system and the autonomy that will be achieved were evaluated; finally, economic factors such as price and availability in the market were evaluated. With this information, the sensors, communication, storage and power systems were determined by means of a selection matrix considering the importance of each of the electronic devices so that together they can meet the energy and weight requirements, you can see the detail in Table 1.

In a second stage, we proceeded to design, build and calibrate the prototype of the portable air quality measurement system, using low-cost sensors that meet the weight and energy requirements of the drone; which measured temperature, humidity, atmospheric pressure, CO<sub>2</sub> and PM. The calibration of the sensors was carried out by comparing the measurements with a standard instrument and in this way to be able to determine their correct operation.

In the third stage, the system was installed on the drone and several tests were carried out to see its correct operation, robustness, autonomy and data transmission. If any problem is found, it will return to the previous step to improve the prototype.

In the fourth stage, we proceeded to collect air quality data in the areas close to the UCEM SA cement company, located in the Guapán parish, during a period of 30 days from 10:00 a.m. to 12:00 p.m. and from 3:00 p.m. 4:30 p.m. with 30 s sampling intervals. These zones were delimited by means of a pollutant dispersion calculation, analyzing meteorological and environmental variables. Immediately, the data obtained was validated and a statistical analysis was carried out to determine if the air breathed in the areas near UCEM SA maintains concentrations of pollutants above the values allowed by Ecuadorian legislation. Finally, with the data collected, information on air quality can be obtained and this information is provided to both the Azogues Municipal Autonomous Decentralized Government (GAD) and the environmental control agencies for their respective analysis.

### 3 Results

The multisensor system was tested through tests carried out at different sites in Azogues. The monitoring time was determined according to the battery time for a safe flight of approximately 25 min. For data acquisition, the system configuration was performed according to the proposed methodology. The multisensor system was attached to the drone and then flew vertically for a range of 15–25 min.

According to the applied methodology and once the data of the measurements carried out in the two locations specified above were obtained, the information was analyzed and interpreted based on the taking of the samples carried out according

to the scheduled sampling, the results can be reviewed most relevant in the graphs shown below.

With the data collected from the embedded system at the sampling points determined was possible to establish the behavior of the particulate material, where the maximum concentration point occurs in the company UCEM SA in comparison with the center from the Guapán parish, the average values are 8.01 and 6.34  $\mu\text{g}/\text{m}^3$ , respectively, in each one of the measurement points; In addition, we can determine that the data is more concentrated in the center of Guapán. In short, it can be indicated that both in the critical zone and in the center of the population the ranges are within those established as acceptable in current regulations.

The concentration of carbon dioxide is higher in the closest point of the cement factory with respect to the center of Guapán, we can also state that the concentration is more dispersed between 50 and 75% of the data in the two measured cases. The average of this gas is 373.26 and 158.45 ppm, being within the levels allowed for the health of people in open spaces.

The temperature is higher at the measurement point close to the company due to the gases emanating from the chimneys due to the production of cement with respect to the populated center, the average is 20.82 and 20.08 °C, respectively. In the same way, it can be determined that the temperature has more variability in the area close to the company in relation to the one in the center of the parish. Finally, we can state that the temperature in the central area of the parish has a symmetrical distribution.

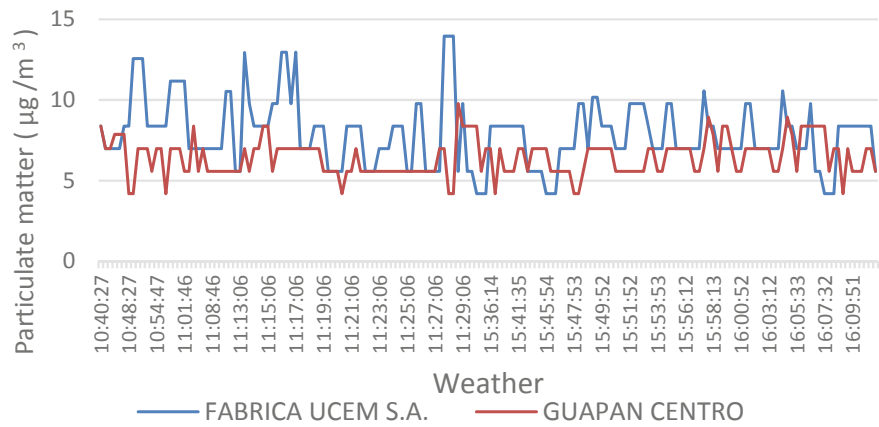
The atmospheric pressure behaves similarly at the two valuation points, with the average being 753,619 hPa at the UCEM SA factory and 753,613 hPa at the other measurement point. Similarly, it can be defined that the dispersion and variability of the data are similar for each of the areas.

Figure 1 shows the variation of the concentration of particulate matter in the two evaluated points and it can be concluded that in the point close to the company UCEM SA the amount of particulate material is greater than in the center of the parish, due to the distance existing between them.

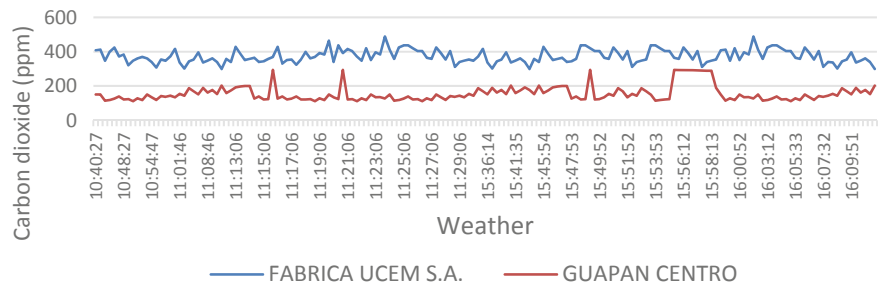
As can be determined, Fig. 2 shows the variation in the behavior of carbon dioxide in the sites determined for data collection and it can be specified that the values are higher in the area close to the industry than in the center of the parish.

Figure 3 shows the behaviour of the temperature in the evaluated points and it can be concluded that in the point close to the company UCEM SA the measured values are higher in relation to the center of Guapán, this is due to the gases that are emitted in the cement manufacturing process.

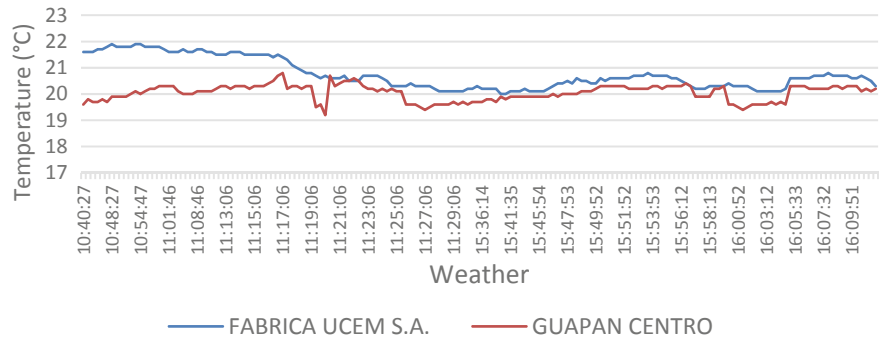
Figure 4 shows the behaviour of the pressure at the chosen points and it can be determined that the variations of this measurement parameter are very similar, since the distance between the two measurement points is not very large.



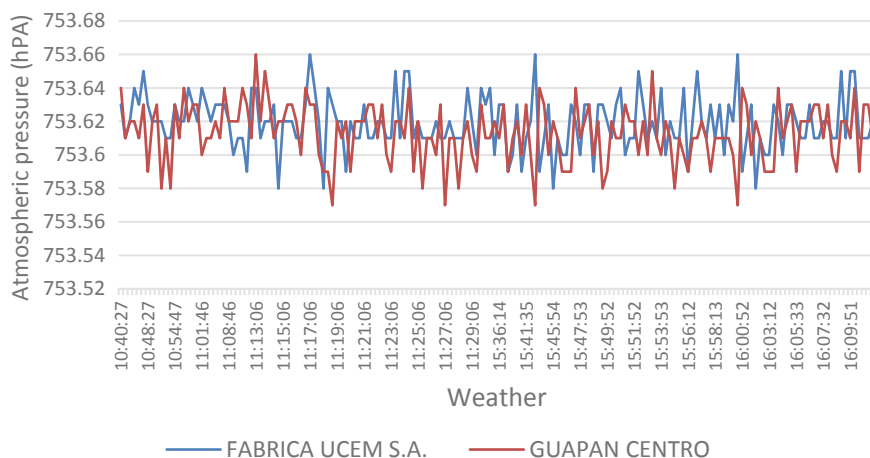
**Fig. 1** Daily average of the behavior of particulate matter



**Fig. 2** Daily average carbon dioxide behavior



**Fig. 3** Daily average of temperature behavior



**Fig. 4** Daily average of atmospheric pressure behavior

## 4 Discussion

Based on the relationship between the results obtained and the information collected, an alternative air quality monitoring system using low-cost sensors is proposed. The low-cost sensors, they are an adequate solution to implement data acquisition systems and, combined with conventional equipment, they allow monitoring air quality more effectively. According to the results obtained in our project, the measurement system based on low-cost sensors proved to be a very useful alternative to help contrast and complement the data obtained or missing from other methods of measuring air quality pollutants, such as monitoring the pollution emitted into the atmosphere by factories and vehicles to identify and control the sources of emission of polluting gases and solid particles. In addition, the proposed system, mounted on a drone, could be used to determine regions of contamination and delimit affected areas in case of disasters or environmental contingencies.

Studies show the symptoms of older adults residing in the surrounding areas of the Guapán cement factory, obtaining as results that 40% present flu symptoms at different times of the day, mostly due to its cold weather rather than pollution. These data coincide with those determined in our study, because the measured variables are within the acceptable parameters for human health.

Data from this study helped government agencies, ministries of health and policy-makers to take proactive steps. Pollution data could be considered by local air quality standards to prevent risks to the population.

## 5 Conclusions

The main conclusion of this work is that it was possible to build a low-cost prototype station to measure several air quality parameters in selected areas, with a particular interest in industrial areas and their surroundings, as an alternative to fixed air quality monitoring stations air quality. It is important to highlight that a successful integration of the sensors has been achieved within the same platform.

The design of the circuit to measure air quality complies with the characteristics of the air vehicle, either in size and weight. In addition, it has electronic devices such as an Arduino, a humidity and temperature sensor, CO<sub>2</sub>, atmospheric pressure and a particulate material sensor, the latter being installed on one side because it is a device that must be free to obtain accurate data without disturbances from other devices. At the same time, it was possible to obtain a system using low-cost sensors that meet the requirements established within the project.

Compared to ground stations or other aerial techniques such as manned aircraft and satellites, UVAs are more versatile and operationally outstanding. Some features of these, such as compact size, weights and power requirements, combined with reduced platform and instrumentation costs, make them extremely suitable for air quality monitoring.

Finally, it can be concluded that the amount of particulate material is within the limits allowed by the laws in force within our country and that a database of these polluting materials can be counted on for future comparisons, due to the fact that the Municipal GAD of Azogues does not have such measurements in real time. In the same way, it is determined that the levels of pollutants are higher in the area near the cement factory in relation to the parish center.

## References

1. Organización Mundial de la Salud (2018) Nueve de cada diez personas de todo el mundo respiran aire contaminado Sin embargo, cada vez hay más países que toman medidas. In: Departamento de Salud Pública, Medio Ambiente y Determinantes Sociales de la Salud, pp 1–5
2. World Health Organization W (2016) Calidad del aire ambiente (exterior) y salud. In: Nota descriptiva, pp 1–8
3. MAE (2011) Norma Ecuatoriana de Calidad del Aire. In: Texto Unificado de Legislación Secundaria Del Ministerio Del Ambiente, pp 402–415
4. Maleki R, Azhdari SS (2022) Measuring the ambient air pollutants in Garmsar industrial district. *J Air Pollut Health* 7:51–60
5. Gouveia N, Kephart JL, Dronova I, McClure L, Granados JT, Betancourt RM, O’Ryan AC, Texcalac-Sangrador JL, Martinez-Folgar K, Rodriguez D, Diez-Roux AV (2021) Ambient fine particulate matter in Latin American cities: levels, population exposure, and associated urban factors. *Sci Total Environ* 772:145035. <https://doi.org/10.1016/j.scitotenv.2021.145035>
6. Parra Ramírez CA (2020) Estrategias de Responsabilidad Social y Ambiental de la Empresa HOLCIM en México. Universidad del Rosario

7. Abu-Allaban M, Abu-Qdais H (2011) Impact assessment of ambient air quality by cement industry: a case study in Jordan. *Aerosol Air Qual Res* 11(7):802–810. <https://doi.org/10.4209/aaqr.2011.07.0090>
8. Segovia Hermoza M (2018) Evaluación de impacto ambiental en la planta de Agregados Oropesa–Concretos Supermix SA-Cusco (Investigación Cuantitativa)
9. Jayadipraja EA, Daud A, Assegaf AH (2016) Air pollution and lung capacity of people living around the cement industry. *Public Health Indonesia* 2(2):76–83
10. Romero Cardenas A (2019) Criterios de implementación ISO 14000:2015 Caso Estudio Sector Fabricación de Cemento
11. Latorre ÁMLR, Tovar MHT (2017) Explotación minera y sus impactos ambientales y en salud El caso de Potosí en Bogotá. *Saúde Em Debate* 41(112):77–91. <https://doi.org/10.1590/0103-1104201711207>
12. Abdul Samad MS, Mohan P, Varghese GK, Shah IK, Alappat BJ (2020) Environmental forensic investigation of the air pollution from a cement manufacturing unit. *Environ Foren* 21(1):37–47. <https://doi.org/10.1080/15275922.2019.1694094>
13. Monteiro A, Vieira M, Gama C, Miranda AI (2017) Towards an improved air quality index. *Air Qual Atmosp Health* 10(4):447–455. <https://doi.org/10.1007/s11869-016-0435-y>
14. García Navarrete G, Rico Soto KG (2020) Sensores de bajo costo para el monitoreo de calidad de aire. *EPISTEMUS* 13(27):30–37
15. Ciobanu C, Istrate IA, Tudor P, Voicu G (2021) Dust emission monitoring in cement plant mills: a case study in Romania. *Int J Environ Res Public Health* 18(17):9096. <https://doi.org/10.3390/ijerph18179096>
16. Petsas C, Stylianou M, Zorpas A, Agapiou A (2021) Measurements of local sources of particulates with a portable monitor along the coast of an insular city. *Sustainability* 13(1):1–17. <https://doi.org/10.3390/su13010261>
17. Liu X, Schnelle-Kreis J, Zhang X, Bendl J, Khedr M, Jakobi G, Schlöter-Hai B, Hovorka J, Zimmermann R (2020) Integration of air pollution data collected by mobile measurement to derive a preliminary spatiotemporal air pollution profile from two neighboring German-Czech border villages. *Sci Total Environ* 722:137632. <https://doi.org/10.1016/j.scitotenv.2020.137632>
18. Kumar MK, Sreekanth V, Salmon M, Tonne C, Marshall JD (2018) Use of spatiotemporal characteristics of ambient PM<sub>2.5</sub> in rural South India to infer local versus regional contributions. *Environ Pollut* 239:803–811. <https://doi.org/10.1016/j.envpol.2018.04.057>
19. Lambey V, Prasad AD (2021) A review on air quality measurement using an unmanned aerial vehicle: water, air, and soil pollution. Springer Science and Business Media Deutschland GmbH, New York. <https://doi.org/10.1007/s11270-020-04973-5>
20. Eslava-Pedraza JE, Martínez-Sarmiento FA, Soto-Vergel AJ, Vera-Rozo EJ, Guevara-Ibarra D (2021) Diseño de un sistema de medición de material particulado mediante un vehículo aéreo no tripulado. *Aibi Revista de Investigación Administración e Ingeniería* 8(S1):1–15
21. Camarillo-Escobedo R, Flores JL, Marin-Montoya P, García-Torales G, Camarillo-Escobedo JM (2022) Smart multi-sensor system for remote air quality monitoring using unmanned aerial vehicle and LoRaWAN. *Sensors* 22(5):1706. <https://doi.org/10.3390/s22051706>
22. Yu S, Chang CT, Ma CM (2021) Simulation and measurement of air quality in the traffic congestion area. *Sustain Environ Res* 31(1):99. <https://doi.org/10.1186/s42834-021-00099-3>
23. Vivar Martínez E (2014) Cuantificación de material particulado PM<sub>10</sub> y su efecto toxicológico-ambiental, en la ciudad de Azogues. Universidad de Cuenca
24. Beketie KT, Angessa AT, Zeleke TT, Ayal DY (2022) Impact of cement factory emission on air quality and human health around Mugher and the surrounding villages, Central Ethiopia. *Air Qual Atmos Health* 15(2):347–361. <https://doi.org/10.1007/s11869-021-01109-4>
25. Wang S, Hao J (2012) Air quality management in China: issues, challenges, and options. *J Environ Sci* 24(1):2–13. [https://doi.org/10.1016/S1001-0742\(11\)60724-9](https://doi.org/10.1016/S1001-0742(11)60724-9)

# The Determinants of ICT Use by University Professors



Mounir Elatrachi and Samira Oukarfi

**Abstract** Technological skills are considered an essential factor in bridging the social gap between individuals. Are they equally important in tertiary education? This article aims to analyze the determinants of ICT use by university professors. The latter is considered a real pillar for the integration of information and communication technologies in education. To this end, we examine the behavior of 223 teachers spread across 8 districts, taking into account regional differences, which is not common in previous research. Using an ordinal probabilistic model, we find that teachers' experience significantly affects their level of instructional use of ICT. In contrast, teachers in less developed regions are more motivated to use ICT than teachers in developed regions.

**Keywords** Higher education · Teacher · Ordered probit · ICT use

## 1 Introduction

Over the past three decades, reforms in Morocco's education system and contingency planning have followed. However, higher education in Morocco is struggling to renew itself. His astonishing performance has sparked more and more graduates' worries about the future. Universities seem unable to respond to the increasingly pressing needs and sectoral development policies initiated by the state. Low internal and external returns remain the dominant feature, mainly in public institutions. According to the High Commission for Planning report (2017), 47.2% of students leave university without a degree and only 13.3% of students earn a bachelor's degree. Moreover, the latter has an unemployment rate of 25%. In this overwhelming situation, digital transformation seems to be the hope of leading the sector out of the long-term crisis.

---

M. Elatrachi (✉) · S. Oukarfi

LARMIG Laboratory, Hassan II University – FSJES Ain Sebaâ, Casablanca, Morocco  
e-mail: [mounirelatrachi@gmail.com](mailto:mounirelatrachi@gmail.com)

S. Oukarfi

e-mail: [samira.oukarfi@gmail.com](mailto:samira.oukarfi@gmail.com)



The current context of COVID-19 has helped reignite a national debate on the importance of integrating new technologies into education. This situation highlights the weaknesses of the university's e-learning system [1]. Noted that during this pandemic, most teachers involved in distance learning have relied essentially on their own means. At the student level, the authors argue that not all benefit from the same learning conditions.

Despite the turmoil of this period, Morocco could use this experience to launch the beginning of a new era in education in general and tertiary education in particular.

Within this general framework, our work includes identifying the determinants of ICT use by teachers and bring out the profile of those professors who are least likely to use new technologies for educational purposes.

To better conduct our research, we will first examine the empirical literature on academics' use of ICT, especially in higher education. We then introduce the databases and econometric model used. Finally, we discuss the analysis of factors influencing the use of ICT by academics in Morocco.

## 2 Empirical Contributions of ICT Use by Teachers

In this section, we present the results of a number of empirical studies examining the determinants of ICT use by university professors.

From the literature review, some researchers seem to be interested in the characteristics of teachers, as they are regarded as the cornerstone of the use of new technologies in the classroom. The aim was to determine the importance of certain determinants of ICT use in teaching. To this end, the empirical study grouped these determinants into three categories, namely, professors' individual factors, their technical profiles, and institutional-related environmental factors.

The first group of individual factors includes gender [2, 3], age, teaching experience [4, 5], and teaching discipline [6, 7].

The second set of variables related to teachers' technical skills includes teachers' experience with computers [8, 9].

Finally, the third set of variables contains establishment-related characteristics, namely their technical equipment (software and hardware) [10, 11].

### 2.1 *The Impact of Individual Teachers' Characteristics*

When researching the literature, the individual characteristics of teachers are very important. The goal was to understand which characteristics described the types of professors who most frequently used technology in classroom practice. Below we examine the impact of these characteristics in different contexts related to gender, age, teacher experience, and teaching discipline.

In research on the use of ICT in education, the focus is on gender issues. The literature does not provide consistent results in this regard. In some studies, the authors indicated no link between gender and ICT use in teaching [2, 7, 12, 13]. Examined the behavior of 1540 male and female teachers in Turkey and calculated gender-related scores related to ICT use and knowledge from 15 and 14 items. The author demonstrates that men score higher than women in both categories. However [12] affirms that there is no distinction between male and female academics in terms of how they use computers in the classroom [14]. Examines how 1209 Dutch teachers used digital learning resources and finds no gender differences in utilization. A Norwegian study [3] of 1072 academics showed no gender differences that were statistically significant. According to studies, all teachers place an equal emphasis on helping their pupils improve their digital information and communication abilities.

On the other hand, according to certain studies [15, 16], men appear to use instructional ICT tools more frequently than women. According to an Australian study [17] involving 929 academics, women use ICT with students less frequently because they experience higher technological fear [5]. Examine the habits of 529 male and 566 female Spanish teachers from various study cycles (from primary to higher education). The writers support the idea that men are superior when it comes to the utilization of ICT in education. They cite the fact that whereas male teachers place a greater emphasis on technological skills, which boosts their comfort utilizing ICT with students, female teachers place a greater emphasis on pedagogical skills. Using data from TALIS 2013 in the same country [8], consider that men are more likely than women to use ICT with students in the classroom.

The researchers' findings about the teachers' ages were quite inconsistent [18]. Reveals that the adoption and usage of technology by teachers varies with age in the United States, with younger professors (under 40 years old) using technology more frequently than their senior colleagues. Another study on academics' use of ICT was conducted in the same nation by [19] with a sample size of 800 teachers. The author discovers that younger teachers (under 44 years old) exhibit a higher degree of computer comfort than more experienced ones. They have trouble implementing new technologies to enhance their instructional strategies. According to [20], younger teachers use ICT more frequently than their elder counterparts.

Laabidi [21] examines the behavior of 163 university professors in the Moroccan context. He comes to the conclusion that teachers under the age of 40 frequently use ICT more than their more senior counterparts. These findings concur with [3, 22–24]. Lubis et al. [25] investigate the technological habits of 260 teachers working in Indonesian private universities. According to these authors, the most significant factor in a research of technology use is age. The majority of teachers who use ICT in the classroom are under the age of 40, whereas teachers above the age of 60 utilize it very little. This outcome is in line with that of [26] in Mauritius, who discovered that younger professors (those under the age of 40) employ ICT more frequently in their educational strategies than their more senior colleagues.

Furthermore, additional findings confirm that the frequency of use of new technology is not much impacted by the age of teachers. According to [27], age does not show the difficulties that teachers perceive. The authors contend that regardless of

age, technology support solutions should be built to fulfill the demands of teachers with limited technological skills.

Limone et al. [28] concludes that age has no bearing on the instructional usage of new technology from a study of 374 Italian instructors. They believe that young instructors with established skills can access ICT more widely. To effectively integrate new technologies in teaching, it appears that all teachers—regardless of age—need pedagogical models and ongoing professional development.

There are no definitive findings in the literature about teachers' experiences. According to other researchers, the experience effect is not particularly significant. According to [29], teachers' prior expertise is not a crucial factor that should be taken into account when deciding how to use ICT in the classroom. In a similar line [30], think that there is no discernible connection between the professors' experience and the pedagogical use of computers.

According to other researchers, though, teaching experience has an impact on how academics accept and use new technologies. Academics with fewer than three years' experience are, in fact, more likely to include ICT into their lessons, according to [31]. Additionally, he discovers that professors with between 10 and 20 years of experience have a tendency to use less computers in their classrooms. Given that young teachers have grown up with technology and that electronic games have been widely accessible since the mid-1970s, the author does not find this result to be surprising.

According to [21], young academics in the Moroccan context who have less teaching experience exhibit higher levels of ICT integration than their older peers who have more classroom experience. Based on this finding, the author draws the conclusion that teaching experience appears to have a significant impact on how well academics integrate computer technology into the classroom.

New professors are less likely to incorporate new technologies into their lessons in the United States, according to [32]. The researchers contend that these academics frequently worry about going against the establishment's norms for education and are unconvinced by the use of new technologies, particularly if it is not a typical practice inside the establishment. Additionally in the US [33], discovers that professors with higher experience are more likely to employ ICT. In contrast to new professors who attempt to assimilate teaching methods, the author notes that these teachers are more at ease with the teaching processes and can therefore experiment with new ones.

Few researchers have focused on how university professors' teaching discipline affect their usage of educational technologies. Waight et al. [7] notes that in the United States, teachers of the hard sciences are more likely than those of the humanities to incorporate technology into their courses. This discrepancy is mostly explained by the incorporation of computer-assisted experiments by earth science, biology, geology, and physics teachers, which allowed them to increase their comfort level with technology. Chowdhury [34] show that teachers of social sciences and the arts use ICT less frequently than professors of the natural sciences, engineering, and technology. Clipa et al. [35] examine the actions of 248 professors from various academic disciplines in the Roman context. They believe that teachers in technical

streams are more likely than teachers in didactic streams to incorporate new technology. However, [36] indicates that teachers of social sciences and humanities are more likely to adopt modern technology in sub-Saharan nations. Additionally, [37] in Singapore and [38] in Wales draw the conclusion that English teachers use ICT more frequently than their colleagues in mathematics. According to these authors, literary professors frequently use word processing software to communicate with their students since it enables them to express themselves more freely and improve their language proficiency.

## ***2.2 The Technological Profile of Teachers***

In what follows we discuss the empirical findings of the professors' technological characteristics of teachers, which include technological experience and ICT skills.

The amount of time a person has spent working with new technologies is used to understand their technological experience [39]. According to [40], it refers to the practice of using technical tools and the growth of skills one learns through technology use.

One of the main barriers to the integration of technology in teaching in Turkey, according to [41], is the lack of expertise professors have using ICT for educational purposes. According to [12] in the same context, teachers who have prior computer skills use ICT more frequently.

Petrogiannis [42], in Greece, investigates the relationship between teachers who have and have not used computers. He discovers that academics with computer experience are more prepared to employ ICT in their courses than those without experience. For their part, [43] demonstrate how teachers' usage of ICT in the classroom is positively impacted technological experience. He and Freeman [44] examines the behavior of 243 professors in the US. They confirm that prior experience significantly raises the likelihood of using technologies in teaching.

The impact of pedagogical experience in ICT use is stronger than that of the individual characteristics, such as age, gender, and teaching experience, according to an Indian study [45] that examines the behaviors of 515 teachers.

In summary, the findings of multiple studies support the notion that teachers' technological expertise and the use of technology with students in the classroom are closely related [10, 16, 46, 47].

The mastery of computer tools is the second factor linked to the technical profile of teachers. Without a basic understanding of computers among teachers, technology adoption models would be ineffective. In fact, the level of computer literacy (or mastery of computers and software) appears to have a considerable influence on the desire to engage in an integration process [48]. Almerich et al. [5] take into account that ICT skills can be viewed as a collection of knowledge and abilities that academics must develop in a variety of technological tools in order to fully integrate them into their teaching methods.

Gil-Flores et al. [8] divide university professors into two groups. The first category consists of teachers who “have a moderate or high demand for ICT training,” while the second group consists of those who “have a low or no requirement.” The researchers observe that the first group uses less devices for schooling than the second group does. They come to the conclusion that pedagogues can employ technological tools more frequently in the classroom since they have mastered them.

The behaviors of 163 university teachers in Morocco were examined by [49], who found that those who lacked fundamental computer knowledge and skills were hesitant to include ICT into their courses. This means that professors who are highly skilled at using ICT tools exhibit higher levels of application of computer technology than do teachers who are less so. In other words, professors who are more proficient with computers use ICT at a higher rate. As a result, limited contact with technological tools and a lack of computer skills generally results in limited confidence in the use of technologies in front of students.

### 3 Database

We chose to create our own databases since there were none available for evaluating the relationships between professor’ behaviors and their usage of ICT. We decided to conduct the survey ourselves using a paper questionnaire (PAPI). Although there are multiple ways to administer a questionnaire (by email, online, etc.), we choose the PAPI survey approach. This approach is more suitable in our situation because teachers—whether or not they use ICT—are our main target. We decided not to distribute the questionnaire online because this mode of administration would oust academics who do not use ICT, which would have posed a problem of selection bias.

Two hundred and twenty three individuals are included in the database, distributed among 6 cities and 4 regions. Professors located in areas with various development’ level were given the questionnaires. We divided the regions into three categories using the HRDI<sup>1</sup>; developed, intermediate, and poor. A developed region is Casablanca-Settat, an intermediate region is Rabat-Sale-Kenitra, Marrakech-Safi and a poor region is Béni Mellal-Khenifra. The objective is to examine how professors’ behaviors vary depending on regional development. Note that men make up 67% of the respondents. In terms of age, the average is 44.5 years old (with a standard deviation of 8 years old).

---

<sup>1</sup> Regional Human Development Index is elaborated by the High Commission for Planning on the basis of the findings of the 2014 General Population and Housing Census.

## 4 Econometric Modeling

### 4.1 Specification of the Ordered Probit Model

The endogenous variable is a qualitative polytomous variable taking 3 modalities, and organized as follows:

$$\text{the educational use of ICT} = \begin{cases} 1 & \text{if the teacher has a moderate use of ICT in teaching} \\ 2 & \text{if the teacher has an often uses ICT in teaching} \\ 3 & \text{if the teacher has continuously used ICT in teaching} \end{cases}$$

An ordered probit model is suggested by the polytomous and ordered character of the dependent variable. Consider the following model with one dependent variable  $y$  and three modalities 1, 2, and 3. The latent variable  $y_i^*$  is the sum of a deterministic component and a random error:

$$y_i^* = \beta' x_i + \varepsilon_i$$

$y_i^*$  is the unobserved dependent variable “Intensity of pedagogical use of ICT” of the teacher  $i$ ,  $x_i$  is a vector of observed professors’ characteristics,  $\beta$  is a group of parameters to be estimated and  $\varepsilon_i$  is a random error term for the teacher  $i$ .

We have:

$$\begin{aligned} y_i &= 1 & \text{if } y_i^* < \mu_1 \\ y_i &= 2 & \text{if } \mu_1 \leq y_i^* < \mu_2 \\ y_i &= 3 & \text{if } y_i^* \geq \mu_2 \end{aligned}$$

Or by replacing  $y^*$  by  $\beta' x_i + \varepsilon_i$

$$\begin{aligned} y_i &= 1 & \text{if } \beta' x_i + \varepsilon_i < \mu_1 \\ y_i &= 2 & \text{if } \mu_1 \leq \beta' x_i + \varepsilon_i < \mu_2 \\ y_i &= 3 & \text{if } \beta' x_i + \varepsilon_i \geq \mu_2 \end{aligned}$$

With:

- $y_i = 1$  if the teacher has a moderate use of ICT in teaching;
- $y_i = 2$  if the teacher has an often uses ICT in teaching;
- $y_i = 3$  if the teacher has continuously used ICT in teaching.

The parameters  $\beta$  and the thresholds  $\mu$  are to be estimated. The latter can be interpreted as being the bounds which separate the different modalities of the “The educational use of ICT” variable, such that:

$\mu_1$ : limit which separates the moderate use category and the often uses of ICT in teaching;

$\mu_2$ : limit which separates the often uses category and the continuously uses ICT in teaching.

## 4.2 Estimation Results

The outcome of the ordered probit model is shown in the Table 1.

The outcome of this estimation shows that the  $R^2$  is equal to 0.2804, it remains satisfactory given the qualitative nature of the individual data. We used the method of White, allowing us to have corrected standard deviations for heteroskedasticity, and hence secure robust standard deviations which will be used in the construction of the corrected student tests.

Then, by reading the Wald chi statistic (92.54) and its probability ( $\text{Pr} = 0.000$ ), we can conclude that the estimated coefficients are different from zero.

Based on these estimations, we will explore the findings in the sections that follow, as well as the weight of independent factors using marginal effects, which gauge the impact of an increase of a given exogenous variable on the endogenous variable, in order to analyze how exogenous factors affect each individual's use of ICT in higher education.

## 5 Discussion

The model's findings indicate the use of technologies according to gender is not significant. So there is no difference in the use of ICT between male and female. This result is consistent with that of [50] who stipulate that the academic use of ICT is independent of the gender of the professors. Similarly, [19] found no gender differences in attitudes in their US study. The same observation can be observed in the study by [51] in the Malaysian context. The authors deny any difference between male and female professors. In light of the evidence of our study, we can affirm that, in the Moroccan context, the use of new technologies in teaching does not differ according to gender, thus belying all stereotypes stipulating that female academics are less likely to use ICT.

It is recognized that age is an essential variable impacting the use of new technologies. Several studies have focused on this aspect because it reflects intergenerational differences. Our empirical results show that age has a negative impact on the degree of use of new technologies. Professors aged 40 years old and under are 5.1% more likely to use ICT continuously in teaching than those aged 50 years old and over. This probability is 2.5% for teachers aged between 41 and 50 years old. This result corroborates those of [11, 52–56]. This result supports that young teachers integrate technological tools more in their teaching than their older counterparts, because they are more familiar with digital spaces. In addition, they have been exposed to ICTs more than their older peers, which allows them to use them with more ease.

**Table 1** Outcome of the ordered probit model

Variables		Coefficients	T-stat	Marginal effects on		
				Y = 1	Y = 2	Y = 3
Gender	Male	−0.158 (n.s.)	0.491	0.076	0.052	−0.060
	Female	Réf	—	—	—	—
Age	< 40 years old	0.839***	0.008	−0.027	0.053	0.051
	Between 41 and 50 years old	0.545***	0.000	−0.033	0.021	0.025
	More than 50 years old	Réf	—	—	—	—
Experiences	10 years and less	0.653***	0.008	0.079	0.167	−0.024
	Between 10 and 20 years	0.245***	0.000	0.102	0.222	−0.025
	More than 20 years	Réf	—	—	—	—
Teacher's grade	Higher education professor	0.323***	0.000	−0.142	−0.109	0.123
	Accredited professor	0.036***	0.003	−0.183	−0.117	0.135
	Assistant professor	Réf	—	—	—	—
Teaching discipline	Economics	0.197***	0.008	−0.117	0.428	0.545
	Sciences	0.286***	0.000	−0.293	0.192	0.486
	Literary	Réf	—	—	—	—
Type of establishments	Public	−0.567***	0.001	0.328	0.520	−0.553
	Private	Réf	—	—	—	—
Type of access to the establishment	Open access	0.232***	0.005	−0.129	−0.072	0.085
	Regulated access	Réf	—	—	—	—
ICT skills	Advanced skills	0.728***	0.000	−0.031	−0.430	0.462
	Moderate skills	0.236***	0.001	−0.040	−0.226	0.267
	Basic skills	Réf	—	—	—	—
Level of development of the region	Low	0.839***	0.001	−0.056	−0.240	0.296
	Intermediate	0.607***	0.004	−0.027	−0.203	0.231
	Developed	Réf	—	—	—	—

n.s., not significant; (\*): Significant at a threshold of 1%, (\*\*): Significant at a threshold of 5%, (\*\*\*): Significant at a threshold of 10%

As for the experience of professors, it positively influences the use of ICT. This implies that as experience increases, the pedagogical use of ICT by the teacher increases. Thus, our results reveal that academics with 10 years' experience or less are 2.4% less likely to use ICT continuously than their peers with more than 20 years



of experience. This probability is 2.5% for professors with between 10 and 20 years of experience.

So and Kim [57] state that although beginning teachers have the skills and tools to use technology, they seem less prepared to use it effectively in teaching. For our part, our econometric results confirm this observation. Moreover, the lack of pedagogical training for young teachers means that they must acquire it with experience, which delays their use of ICT in classroom.

Few research has focused on the impact of the teacher's grade on their use of ICT for academic purposes. Our results conclude that Accredited Professor are 13.5% more likely to opt for ICT than Assistant Professor. This probability is 12.3% for Higher Education Professor. This result is explained by the graduation system applied in Morocco. Indeed, teachers aspiring to promotion must meet certain criteria, including pedagogical innovation. This section includes NTIC media, namely Slideshows, Tutorials, Educational web page, Production of online content (MOOC). These criteria encourage Accredited Professor to use new technologies more, and continue to use them by moving to Higher Education Professor status. However, it should be noted that this practice was only initiated from 2015. We can conclude that Higher Education Professor after 2015 use ICT more than their counterparts before 2015. This conclusion is supported by the impact of the age variable (academics aged between 41- and 50-years old use ICT more than those aged 50 years old and over).

Researchers have given particular importance to the impact of the teaching discipline on ICT use, but without consensus in the results. Howard and Maton [38] and Tay et al. [37] find a marked disparity in the use of new technologies between teachers of literary and scientific subjects, and conclude that ICT use is higher by teachers of English compared to their peers in mathematics. On the other hand, [58] affirm that teachers in physical sciences, social sciences and geography have, on average, a strong ICT use compared to those of literary subjects (English, French, and history). Hennessy et al. [59] also show that English teachers are less likely to use new technologies than math and science teachers.

As far as we are concerned, our econometric results show that teachers of economics subjects are 54.5% more likely to continuously use ICT in their subjects than professors of literary subjects. This probability is 48.6% for teachers of science subjects. This conclusion is consistent with that of [58, 59]. This difference in use between teachers of science and literature is attributed to the history of the subject. Professors may not be willing to use technology that does not seem compatible with the norms of their discipline's culture. Selwyn [60] describes that the computer is more consistent with the history of certain subjects and more integrated into their practice than others. The author adds that there is a feeling of appropriation of ICT among teachers of science subjects and a suspicion toward these tools among professors of didactic disciplines. We can conclude that this statement by [60] still seems relevant in the Moroccan context.

Regarding the influence of ICT skills, researchers agree on its close link with the pedagogical use of new technologies [8, 10, 16, 46, 47]. Ahmed and Kurshid [61] concluded that professors with higher computer skills use more ICT in the learning

process than academics with less computer and software skills. Our results confirm this relationship. Indeed, the more ICT skills of teachers increases, the more the probability of using ICT increases. Professors with moderate ICT skills are 26.7% more likely to use new technologies continuously than their colleagues with a basic ICT skills. This probability increases to 46.2% for teachers with an advanced ICT skills. As for the impact of the establishment' type on the educational use of ICT, the researchers find mixed results. On the one hand, [62] state that teachers in private universities are better trained in the use of information technologies than their peers in public universities. Their result is consistent with that of [52, 63]. On the other hand, [61] show that professors in public universities frequently use ICT in teaching compared to those in the private sector. [64], for their part, find that private establishments are better equipped with ICT than public establishments, but with no difference in use at professors' level. In the Moroccan context, we note that teachers in public establishments are less likely to use ICT than their private colleagues, with a probability of 55.3%. The same observation concerning the type of access to the establishment. Professors working in open access establishments are 8.5% more likely to continuously opt for ICT compared to teachers in regulated-access establishments. It should be noted that, to our knowledge, no study has taken into consideration the difference in ICT use according to the type of access to the establishment. This is an innovative and major contribution of our study. We are the first to have introduced this variable given the different operating conditions of these establishments.

In Morocco, establishments with regulated access enjoy special attention from the government, given the nature of the profiles they train, and their importance in sectoral strategies. For example, establishments training engineers and physician benefit from substantial budgetary envelopes in order to achieve the objectives of government initiatives aimed at training 10,000 engineers<sup>2</sup> and 3300 physicians<sup>3</sup> per year by 2020. As a result, the unit operating cost per student in engineering sciences is 28.216 Dhs per year. In technology, it is 26.141 Dhs, for medical sciences it is up to 37.000 Dhs.<sup>4</sup> In addition, 40% of university professors work in establishments with regulated access, for the 13% of student numbers in these establishments. This shows that the student/teacher ratio is much higher in institutions with restricted access than in those with open access.

In conclusion, our econometric modeling shows that despite the favorable conditions enjoyed by regulated access establishments, their professors use new technologies less compared to those in open access establishments.

---

<sup>2</sup> The government program "Initiative 10,000 engineers" was designed to support the Industrial Emergence Plan in 2010. The number of engineering graduates in 2018 was 7366 laureates, i.e., an achievement rate of 74% of the program's objectives (Ministry of Higher Education and Scientific Research).

<sup>3</sup> Launched in 2007, this government initiative aims to train 3300 doctors per year by 2020. The objective is to increase medical density from 6 to 10 doctors per 10,000 inhabitants. In 2018, the number of winners reached 2051 doctors and pharmacists, i.e., an achievement rate of 62% of the objectives set (Ministry of Higher Education and Scientific Research).

<sup>4</sup> 2018 statistics from the Ministry of Higher Education and Scientific Research.

Finally, the region's development level impacts ICT use by professors [65, 66]. Several researchers support this postulate. Lu et al. [67] state that establishments based in urban areas favor the use of ICT, because they have better technological infrastructures, and attach importance to the construction of digital educational resources in order to facilitate their use by teachers. In Morocco, higher education establishments are generally based in urban areas. Therefore, we opted for the Regional Human Development Index in order to differentiate the region's development level. The modeling reveals a counter-intuitive result, professors in regions with low and intermediate development' level are more likely (29.6 and 23.1%, respectively) to use ICT continuously in teaching than those in regions with a high development' level.

The lack of basic educational infrastructure in less developed regions pushes teachers to make greater use of new technologies in order to better assist students in their teaching. This result is in line with that found by [68] in Morocco, in which we they found that students in poor and intermediate regions use ICT more in studies than those in developed regions. This finding suggests that teachers in poor and intermediate regions behave similarly to students.

The results obtained allow us to draw a typical profile of professors who are less likely to use ICT in their pedagogy. In general, these are teachers (male and female) of literary disciplines based in developed regions, working in the public sector and who have just started their professional careers.

The configuration of this profile is relevant since it can help guide educational policy measures aimed at increasing the presence of ICT in higher education, and take advantage of its potential. The programs implemented have focused exclusively on the "technical" aspect of ICT in terms of the quality of platforms and Internet connections. This technicist approach has led to the neglect of much more important factors, namely the age and experience of the teachers, their grade, the teaching discipline, their ICT skills, as well as the establishment and the region in which they teach. Our results clearly show that infrastructure alone is not enough. The ICT use is limited by the professors' characteristics, in particular because of the high training needs for the pedagogical use of new technologies. Therefore, to compensate the lack of training for professors at the start of their careers, we propose to integrate training on the pedagogical use of ICT in the last year of Ph.D., since 88% of Ph.D. students wish to integrate the sector of Higher Education. These trainings will enable them to acquire the pedagogical skills necessary for the effective use of new technologies. Also, ongoing training should be offered to update and improve the technological skills of professors. These tools will eventually make it possible to create synergies between novice teachers (with technological skills) and experienced teachers (with pedagogical skills).

## 6 Conclusion

The aim of the study was to analyze the determinants of ICT use by professors in tertiary education, as well as to highlight the profile of academics least likely to use new technologies for educational purposes.

Regarding the determinants of ICT use in teaching, we have noticed that the key factors are age, experiences, professors' grade, teaching discipline, ICT skills, characteristics of establishment, and region' level of development. Finally, the professors that are the less likely to use ICT in classroom are professors (male and female) of literary disciplines based in developed regions, working in the public sector and who have just started their professional careers.

Obviously, these results need to be interpreted with caution. There are various limitations on this work, some of which are connected to the data used. These results must obviously be taken with caution. This work has certain limitations, some of which are related to the nature of data used. For instance, only considering qualitative variables while determining ICT-related factors. Although it is not possible, we would have preferred to retain other quantitative characteristics, such as the budget allocated per establishment or region. Additionally, these restrictions represent potential directions for future research, whose investigation can give rise to fresh scientific contributions.

## References

1. Benseddik M (2020) L'université marocaine à l'épreuve du Covid-19. AEGIS, p 37
2. Erdogdu F, Erdogdu E (2015) The impact of access to ICT, student background and school/home environment on academic success of students in Turkey: an international comparative analysis. *Comput Educ* 82:2649
3. Siddiq F, Scherer R, Tondeur J (2016) Teachers' emphasis on developing students' digital information and communication skills (TEDDICS): a new construct in 21st century education. *Comput Educ* 92:1–14
4. Alazam A, Bakar AR, Hamzah R, Asmiran S (2012) Teachers' ICT skills and ICT integration in the classroom: the case of vocational and technical teachers in Malaysia. *Sci Res* 3:70
5. Almerich G, Orellana N, Suarez-Rodríguez J, Díaz-García I (2016) Teachers' information and communication technology competences: a structural approach. *Comput Educ* 100:110–125
6. Greenberg MT, Kusche CA (1998) Blueprints for violence prevention: the PATHS project, vol 10. Institute of Behavioral Science, Regents of the University of Colorado, Boulder, CO
7. Waight N, Chiu MM, Whitford M (2014) Factors that influence science teachers' selection and usage of technologies in high school science classrooms. *J Sci Educ Technol* 23(5):668–681
8. Gil-Flores J, Rodríguez-Santero J, Torres-Gordillo J-J (2017) Factors that explain the use of ICT in secondary-education classrooms: the role of teacher characteristics and school infrastructure. *Comput Hum Behav* 68:441–449
9. Zhao Y, LeAnna Bryant F (2006) Can teacher technology integration training alone lead to high levels of technology integration? A qualitative look at teachers' technology integration after state mandated technology training. *Elect J Integ Technol Educ* 5:53–62
10. Bingimlas KA (2009) Barriers to the successful integration of ICT in teaching and learning environments: a review of the literature. *Eurasia J Math Sci Technol Educ* 5(3):235–245

11. Hermans R, Tondeur J, van Braak J, Valcke M (2008) The impact of primary school teachers' educational beliefs on the classroom use of computers. *Comput Educ* 51(4):1499–1509
12. Tezci E (2009) Teachers' effect on ICT use in education: the Turkey sample. *Soc Behav Sci* 1:1285–1294
13. Vekiri I, Chronaki A (2008) Gender issues in technology use: perceived social support, computer self-efficacy and value beliefs, and computer use beyond school. *Comput Educ* 51(3):1392–1404
14. Kreijns K, Van Acker F, Vermeulen M, Van Buuren H (2013) What stimulates teachers to integrate ICT in their pedagogical practices? The use of digital learning materials in education. *Comput Hum Behav* 29(1):217–225
15. Papanastasiou EC, Angeli C (2008) Evaluating the use of ICT in education: psychometric properties of the survey of factors affecting teachers teaching with technology (SFA-T3). *Educ Technol Soc* 11(1):69–86
16. Van Braak J (2001) Individual characteristics influencing teachers' class use of computers. *J Educ Comput Res* 25(2):141–157
17. Jamieson-Proctor R, Burnett P, Finger G, Watson G (2006) ICT integration and teachers' confidence in using ICT for teaching and learning in Queensland stateschools. *Aust J Educ Technol* 22:511–530
18. Prensky M (2001) Digital natives, digital immigrants. *Horizon* 9(5):1–6
19. Ahadiat N (2008) Technologies used in accounting education: a study of frequency of use among faculty. *J Educ Bus* 15:123–133
20. Sriyono H, Ino L (2019). Correlation between English teachers' computer skills, perceptions, demographics, and ICT integration at vocational schools of south Konawe regency. *J Pendidikan Bahasa*
21. Laabidi H (2017) Investigating the influence of teaching experience on the use of ICT in education. *EFL J* 2:15–31
22. Cabero J, Barroso J (2016) ICT teacher training: a view of the TPACK model. *Cult Educ* 28:633–663
23. Gudmundsdottir G, Hatlevik O (2018) Newly qualified teachers' professional digital competence: implications for teacher education. *Eur J Teach Educ* 41:214–231
24. Mohamad A, Idrus S, Ibrahim A (2019) The impact of age, gender, culture and language toward the use of ICT for teaching and learning by lecturers in university of Tripoli, Libya. *Int J Acad Res Bus Soc Sci* 14:71–82
25. Lubis A, Idrus S, Sarji A, Lubis Z, Sutrisno S (2020) Investigating the moderating effect of demographic variables on ICT usage and learning process quality of higher education in Medan, Indonesia. *Int Conf Emerg Comput Technol Sports*
26. Perienen A (2020) Frameworks for ICT integration in mathematics education: a teacher's perspective. *J Math Sci Technol Educ* 16:em1845
27. Lane C, Lyle H (2011) Obstacles and supports related to the use of educational technologies: the role of technological expertise, gender, and age. *J Comput Higher Educ* 23:38–59
28. Limone P, Sinatra M, Tanucc G, Monacis L (2019) The utilitarian versus hedonic teacher acceptance of ICT use. *Turkish Online J Dist Educ* 20(4):1–10
29. Becker H (1999) Internet use by teachers: conditions of professional use and teacher-directed student use. Center for Research on Information Technology and Organizations, Irvine
30. Dusick D, Yildirim S (2000) Faculty computer use and training: identifying distinct needs for different populations. *Commun College Rev* 27(4):33–47
31. Adams NB (2002) Educational computing concerns of postsecondary faculty. *J Res Technol Educ* 34(3):285–303
32. Conway C, Micheel-Mays C, Micheel-Mays L (2005) A narrative study of student teaching and the first year of teaching: common issues and struggles. *Bull Council Res Music Educ* 14:65–77
33. Gorder L (2008) A study of teacher perceptions of instructional technology integration in the classroom. *Delta Pi Epsilon J* 50:63–76

34. Chowdhury M (2009) The relationship between information and communication technologies integration and improvement in teaching as perceived by college instructors. Thèse de doctorat en sciences de l'éducation
35. Clipa O, Colomeischi A, Mari D (2014) Students perceptions upon ICT in university training process. In: Proceedings of the 10th international scientific conference elearning and software for education. Bucharest
36. Meso D, Musa F, Mbarika V, Okoli C, Byrd T, Delta P (2005) Toward sustainable adoption of technologies for human development in Sub-Saharan Africa: precursors, diagnostics, and prescriptions. *Commun Assoc Inform Syst* 15:33
37. Tay L, Lim S, Lim C, Koh J (2012) Pedagogical approaches for ICT integration into primary school English and mathematics: a Singapore case study. *Aust J Educ Technol* 28:740–754
38. Howard S, Maton K (2013) Technology knowledge: an exploration of teachers' conceptions of subject-area knowledge practices and technology integration. In: AERA 2013 SIG-computer and internet applications in education, pp 1–8
39. Ropp MM (1999) Exploring individual characteristics associated with learning to use computers in preservice teacher preparation. *J Res Comput Educ* 31:402–416
40. Thompson R, Compeau D, Higgins C (2006) Intentions to use information technologies: an integrative model. *J Org End User Comput* 27:14–15
41. Usluel K, Demiraslan Y, Kuskaya Mumcu F (2007) Integrating ICT into classrooms: a note from Turkish teachers. In: Carlsen R, McFerrin K, Price J, Weber R, Willis D (eds) Proceedings of SITE 2007: society for information technology teacher education international conference. Association for the Advancement of Computing in Education (AACE), San Antonio, TX, pp 1569–1575
42. Petrogiannis K (2010) The relationship between perceived preparedness for computer use and other psychological constructs among kindergarten teachers with and without computer experience in Greece. *J Inform Technol Impact* 10:99–110
43. Seraji N, Ziabari R, Rokni S (2017) Teacher's attitudes towards educational technology in English language institutes. *Int J English Linguist* 7:176–185
44. He J, Freeman L (2019) Are men more technology-oriented than women? The role of gender on the development of general computer self-efficacy of college students. *J Inform Syst Educ* 21:203–212
45. Chandan S, Prema B (2019) Demographic and other influencers of teachers' perception about ICT adoption in the curriculum. *Int J Bus Insights Transf* 12:41–45
46. Keijo S (2013) No pain, no gain? Educational use of ICT in teaching, studying and learning processes: teachers' and students' views. Lapland University Press, London
47. Vanderlinde R, Aesaert K, Van Braak J (2014) Institutionalised ICT use in primary education: a multilevel analysis. *Comput Educ* 72:1–10
48. Larose F, Grenon V, Lafrance S (2002) Pratiques et profils d'utilisation des TICE chez les enseignants d'une université. In: Guir R (ed) *Pratiquer les TICE, Former les enseignants et les formateurs à de nouveaux usages*. De Boeck, Bruxelles, pp 23–47
49. Khaloufi A, Laabidi H (2017) An examination of the impact of computer skills on the effective use of ICT in the classroom. *Indon J EFL Linguist* 2(1):53–69
50. Sieverding M, Koch SC (2009) Self-evaluation of computer competence: how gender matters. *Comput Educ* 52(3):696–701
51. Wong SL, Hanafi A (2007) Gender differences in attitudes towards information technology among Malaysian student teachers: a case study at Universiti Putra Malaysia. *Educ Technol Soc* 10(2):158–169
52. Gómez-Fernández N, Mediavilla M (2019) What are the factors that influence the use of ICT in the classroom by teachers? Evidence from a census survey in Madrid. Institut d'Economia de Barcelone
53. Mamun S, Rahman MM, Danaher PA (2015) The determinant of faculty attitude to academic (over-) work load: an econometric analysis. *J Dev Areas* 49:373–385
54. Rahimi M, Yadollahi S (2011) ICT use in EFL classes: a focus on EFL teachers' characteristics. *World J English Lang* 1:17

55. Suarez JM, Almerich G, Orellana N, Belloch C (2012) The use of ICTs by non university's faculty. Basic model and influence of personal and contextual factors. *Revista Iberoamericana de Evaluacion Educativa* 5(1):249–265
56. Wong E, Li S (2008) Framing ICT implementation in a context of educational change: a multilevel analysis. *School Effect School Improv* 19(1):99–120
57. So HJ, Kim B (2009) Learning about problem based learning: student teachers integrating technology, pedagogy and content knowledge. *Aust J Educ Technol* 25(1):101–116
58. Cuckle P, Clarke S (2002) Mentoring student-teachers in schools: views, practices and access to ICT. *J Comput Assist Learn* 18:330–340
59. Hennessy S, Ruthven K, Brindley S (2014) Teacher perspectives on integrating ICT into subject teaching: commitment, constraints, caution, and change. *J Curricul Stud* 37:155–192
60. Selwyn N (1999) Why the computer is not dominating schools: a failure of policy or a failure of practice? *Camb J Educ* 29(1):77–91
61. Ahmed H, Kurshid F (2016) Usage of information and communication technology among public and private sector university faculty. *J Element Educ* 26:39–55
62. Yasmeen S, Alam M, Mushtaq M, Bukhari M (2015) Comparative study of the availability and use of information technology in the subject of education in public and private universities of Islamabad and Rawalpindi. *SAGE Open*, New York, pp 1–7
63. Afridi T, Chaudhry H (2019) Technology adoption and integration in teaching and learning at public and private universities in Punjab. *Bull Educ Res* 14:121–143
64. Haque H, Shahriar F (2016) Quality of ICT facilities at the tertiary level education in Bangladesh: public versus private university. *J Sci Eng* 36:13
65. Rao J, Wu L, Xu J, Liu Y (2017) Empirical study on the difference of teachers' ICT usage in subjects, grades and ICT training. In: *International symposium on educational technology*, pp 58–61
66. Wu D, Li C-C, Zhou W-T, Tsai C-C, Lu C (2019) Relationship between ICT supporting conditions and ICT application in Chinese urban and rural basic education. *Asia Pacific Educ Rev* 20:147–157
67. Lu C, Tsai C-C, Wu D (2015) The role of ICT infrastructure in its application to classrooms: a large scale survey for middle and primary schools in China. *Educ Technol Soc* 18(2):249–261
68. Elatrachi M, Sattar H, Oukarfi S (2022) The impact of ICT use on the academic student performance in Morocco. In: Arai K (eds) *Advances in information and communication*. FICC 2022. Lecture notes in networks and systems, vol 438. Springer, Cham

# Load Capacity Study on the Flora Path of the Manglares Churute Ecological Reserve



Miriam Vanessa Hinojosa-Ramos , Marcelo Leon , Paulina Leon ,  
Viviana Tomala , and José Maldonado-Quezada 

**Abstract** This article precedes a research project which focuses on the design of a nature tourism product in the Churute Manglares Ecological Reserve, as a means to continue increasing the tourism potential of our country. Based on this purpose, the development of the work was framed in one of the existing methodologies for this purpose, serving as input the relevant, objective and reliable information obtained from the key actors linked to the reserve. In this sense, Sendero La Flora was chosen as a strategic location for the development of the proposal, seeking to relate the conditions of the tourist environment with the type of product to be offered, in order to satisfy the needs of current and potential visitors, using for this very useful techniques such as interviews, surveys and focus group. The work carried out allowed us to analyze the destination, conceptualize the tourist product; and finally, propose its design, specifying all its respective components.

**Keywords** Ecotourism · Tourist product · REMCH · Ecological reserve

---

M. V. Hinojosa-Ramos (✉)

Instituto Superior Tecnológico Vicente Rocafuerte, Guayaquil, Ecuador

e-mail: [mhinojosa@istvr.edu.ec](mailto:mhinojosa@istvr.edu.ec)

M. Leon

Universidad ECOTEC, Samborondon, Ecuador

P. Leon

University of Malaga, Malaga, Spain

V. Tomala

Universidad Cesar Vallejo, Trujillo, Perú

J. Maldonado-Quezada

Universidad Nacional de Loja, Loja, Ecuador



## 1 Introduction

The Churute Mangroves Ecological Reserve (REMCH) has been part of the National System of Protected Areas since 1979, the year it was created by Interministerial Agreement No. A-322 of July 26. In 1990, REMCH gained global importance by being included among the RAMSAR sites for mangrove protection within the framework of wetland conservation. In addition to this, it stands out especially for being an important area for bird conservation in the Tropical Andes (IBA). In turn, in 1998, the Reserve Management Plan was created in order to achieve sustainable development of tangible and intangible resources from it [1].

REMCH is located to the south of the Guayas province, in the Naranjal and Guayaquil cantons, in the Taura parish. Currently, it covers approximately 55,212 hectares, a characteristic that makes it one of the largest marine-coastal reserves in Continental Ecuador, located in the interior estuary of the Gulf of Guayaquil and the lower basin of the Guayas River. Six islands are included in the area of the reserve: Matorrillos, Los Ingleses, Los Álamos, Malabrigo, Cabeza de Mate and Churutillo [2].

Being located where the fresh water from the rivers that descend from the mountains and the saline water that comes from the sea converge, it is identified as the first mangrove protected area on the Ecuadorian continental coast. In addition to covering the largest extension of mangroves in Ecuador, the reserve also protects a freshwater lagoon and dry and mist ecosystems found in the hills of the Churute mountain range that reach up to 680 m.s.n.m, serving as a refuge for many species [3].

From the structural point of view of the Reserve, there are four plant formations: mangrove forests, dry forests, humid cloud forests and aquatic vegetation; although the largest percentage of the Reserve's territory corresponds to areas covered by estuaries and mangroves: red, black, white, jelf and red or crawling. In addition, there are 37 species of timber trees declared in the process of extinction, such as: balsam, cedar, guayacán, bobo berry, palo santo and all mangrove species. Its fauna houses 45 species of mammals, more than 300 species of birds, of which 27 are endemic. The most representative aquatic bird of the Reserve is the canclón, which has been widely studied in terms of its distribution, behavior and abundance; and that, in addition, it is threatened with extinction because it is an uncommon species in the Southwest, and rare in the eastern part of Ecuador [1].

Additionally, there is a great diversity of fish, molluscs and crustaceans, especially the red crab or guariche. Within the Reserve, artisanal fishermen, crab collectors and shrimp larvae coexist, which carry out a sustainable management of the resources, although there are certain identified problems associated with the extraction of these species by unauthorized outsiders [4, 5].

Among the tourist facilities of the Reserve are: camping area, picnic area and cabins to stay in the administrative area. Its main trails are: Laguna El Canclón, La Flora (low difficulty), El Mirador, Aulladores (medium difficulty), El Mate (medium difficulty), Route in the Churute river estuary; and, Route to the Ulpiano River. All require the assistance of a guide authorized to operate within the area [3] (Table 1).

**Table 1** Internal visitor activity preferences

Activity performed	Total visitors	Percentage
Practice sports	326,317	12.2
Observe flora and fauna	79,232	3.0
Visit, naturalize in protected areas	695,169	26.1
Visit communities	4,473	0.2
Visit shamans, healers	1,608	0.1
Visit archaeological, historical sites	77,126	2.9
Fun	1'214,990	45.6
Gastronomy	170.82	6.4
Make purchases	52,381	2.0
Others	44,137	1.7
Whole of the universe	2'666,315	100.0

Source:[6]

Coincidentally, the PIMTE 2014 reflects that at the national level the most commercialized tourist product is ecotourism and nature tourism, while in the international market this product is the second, constituting 21% of the offer [6]. As evidenced in the Strategic Plan for the Development of Sustainable Tourism for Ecuador “PLANDETUR 2020”, our country has a privileged position to develop tourism thanks to its mega-biodiversity housed in its protected areas [7].

This natural wealth contrasts with consumption and production systems that are not sustainable and that threaten the integrity of the ecosystems present in Ecuador. This reality presents an enormous challenge in terms of the need to introduce better practices, such as cleaner production, eco-efficiency and the application of more responsible behaviors with the environment [7].

The search for economic growth without considerations of an environmental nature has established production cycles that generate among other impacts: increasing contamination of water, soil and air that contributes to the alteration and degradation of ecosystems that could be prevented; the overexploitation of natural resources; loss of water and soil quality; and, damages associated with the introduction of exotic species. The opportunities linked to tourism need to be taken into account from a sustainability perspective, otherwise unique and fragile destinations such as the Galapagos Islands, the mangrove swamps of the coastal marine areas, the Cuyabeno lake systems or the Andean páramos, will not be able to withstand greater pressure [7].

On the other hand, in the technical report of the Project for the Conservation of Marine and Coastal Biodiversity of Ecuador, it was evidenced that the management plan of the Churute Manglares Ecological Reserve does not have an expiration date, despite the fact that it is mentioned as the last update. The year it was created (1998). At the legal level, the management plan is still in force, however, at the technical level of the Ministry of the Environment of Ecuador, it has been recommended to update it. Therefore, in 2015 its updating was planned and currently, according to

experts from the protected area, this stage is still being carried [8]. To date, since there is no updated management plan, it is not possible to have an effective visitor management system up to date that reflects the realities of the area and all its visitor sites.

Given that tourism is one of the main sources of development in Ecuador; protected areas are natural spaces that have the conditions to promote activities within the framework of ecotourism; and that there is still no updated management plan for the protected area in question, it is necessary to carry out a tourist carrying capacity study on one of its main trails.

Fourteen years later, considering 800 m of trail, 2133 visits per day were obtained as physical load capacity. Regarding the real carrying capacity, correction.

## 2 Methodology

The research developed in this project had a mixed approach: quantitative and qualitative, with the purpose of determining the calculation of tourist carrying capacity and designing a proposal that contributes to the management of the protected area. For the development of the research, three techniques were used: load capacity calculation methodology, survey and interview.

The carrying capacity calculation was carried out using the methodology described by Cifuentes, which seeks to establish the maximum number of visits that a protected area can receive based on the physical, biological and management conditions that occur in the area at the time of the study [9]. Queries about the technical specifications to be considered in the analysis were reviewed in conjunction with Atty. Rómulo Gaínza Torres, an expert who works in the area.

The survey was carried out on 330 tourists, 165 nationals and 165 foreigners, a rounded value from a representative sample of 323.83. For the selection of the sample size, the following expression corresponding to finite populations was used:

$$n = \frac{N * z_{\alpha}^2 * p * q}{d^2 * (N - 1) + z_{\alpha}^2 * p * q}$$

where  $n$  constitutes the sample size,  $N$  is the population size,  $z_{\alpha}$  is the percentile  $(1 - \alpha) * 100$  of the standard normal distribution corresponding to the confidence level,  $p$  represents the probability of success or occurrence of an event,  $q$  represents the probability of failure or not occurrence of an event;  $y, d$  is the maximum admissible error in terms of proportion.

Substituting in the expression, we obtained:

$$n = \frac{2,057 * 1.96^2 * 0.5 * 0.5}{0.05^2 * (2,057 - 1) + 1.96^2 * 0.5 * 0.5} = 323.83$$

Considering that the population  $N$  was made up of 2057 visitors during the year 2018 and a value  $z_{0.05}$  of 1.96 corresponding to the 95th percentile of the standard normal distribution (95% confidence) was considered; where  $p$  and  $q$  equal to ensure that the proportion of success or failure is the same, that is, 0.5. Additionally, a 5% admissible error was considered.

Through the survey, we sought to know about the potential visitation in the Reserve; however, this information gathering was complemented with the interview technique. For this purpose, 21 tour operators and travel agencies were interviewed, with the purpose of inquiring about supply and demand.

The interview was also used to examine the profile and perception of the current visitor to the La Flora trail. Between the months of August and September of the current year, 42 visitors were consulted, between nationals and foreigners.

The questionnaires used for the surveys and interviews, both in Spanish and English.

### 3 Analysis of Results

#### 3.1 Load Capacity Calculation

The calculation of tourist carrying capacity was carried out on the La Flora trail of the Churute Mangroves Ecological Reserve. Said trail has a welcome signage that indicates the travel time, distance and pictograms of permitted and not permitted activities as shown in image 1. This trail runs from Puerto La Flora to the pier, where you can see the forest of the mangrove, burrows of crabs and birds on the banks of the river. It is a low difficulty route where often, the visitor can coincide with community members of the area extracting crab from the mangrove swamp.

Although the trail has been classified as a natural rustic subzone, the surrounding infrastructure includes restrooms that are not 100% operational and a building that is not currently used. This building could become an environmental interpretation center or restaurant. Within the trail, the infrastructure of the shade house is in disuse; however, the bridge to the dock is in good condition for visitation. In addition, it does not have any interpretive signage on the species of flora and fauna, or on the ecosystem. In Fig. 1, a sub-zoning map of the trail is presented.

The associated calculations were based on the following assumptions:

- Visitor flow has been considered to be one-way on the trail.
- A person normally requires  $1 \text{ m}^2$  space to move freely. In the case of trails it is translated into 1 m linear, as long as the width of the trail is  $<2 \text{ m}$ .

### 3.2 *Physical Carrying Capacity (CCF)*

It was calculated using the following equation:

$$CCF = \frac{S}{sp} * \frac{h}{t}$$

where:

$S$  = available area in linear meters (300 m).

$sp$  = surface used per person (1 m).

$h$  = visiting hours (08:00 a.m. to 04:00 p.m., that is, 8 h/day).

$t$  = time needed to visit (30 min, that is, 0.5 h/visit/visitor).

Then:

$$CCF = \frac{300m}{1m} * \frac{8 \frac{h}{\text{día}}}{0.5 \frac{h}{\text{visita}}}$$

$$CCF = 4800 \frac{\text{visitas}}{\text{día}}$$

### 3.3 *Actual Carrying Capacity (CCR)*

The physical carrying capacity was subjected to a series of correction factors, specific to each site. The correction factors considered in this study were:

- Social Factor
- Precipitation Factor
- Sun Brightness Factor
- Fauna Disturbance Factor

Factors such as erodibility, accessibility, temporary closures and waterlogging were not considered in the analysis, since they are not applicable to the reality of the trail studied because the trail has a low slope and does not show traces of erosion or stagnation.

Additionally, the trail remains open 365 days a year, so it does not have temporary closures scheduled for certain days a week.

#### *Social Factor (FCsoc)*

Considering aspects related to the quality of visitation, the need to manage visits by groups was raised. For a better control of the flow of visitors and, at the same time,

to ensure their satisfaction, it was proposed that the visitation be managed under the following assumptions:

- Groups must be a maximum of 16 people.
- The distance between groups must be at least 50 m.
- Each group requires 66 m.

The number of groups ( $NG$ ) that can be simultaneously on the trail was calculated as follows:

$$NG = \frac{S}{dr}$$

$S$  = available area in linear meters (300 m).

$dr$  = distance required by each group (66 m).

So:

$$NG = \frac{300 \text{ m}}{66 \text{ m}} = 4.55 \text{ grupos}$$

After this, it was necessary to know how many people ( $P$ ) can be simultaneously on the trail. This was done through:

$$P = NG * pg$$

$NG$  = number of groups (grupos 4.55).

$pg$  = number of people for each group (16 people per group).

So:

$$P = 4.55 \text{ grupos} * 16 \frac{\text{personas}}{\text{grupo}} = 72.80 \text{ personas}$$

To calculate the social correction factor, it was necessary to identify the limiting magnitude which, in this case, was that portion of the trail that cannot be occupied because a minimum distance between groups must be maintained.

Therefore, since each person occupies 1 m of the path, the limiting magnitude ( $ml$ ) was equal to:

$$ml = S - P = 300 \text{ m} - 72.80 \text{ m} = 227.20 \text{ m}$$

Then:

$$FC_{soc} = 1 - \frac{ml}{mt}$$

$$ml = 227.20 \text{ m}$$

$$mt = 300m$$

Substituting we had:

$$FC_{soc} = 1 - \frac{227.20m}{300m} = 0.2427$$

#### *Precipitation Factor (FCpre)*

This factor prevents normal visitation, since the vast majority of visitors are not willing to hike in the rain. The months with the highest rainfall were considered (from January to April), in which the rain occurs more frequently in the morning hours [1, 10]. From this, it was determined that the limiting rain hours per day in this period are 2 h (from 08:00 to 10:00). Based on these considerations, the factor was calculated as follows:

$$FC_{pre} = 1 - \frac{hl}{ht}$$

where:

$hl$  = limiting rain hours per year.

$ht$  = hours per year that the protected area is open

$$hl = 119\text{días} * 2\frac{\text{horas}}{\text{día}} = 238\text{horas}$$

$$ht = 365\text{días} * 8\frac{\text{horas}}{\text{día}} = 2920\text{horas}$$

$$FC_{pre} = 1 - \frac{238\text{horas}}{2920\text{horas}} = 0.9185$$

#### *Solar Brightness Factor (FCsol)*

At certain times of the day, when the sun is very bright between 10:00 and 15:00, visits to places without coverage are difficult or uncomfortable. In the case of REMCH, this factor is limiting only in Sendero La Flora, where there is a stretch of 144 without coverage  $m$ . The months where the sun shines mostly begin in May and last until December.

During the eight months with little rain, the five limiting hours were taken into account and during the four rainy months the same five limiting hours in the morning were taken into account, since they are not affected because it regularly rains from 08:00 a.m. to 10:00 a.m. Therefore, the limiting sunshine hours were calculated as follows:

$$hsl = 365 \frac{\text{días}}{\text{año}} * 5 \frac{\text{horas}}{\text{día}} = 1,825 \frac{\text{horas}}{\text{año}}$$

Additionally, these calculations were applied only to uncovered tranches. Based on this, the factor was calculated as follows:

$$FC_{sol} = 1 - \left( \frac{hsl}{ht} * \frac{ms}{mt} \right)$$

where:

$hsl$  = limiting hours of sunshine/year (1825 horas).

$ht$  = hours per year that the protected area is open (2920 horas).

$ms$  = meters of path without coverage (144 m).

$mt$  = total meters of the path (300 m)

$$FC_{sol} = 1 - \left( \frac{1825 \text{horas}}{2920 \text{horas}} * \frac{144 \text{m}}{300 \text{m}} \right) = 0.70$$

#### *Fauna Disturbance Factor (FC<sub>fau</sub>)*

Since the trail is in the heart of the mangrove ecosystem, the red crab was selected as a representative species. Every year there are two important months for the red crabs of the Ecuadorian coast (*Ucides occidentalis*) called “vedas”.

The first date corresponds to the breeding ban that begins in February and ends in March. The second date corresponds to the growth ban where the species shed their shells, which begins on August 15 and ends on September 15. Accordingly, the factor was calculated as follows:

$$FC_{fau} = 1 - \frac{msl}{mst}$$

where:

$msl$  = Limiting months per year (2 meses).

$mst$  = total months in the year (12 meses)

$$FC_{fau} = 1 - \frac{2}{12} = 0.8333$$

From the application of the correction factors mentioned for the path, the actual carrying capacity was calculated, as shown below:

$$CCR = CCF * (FC_{soc} * FC_{pre} * FC_{sol} * FC_{fau})$$

$$CCR = 4800 \frac{\text{visitas}}{\text{día}} * (0.2427 * 0.9185 * 0.70 * 0.8333)$$



$$CCR = 623.93 \frac{\text{visitas}}{\text{día}}$$

### 3.4 Effective Carrying Capacity (ECC)

It is calculated considering the management capacity of the protected area. Said handling capacity was calculated based on three variables: personnel, infrastructure and equipment, as evidenced in Tables 2, 3, 4.

Each variable was evaluated with respect to four criteria: quantity, state; location and functionality. The personal category was only qualified taking into account the quantity criterion, because the knowledge and time for a personal evaluation were insufficient.

Therefore, the handling capacity is an average calculated from the three variables, as indicated:

$$CM = \frac{\text{Infr} + \text{Eq} + \text{Pers}}{3}$$

$$CM = \frac{0.491 + 0.382 + 0.667}{3}$$

$$CM = 0.513$$

In this way, the effective load capacity was obtained according to the following equation:

$$CCE = CCR * CM$$

where:

CCR = actual load capacity

CM = handling capacity

$$CCE = 623.93 \frac{\text{visitas}}{\text{día}} * 0.513$$

$$CCE = 320.08 \frac{\text{visitas}}{\text{día}}$$

From the effective carrying capacity calculated, it was possible to obtain the number of daily and annual visitors.

$$VD = CCE / \left( \frac{h}{t} \right)$$

**Table 2** REMCH management capacity: variable infrastructure

Infrastructure	Current amount (A)	Optimal amount (B)	Relationship (A/B)	Condition	Location	Functionality	Addition (S)	Factor (S/I6)
Administrative office	3	6	1	two	4	4	Eleven	0.688
House for staff	two	two	4	two	4	4	14	0.875
Conference room or exhibitions	1	two	1	two	4	two	9	0.563
Parking	two	two	4	two	4	two	12	0.750
Camping area	0	1	0	0	0	0	0	0,000
Picnic area	0	two	0	0	0	0	0	0,000
Garbage cans	3	5	two	1	3	two	8	0.500
	8	10	3	two	4	3	12	0.750
Toilets	1	two	1	two	3	3	9	0.563
Showers	1	two	1	two	3	3	9	0.563
Washbasin	10	twenty	1	3	3	3	10	0.625
Toilets	10	twenty	1	3	3	3	10	0.625
Urinals	3	6	1	3	3	3	10	0.625
Cellar	1	two	1	two	4	3	10	0.625
Walking trails	4	4	4	3	4	4	Fifteen	0.938
Lookout	1	1	4	two	4	4	14	0.875
Signaling	10	twenty	1	two	4	4	Eleven	0.688
Interpretive system	0	1	0	0	0	0	0	0,000
Sketch	0	1	0	0	0	0	0	0,000
Model	1	two	1	two	3	3	9	0.563

(continued)

Table 2 (continued)

Infrastructure	Current amount (A)	Optimal amount (B)	Relationship (A/B)	Condition	Location	Functionality	Addition (S)	Factor (S/16)
Rest areas	6	6	4	4	5	5	18	0,000
Docks	1	1	4	two	5	5	16	0,000
Average								

Prepared by: Authors

Table 3 REMCH handling capacity: variable equipment

Equipment	Current amount (A)	Optimal amount (B)	Relationship (A/B)	Condition	Location	Functionality	Addition (S)	Factor (S/16)
Vehicle	1	two	1	1	3	0	5	0,313
Radio	0	25	0	two	4	4	10	0,625
Fire extinguisher	5	10	1	two	3	3	9	0,563
First aid kit	1	10	0	1	3	1	5	0,313
Slide projector	0	two	0	0	0	0	0	0,000
Computer	two	10	0	1		two	3	0,188
Chainsaws	1	two	1	two	two	two	7	0,438
Surveillance Camera	0	40	0	0	0	0	0	0,000
GPS	4	4	4	4	4	4	16	1,000
Average								0,382

Prepared by: Authors

**Table 4** REMCH management capacity: personal variable

Staff	Current amount (A)	Optimal amount (B)	A/B ratio (C)	Factor (C/4)
Administrator	1	1	4	1,000
Environmental education	10	18	two	0.500
Park rangers	18	18	4	1,000
Technicians	two	3	two	0.500
Administrative staff	3	3	4	1,000
Guides	two	10	0	0,000
Average				0.667

Prepared by: Authors

$$VD = 320.08 \frac{\text{visitas}}{\text{día}} / \left( \frac{8 \frac{h}{\text{día}}}{0.5 \frac{h(\text{visitante})}{\text{visita}}} \right) = 20 \frac{\text{visitantes}}{\text{día}}$$
$$VA = VD * 365 \text{días/año}$$
$$VA = 20 \frac{\text{visitantes}}{\text{día}} * 365 \frac{\text{días}}{\text{año}} = 7300 \frac{\text{visitantes}}{\text{año}}$$

Based on the results, it was observed that the number of daily visitors is close to the maximum number of people that make up a group. This would imply that only a group of up to 16 people could be found on the trail; however, on certain occasions more than one group could coincide on the trail. This situation would put the sustainability of the area and its resources at risk, since if the limit of human activity is exceeded, the natural resource would deteriorate. Regarding the number of visitors per year, according to the visitor history, in 2018, there were 2,057 annual visitors. Therefore, it was evidenced that the tourist carrying capacity was not exceeded.

Possible solutions to improve handling capacity within the La Flora trail; and depending on the current and potential visitation to increase the tourist carrying capacity, there is the installation of informative signs, signage and signaling along the trail, installation of complementary infrastructure such as garbage cans and wooden bridges, installation of latrines or bathrooms at the entrance to the trail, identification and warning about risk or danger sites for visitors and design of interpretive trails, in such a way that all the points to be discussed on the walk are of interest, becoming a memorable experience about interpretation of natural or cultural history, bird and wildlife observation.

## 4 Conclusions

The number of 20 daily visitors obtained is close to the maximum number of people that make up a group, which would imply that only a group of up to 16 people could be found on the trail; Being able to count on certain occasions with more than one group on the trail, a situation that would jeopardize the sustainability of the area and its resources, since, if the limit of human activity is exceeded, the natural resource would deteriorate.


Regarding the number of visitors per year, according to the visitor history, in 2018, there were 2057 annual visitors, evidencing that the tourist carrying capacity of 7300 visitors per year has not been exceeded.

## References

1. Rivadeneira-Roura C, Rivera J (2007) Reserva Ecológica Manglares Churute. In: Ecolap MAE (ed) Guía del Patrim Áreas Nat Protegidas del Ecuador. DarwinNet, Quito
2. Sánchez A (2019) Propuesta de Actualización del Plan de Marketing Turístico para la Reserva Ecológica Manglares Churute. Universidad Católica de Santiago de Guayaquil
3. MAE (2015) Sistema Nacional de Áreas Protegidas. Reserva Ecológica Manglares Churute. Recuperado a partir de. <http://areasprotegidas.ambiente.gob.ec/es/areas-protegidas/reserva-ecologica-manglares-churute>
4. Navarrete R (2000) Atractivos Turísticos Naturales de la Reserva Ecológica Manglares Churute. CEDEGE, Ministerio del Ambiente, Fundación Natura, editores. Guayaquil
5. Navarrete R (2000) Capacidad de carga turística de los Sitios de Visita de la Reserva Ecológica Manglares Churute. CEDEGE, Ministerio del Ambiente, Fundación Natura, editores. Guayaquil
6. Ministerio de Turismo, Tourism and Leisure Advisory Services, Fondo de Promoción Turística (2009) Plan Integral de Marketing Turístico de Ecuador PIMTE 2014. Recuperado a partir de. <http://www.competencias.gob.ec/wp-content/uploads/2017/06/c.-2014.-PLAN-INT-EGRAL-MARKETING-TURISTICO.pdf>
7. Ministerio de Turismo (2007) Diseño del Plan Estratégico de Desarrollo de Turismo Sostenible para Ecuador PLANDETUR 2020. Recuperado a partir de. <https://www.turismo.gob.ec/wp-content/uploads/downloads/2013/02/PLANDETUR-2020.pdf>
8. Paguay J (2016) Evaluación de la Efectividad de Manejo de AMCP's del Ecuador Continental. Ecuador
9. Cifuentes M (1992) Determinación de Capacidad de Carga Turística en Áreas Protegidas. In: CATIE. Turrialba, Costa Rica
10. Climate-Data.Org (2020) Clima Naranjal. Climograma y Tabla climática. Recuperado a partir de. <https://es.climate-data.org/americadel-sur/ecuador/provincia-de-los-rios/naranjal-180158/>

# Convergent Fuzzy Cognitive Modelling of Regional Youth Policy Strategy



Aleksandr Raikov 

**Abstract** The article aims to accelerate strategic meetings in the country's regional governments. The fuzzy cognitive modelling and fuzzy cognitive maps (FCM), along with the author's convergent methodology, have helped to make this goal achievable. The convergent methodology suggests a unique approach for structuring information generated during the strategic meeting by applying the inverse problem-solving method to the fuzzy cognitive modelling to transform the divergent strategic discussion into a convergent one and make a strategic meeting sustainable and purposeful. The FCM helps to represent the non-formalisable cognitive semantics of computer models. The relevant big data analysis was used to justify fuzzy cognitive models and create them automatically. A high level of accuracy was achieved for verifying models; however, the accuracy of the latter process needed to be higher. The practical objective was to speed up the collective creation of a strategy for youth policy development for one of the country's regions. Three dozen participants, including remote experts and the youth policy department authorities, took part in the strategic meeting. It took five hours to collective building the draft of the strategy.

**Keywords** Artificial intelligence · Cognitive modelling · Fuzzy cognitive maps · Convergent methodology · Youth policy strategy

---

A. Raikov (✉)

National Supercomputer Centre in Jinan, Shandong 250103, China

e-mail: [aleksandr@jnist.cn](mailto:aleksandr@jnist.cn)

Institute of Control Sciences of Russian Academy of Sciences, Profsoyuznaya St. 65,  
Moscow 117997, Russia

MIREA – Russian Technological University, Vernadsky Avenue, 78, building 4, Moscow 119526,  
Russia

## 1 Introduction

There are many ways of strategic thinking and planning processes [1]. The collective strategic process involves teams for discussions of the current and future situation. As a result of the debate, participants should agree on strategic goals and courses of action. Non-formalisable and formalisable factors influence these processes. The former include chaotic and unforeseen circumstances as the participants' interests, thoughts, feelings, experiences, and intentions. The latter contains schemes, technologies, responsibilities, standards, instructions, material, financial, and information resources.

Many factors account for strategic decisions, and regional leaders should chip away at problems by creating effective strategies using artificial intelligence (AI) modelling. However, the possibilities of classical AI approaches, such as logic, metric spaces, and statistical instruments, are limited because they consider only formalisable factors.

The fuzzy cognitive modelling and fuzzy cognitive maps (FCM) can help to describe ill-defined social-economic situations [2] by using reconciliation approaches that exploit evolutionary (genetic) algorithms [3, 4], decision tree-based models [5], and fuzzy systems [6]. Furthermore, the author's convergent methodology for holding a strategic meeting using cognitive modelling and inverse problem-solving helps accelerate the collective strategic process [7].

The paper suggests the author's convergent approach to strategic regional youth policy development planning. Next, it reviews current research in the field. Finally, the implementation of the approach is demonstrated.

## 2 Analytical Review

The criteria for the analytical review, which accounts for collaborative strategic planning, are as follows:

- the goals are ill-defined and cannot be extrapolated from history by using statistical or neural network methods;
- an agreement between the team's participants must be achieved by helping cognitive modelling in a non-formalisable conceptual space;
- the cognitive semantics of computer models must be considered, which can be represented by FCM;
- the strategic decision-making process has an inverse character.

Cognitive modelling and the FCM method help to support collaborative strategic planning. The FCM method uses fuzzy logic and recurrent neural networks. It has been intensively discussed since the mid-1970s [8, 9]. There are some examples in which cognitive modelling and fuzzy methods have been applied: ecosystem building [10], the development of economic systems [11], strategic tourism development [12],



project risk management [13], security [14], classification [15], forecasting [16], industrial control [17]. The fields can intersect; e.g. a review [18] noted the deficit of studies devoted to using the FCM approach for classification.

The FCM model consists of concepts connected by causal relations with weights. Experts can estimate weights, or they can be automatically trained. For example, in [15], the FCM is represented by the graph of causal-directed relations, which indicates a negative or positive influence with continuous values. This graph can be represented by an influence matrix whose elements are the weights of edges. The author of [19] applied a formalised algorithm for automatic model learning, transforming a symbolic feature space step by step. Finally, the FCM with weights that keep good convergence properties during classification is introduced [20].

Many approaches and methods for training the weights of the edges between graph's nodes have been proposed, e.g. Hebbian learning [21], a balanced differential algorithm [22], evolutionary-based algorithms [23–25], and a multi-modal multi-agent learning genetic algorithm [26]. The paper [27] proposed an algorithm that generates models from big data. This algorithm uses a local search approach to reduce premature convergence. Intuitionistic FCM [28] can process imperfect facts and missing information. Granular Cognitive Maps [29] represent the arcs' weights using fuzzy sets, and so on.

A fuzzy multi-criteria decision-making approach helps to construct management systems in an uncertain environment by decision-makers with divergent interests and conflicting aims [30]. Project management was identified in the latter paper as the most popular field for the Fuzzy-Analytical Hierarchy Process approach [31, 32].

The causality of events, planning, and the creation of concepts cannot be effectively represented by traditional methods [33], e.g. by logical cognitive architectures [34] or a logical representation of AI tools [35] because they do not cover the full depth of the consciousness. For example, corporate social responsibility (CSR) is significant in strategic planning [36, 37]. CSR has a multifaceted impact, including improving decision-making in interregional affairs, generating positive social effects, increasing privacy protection, and improving the well-being of citizens.

In these ill-defined conditions, the collective strategic process must be convergent, i.e. it has to ensure a purposeful and stable progression to the goals. The spaces for decision-making can be conceptual or non-metric, for example, topological spaces [38–40]. The analytical instruments must be more vital to model human cognitive processes, thoughts, feelings, and consciousness.

For a long time, researchers have represented the semantics of AI models by using formalised contextual grammar [41]. Currently, the cognitive semantics of AI models can be taken into consideration by using the following components of the strategic meeting process with experts assessing a social-economic situation [42]:

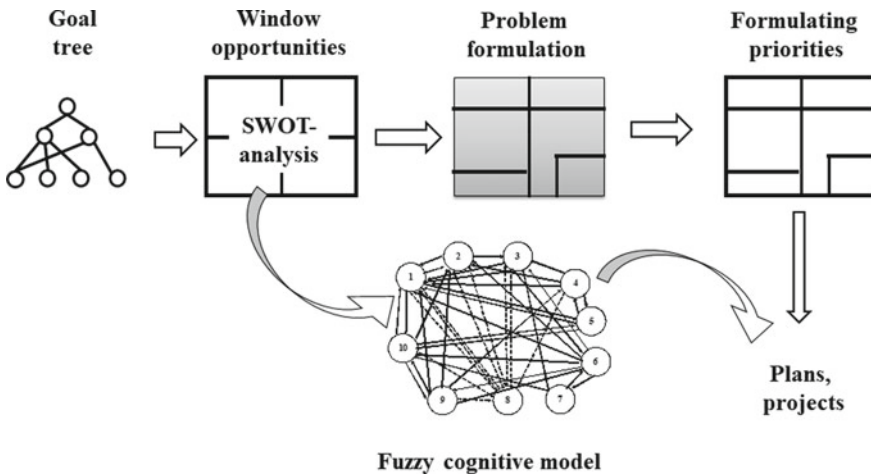
- regional leaders, experts, and representatives of civil society;
- big data, papers, patents, and Internet of thing information;
- mental space of participants (latent thoughts, motivations, emotions).

The convergence modelling approach [3], the ontologising operator [39], the big data analysis method for creating cognitive models [43], and the above-mentioned inverse problem-solving on topological spaces method—help to cover the non-formalisable mental space of participants and ensure the process of strategic conversation purposeful.

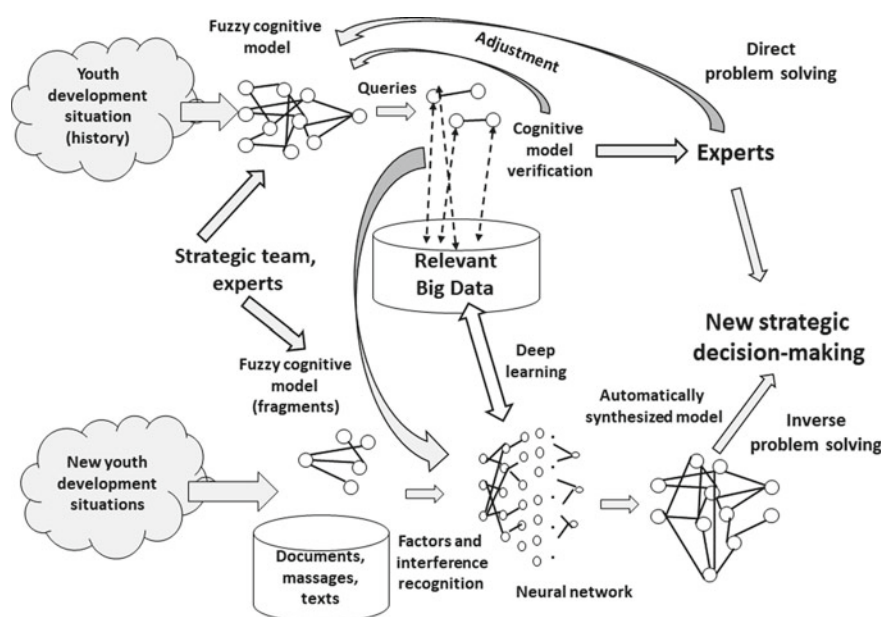
### 3 Collective and Automatic Fuzzy Cognitive Modelling

In the strategic meeting, which claims to take non-formalisable cognitive aspects of the situation into account, the ability to use traditional formalised AI models is limited. This is because the strategic process can have a chaotic and non-causality character. Cognitive modelling helps to accelerate the achievement of a group insight by representing the participants’ thoughts during the strategic meeting in a formalised way. Automating cognitive modelling can help in this process [43]. Strategic planning includes steps (Fig. 1): goal tree creation; strategic Strengths, Weaknesses, Opportunities, and Threats (SWOT) analyses; strategic problems and priorities formulating; fuzzy cognitive model creation; direct and inverse problem-solving.

The neural network (NN) and deep learning technologies can automatically construct the cognitive model (Fig. 2). However, the experimental assessments of the quality level of the automatic cognitive modelling process have yet to show high-level quality [43]. Several experiments have been conducted with recurrent and convolutional NNs. Encoding words using vectors and feeding them into NNs was done using Python, the open-source software library Keras, and the end-to-end open-source platform TensorFlow. The training data consisted of several thousand documents.



**Fig. 1** The organisation of the networked expertise



**Fig. 2** Automatic verification and creation of the cognitive model

The results were estimated using the metric of ‘accuracy’, i.e. the ratio of correctly classified documents to the total number of copies. The testing shows that the accuracy of the NN in creating a cognitive model’s factors is about 0.33. It is too small for the following reasons:

- the intersection of documents’ sets of words suited to factors;
- the imperfection of class recognition algorithms;
- the different semantics of economic sectors’ terminology, etc.

The idea for increasing the quality of this process may consist of creating cognitive semantics of AI models based on the quantum, relativistic and wave theories [44].

## 4 Collective Strategic Meeting and Modelling

The implication of the approach consisted of its use for strategic meeting support. It was held under the leadership of the regional youth policy department for drafting a regional youth policy strategy for three years. The department is the body of executive authority providing services concerning the regional youth policy and implementing projects to encourage young people’s healthy lifestyles, moral principles, and professional aspirations. Three dozen participants took part in this meeting, including remote experts and the regional authorities.

## 4.1 Goal Tree

According to the steps in Sect. 4, the strategic conversation started with creating and rating a goal tree. It can have more than three levels. The first level is the main goal (MG), which can be the mission or vision of youth policy development. The MG is understood as why and which the participants would like to see the situation in the future. The second level includes outside goals, and the third consists of inside goals, i.e. youth policy department development tasks. Every goal must obey one or more goals of the previous level. At each level, the goals must be ordered by importance.

The MG of the regional youth policy development should express young people's basic needs, interests, views, and dreams. Moreover, this goal should also fit into the country's national development goals. Therefore, as a vision for youth policy development, it was chosen: 'Constant improvement of the position of regional youth in civil society (Civic participation)'. Furthermore, the slogan was selected as the mission: 'Youth is a strategic region resource'.

The goals of the second level serve to realise the MG and reflect the directions of working in the external environment, among regional youth, the population, business, etc. External purposes have a good sound, a form of challenge. Within the framework of this work, the following external goals are defined:

- formation of a creative atmosphere among the youth for the growth of the competitiveness of the region;
- involvement of youth in social practice and creation of conditions for its self-organisation;
- development of infrastructure to ensure accessibility of services and awareness of young people;
- creation of an atmosphere of trust, promoting spiritual and cultural values and ethical standards.

The goals of the third level (internal tasks) ensure external goals. Internal tasks relate mainly to the youth policy management system; they were suggested as follows:

- development of a network of information and marketing centres;
- formation and promotion of a positive image of a young person;
- integration of the interaction of state, municipal and non-state resources;
- creation of a system for assessing the quality of life of young people;
- creation of conditions for the career growth of young talents through regional competitions, festivals, etc.

Experts, including leaders of youth organisations and the region's authorities, ordered the goals according to their importance (weight) using the well-known Fuzzy Analytic Hierarchy Process method. The author's computer tool helped to do it (Fig. 3) [42]. The pairwise goal comparison consistency level ('Experts' consistency') was automatically controlled by calculating an eigenvector of the goals' comparison matrices.

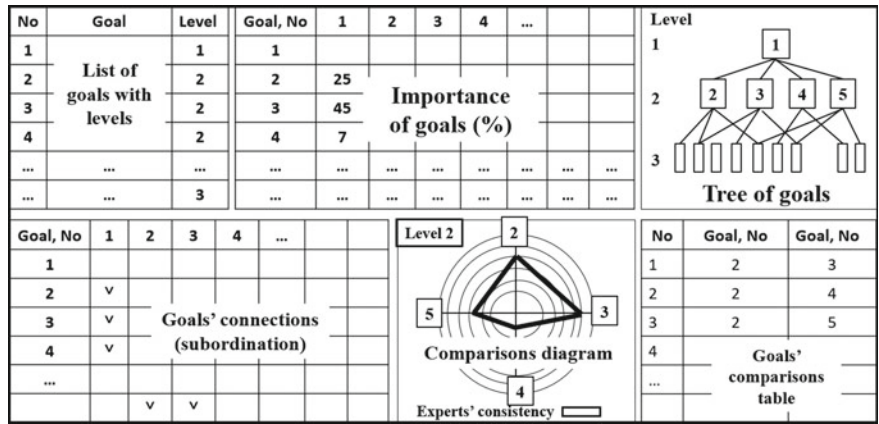


Fig. 3 Components (template) of the interface for creating the weighted goal tree

4.2 SWOT Analysis

The collective SWOT analysis was used to identify strategic factors in regional youth policy development and create strategic priorities (see Fig. 1). Fourteen external opportunities (Opportunities) for the development of regional youth policy were identified, for example, state support for youth initiatives, a favourable economic situation in the region.

Twelve factors were classified as favourable internal factors (Strengths). These are the factors that the meeting participants can directly influence themselves, for example, through the preparation of regulatory documents, the implementation of plans, decision-making, and the introduction of new management methods.

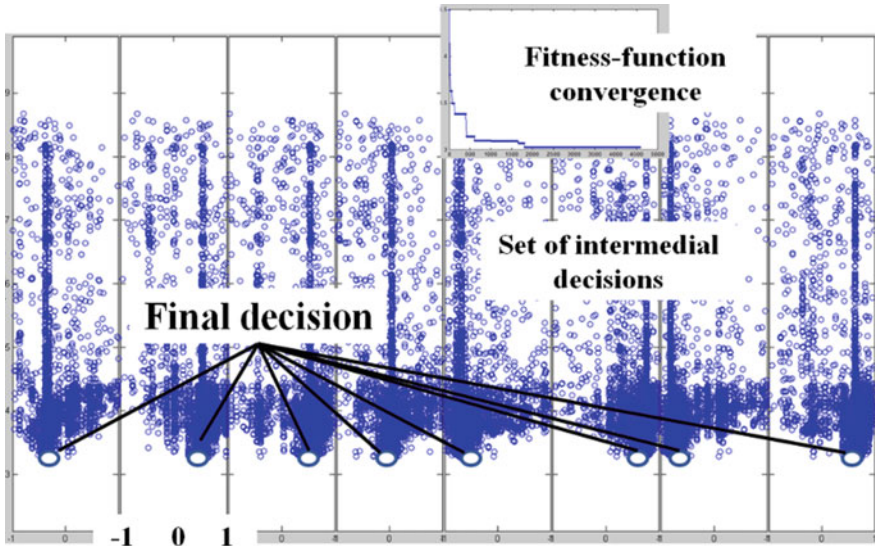
Eleven factors were classified as external threats (Threats) to the development of regional youth policy, including a small proportion of young people in the sphere of state and municipal administration, low political activity of young people, weak civil control and expertise, criminal activity of young people, etc.

Participants identified five main factors as internal weaknesses (Weaknesses), including weak cross-sectoral coordination of organisations working with youth; lack of a unified approach to forming programmes for implementing youth policy.

The selected factors became the basis for formulating strategic priorities. To do this, the participants of the strategic meeting jointly filled in the correlation matrix of the relationship of the factors.

4.3 Collective Fuzzy Cognitive Modelling

After the SWOT analysis, the ten most critical factors were bolded for creating the cognitive model. Then, experts assessed the strength of the mutual influence of these



**Fig. 4** The results of inverse problem-solving

factors by using the fuzzy scale  $(-1;1)$ . Cognitive modelling helps to answer such questions as ‘What decisions must be made to achieve the goals?’ (Fig. 4).

Collective strategic analysis and cognitive modelling showed the feasibility of identifying the following three priority areas of activity:

- ‘Active life position of youth: civil, social, and innovative’.
- ‘The New Business Youth: Professionalism and Entrepreneurship’.
- ‘Healthy generation: spiritually, morally, and physically’.

Projects (strategic measures) were selected within the above priorities. For example, implementing the ‘Youth Civil Expertise’ initiative can be a project.

## 5 Discussion and Conclusion

The most discussed issue in collective fuzzy cognitive modelling is considering the non-formalisable cognitive semantics of AI models. The model’s factors can be verified by automating mapping them on relevant big data. As a result, it gives a high level of adequacy to created cognitive models and accelerates the collective strategic process.

However, the accuracy of automatic finding out of cognitive models’ factors did not exceed 33%. Therefore, a low accuracy level needs to be studied in more detail. For example, creating the cognitive semantics of AI models based on the quantum, relativistic, and wave theories described in [44] can be the way to raise this accuracy.

The inverse problem-solving method immerses the strategic conversation's participants in fuzzy cognitive modelling process. It helps to consider the non-formalisable (cognitive, chaotic, and uncaused) semantics, accelerate the strategic meeting, and make it more convergent, purposeful, and sustainable.

**Acknowledgements** The Russian Science Foundation supports this work, Grant No 21-18-00184.

## References

1. Krogerus M, Tschäppeler R (2012) *The decision book: fifty models for strategic thinking*. W. Norton & Company, New York
2. Cherenkov IV, Feyzov VR (2021) The study of the possibilities of constructing cognitive models of complex systems as a result of the analysis of time series of a limited number of factors on the example of financial markets. *IFAC-PapersOnLine*. 54(13):166–171. <https://doi.org/10.1016/j.ifacol.2021.10.439>
3. Raikov AN, Panfilov SA (2013) Convergent decision support system with genetic algorithms and cognitive simulation. In: *Proceedings of the IFAC conference on manufacturing modelling, management and control, MIM'2013*, pp 1142–1147. <https://doi.org/10.3182/20130619-3-RU-3018.00404>
4. Przewozniczek MW, Komarnicki MM (2021) Empirical problem decomposition: the key to the evolutionary effectiveness in solving a large-scale non-binary discrete real-world problem. *Appl Soft Comput* 113:107864. <https://doi.org/10.1016/j.asoc.2021.107864>
5. Spiliotis E, Abolghasemi M, Hyndman RJ, Petropoulos F, Assimakopoulos V (2021) Hierarchical forecast reconciliation with machine learning. *Appl Soft Comput* 112:107756. <https://doi.org/10.1016/j.asoc.2021.107756>
6. Laengle S, Lobos V, Merigó JM, Herrera-Viedma E, Cobo MJ, De Baets B (2021) Forty years of fuzzy sets and systems: a bibliometric analysis. *Fuzzy Sets Syst* 402:155–183. <https://doi.org/10.1016/j.fss.2020.03.012>
7. Raikov AN (2008) Convergent cognotype for speeding-up the strategic conversation. *Proc World Cong Int Feder Autom Control (IFAC)* 41(2):8103–8108. <https://doi.org/10.3182/20080706-5-KR-1001.01368>
8. Axelrod RM (1976) *Structure of decision: the cognitive maps of political elites*, the structure of decision the cognitive maps of political elites. Princeton University Press, Princeton, p 404
9. Kosko B (1986) Fuzzy cognitive maps. *Int J Mach Stud* 24:65–75
10. Ozesmi S, Ozesmi U (2004) Ecological models based on people's knowledge: a multi-step fuzzy cognitive mapping approach. *Ecol Model* 176(1–2):43–64
11. Jetter A, Schweinfurt W (2011) Building scenarios with fuzzy cognitive maps: an exploratory study of solar energy. *Futures* 43(1):52–66
12. Raikov A (2020) Megapolis tourism development strategic planning with cognitive modelling support. In: Yang XS, Sherratt S, Dey N, Joshi A (eds) *Proceedings of the 4th international congress on information and communication technology. Advances in intelligent systems and computing*, vol 1041. Springer, Singapore. [https://doi.org/10.1007/978-981-15-0637-6\\_12](https://doi.org/10.1007/978-981-15-0637-6_12)
13. Lazzerini B, Mkrtchyan L (2011) Analyzing risk impact factors using extended fuzzy cognitive maps. *IEEE Syst J* 5(2):1480
14. Ryjov AP, Mikhalevich IF (2021) Hybrid intelligence framework for improvement of information security of critical infrastructures. In: Cruz-Cunha MM, Mateus-Coelho NR (eds) *Handbook of research on cyber crime and information privacy*, Hershey, PA, US, pp 310–337
15. Szwed P (2021) Classification and feature transformation with fuzzy cognitive maps. *Appl Soft Comput* 105:107271. <https://doi.org/10.1016/j.asoc.2021.107271>



16. Yang S, Liu J (2018) Time-series forecasting based on high-order fuzzy cognitive maps and wavelet transform. *IEEE Trans Fuzzy Syst* 26(6):3391–3402
17. de Souza LB, Soares PP, Mendonza M, Mourhir A, Papageorgiou EI (2018) Fuzzy cognitive maps and fuzzy logic applied in industrial processes control. In: *Proceedings of the 2018 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, IEEE, pp 1–8
18. Nápoles G, Espinosa ML, Grau I, Vanhoof K, Bello R (2017) Fuzzy cognitive maps-based models for pattern classification: advances and challenges. *Soft computing based optimization and decision models*, vol 360. Springer, Berlin, pp 83–98
19. Bueno S, Salmeron J (2009) Benchmarking main activation functions in fuzzy cognitive maps. *Exp Syst Appl* 36(3):5221–5229
20. Karatzinis GD, Boutalis YS (2021) Fuzzy cognitive networks with functional weights for time series and pattern recognition applications. *Appl Soft Comput* 106:107415. <https://doi.org/10.1016/j.asoc.2021.107415>
21. Dickerson JA, Kosko B (1994) Virtual worlds as fuzzy cognitive maps. *Presence Teleoper Virt Environ* 3(2):173–189
22. Hueriga AV (2002) A balanced differential learning algorithm in fuzzy cognitive maps. In: *Proceedings of the 16th international workshop on qualitative reasoning*
23. Stach W, Kurgan L, Pedrycz W, Reformat M (2005) Genetic learning of fuzzy cognitive maps. *Fuzzy Sets Syst* 153(3):371–401
24. Chi Y, Liu J (2015) Learning of fuzzy cognitive maps with varying densities using a multiobjective evolutionary algorithm. *IEEE Trans Fuzzy Syst* 24(1):71–81
25. Parsopoulos KE, Papageorgiou EI, Groumpos P, Vrahatis MN (2003) A first study of fuzzy cognitive maps learning using particle swarm optimization. In: *Proceedings of the 2003 congress on evolutionary computation. CEC'03*, IEEE, vol 2, pp 1440–1447
26. Yang Z, Liu J (2019) Learning of fuzzy cognitive maps using a niching-based multi-modal multi-agent genetic algorithm. *Appl Soft Comput* 74:356–367
27. Acampora G, Pedrycz W, Vitiello A (2015) A competent memetic algorithm for learning fuzzy cognitive maps. *IEEE Trans Fuzzy Syst* 23(6):2397–2411
28. Papageorgiou EI, Iakovidis DK (2012) Intuitionistic fuzzy cognitive maps. *IEEE Trans Fuzzy Syst* 21(2):342–354
29. Pedrycz W, Homenda W (2013) From fuzzy cognitive maps to granular cognitive maps. *IEEE Trans Fuzzy Syst* 22(4):859–869
30. Chen L, Pan W (2021) Review fuzzy multi-criteria decision-making in construction management using a network approach. *Appl Soft Comput* 102:107103. <https://doi.org/10.1016/j.asoc.2021.107103>
31. Lee S (2014) Determination of priority weights under multi-attribute decision-making situations: AHP versus fuzzy AHP. *J Constr Eng Manag* 141(2):897. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0000897](https://doi.org/10.1061/(ASCE)CO.1943-7862.0000897)
32. Jaskowski P, Biruk S, Bucon R (2010) Assessing contractor selection criteria weights with fuzzy AHP method application in group decision environment. *Autom Constr* 19(2):120–126. <https://doi.org/10.1016/j.autcon.2009.12.014>
33. LeCun Y (2017) Power and limits of deep learning. <https://www.youtube.com/watch?v=0tEhw5t6rhc>. Accessed 15 July 2022
34. Kotseruba I, Tsotsos JK (2020) 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artif Intell Rev* 53(1):7–94. <https://doi.org/10.1007/s10462-018-9646-y>
35. Wang P (2019) On defining artificial intelligence. *J Artif Gen Intell Philadelphia USA Temple Univ* 10(2):1–37. <https://doi.org/10.2478/jagi-2019-0002>
36. Du S, Xie C (2020) Paradoxes of artificial intelligence in consumer markets: ethical challenges and opportunities. *J Bus Res* 129:961–974. <https://doi.org/10.1016/j.jbusres.2020.08.024>
37. Perko I (2020) Hybrid reality development-can social responsibility concepts provide guidance? *Kybernetes* 50:676–693. <https://doi.org/10.1108/K-01-2020-0061>
38. Ivanov VK (1969) Incorrect problems in topological spaces. *Siberian Math J* 10(5):785–791. <https://doi.org/10.1007/BF00971654>



39. Raikov A (2021) Convergent ontologization of collective scientific discoveries. In: Proceedings of the 14th international conference management of large-scale system development (MLSD), pp 1–5. <https://doi.org/10.1109/MLSD52249.2021.9600184>
40. Raikov AN (2020) Accelerating decision-making in transport emergency with artificial intelligence. *Adv Sci Technol Eng Syst J* 5(6):520–530. <https://doi.org/10.25046/aj050662>
41. Rumelhart D (1975) Notes on a schema of stories. In: Bobrov DC, Collins A, et al (eds) *Representation and understanding*. Academic Press, New York
42. Gubanov D, Korgin N, Novikov D, Raikov A (2014) E-expertise: modern collective intelligence, vol 558. Springer, Series: Studies in Computational Intelligence, New York, p XVIII. <https://doi.org/10.1007/978-3-319-06770-4>
43. Raikov A, Ermakov A, Merkulov A, Panfilov S (2023) Automatic synthesis of cognitive model for revealing economic sectors' needs in digital technologies. In: Yang XS, Sherratt S, Dey N, Joshi A (eds) *Proceedings of seventh international congress on information and communication technology. Lecture notes in networks and systems*, vol 448. Springer, Singapore. [https://doi.org/10.1007/978-981-19-1610-6\\_20](https://doi.org/10.1007/978-981-19-1610-6_20)
44. Raikov A (2021) Cognitive semantics of artificial intelligence: a new perspective. Springer Singapore, Topics: Computational Intelligence, Singapore, p XVII. <https://doi.org/10.1007/978-981-33-6750-0>

# Realistic Modeling of Computer Systems in Gem5 Simulator



Amit Mankodi

**Abstract** Researchers have built computer systems in system-level simulators for decades to model real-world systems. Simulated systems developed in the simulator are utilized for architectural research to understand the effect of system feature changes on the application's performance. Generally, system-level simulators use processor, cache and memory features to simulate a system. Therefore, it is crucial to construct computer systems in a simulator that matches the features of real computer systems. One approach is to build simulated systems using a full-system simulator (FSS) that provides performance with higher accuracy but constructing FSS is complex and slow. The Gem5 simulator offers an alternative approach that speeds up the construction of systems by emulating a system using the system-Call emulation mode. However, there are several challenges in representing real-world systems in Gem5. In this paper, we identify those challenges and provide solutions by modifying the code of the open-source Gem5 simulator.

**Keywords** Gem5 simulator · System Modeling · System-call emulation

## 1 Introduction

Architectural research generally requires a large number of computer systems for evaluation purposes. Either acquiring systems is not cost-effective or a system with needed hardware features is not designed yet. Therefore, researchers use system-level simulators such as Gem5, PTLSim, SimpleScalar, BookSim, and MultiSim to simulate systems required for research. Full-system simulator PTLSim is a cycle-accurate simulator with the support of only x86 instruction set according to [1]. SimpleScalar simulates only processors with various instruction sets and designs used by [2] and [3]. The BookSim and MultiSim simulators simulate network on-

---

A. Mankodi (✉)

Dhirubhai Ambani Institute of Information and Communication Technology, Near Indroda Circle, Gandhinagar, Gujarat 382007, India

e-mail: [amit\\_mankodi@daiict.ac.in](mailto:amit_mankodi@daiict.ac.in)

chip used by [4] and [5]. On the other hand, we require a simulator that simulates out-of-order processors with multiple instruction sets (x86, ARM, etc.), multi-level caches, and various memory modules.

Gem5 simulator [6] provides the required functionality to build systems with different instruction sets, cache levels and memory modules. We can build full systems in Gem5; however, building a large number of full systems in Gem5 is time-taking and complex, according to [7]. Gem5 provides a solution in the form of emulating systems using system-call emulation mode instead of building full systems with some compromise on accuracy. While emulating a large number of systems in Gem5, we found that the current implementation of Gem5 does not support three levels of cache and some memory modules used in real-world systems today. In this paper, we address these challenges and provide solutions by modifying open-source Gem5 source code, which is our contribution.

This paper is organized as follows: Sect. 2 provides details of the real-world computer systems selected to be built in the Gem5 simulator. Section 3 provides reasoning for the selection of Gem5 simulator. Section 4 lists out the challenges faced during the construction of systems in Gem5 and the implemented solution. Section 5 provides details of the experiment performed and its results. Section 6 provides concluding remarks.

## 2 Selection of Computer Systems

The simulated systems must represent the general population of physical computer systems available in the market today. Furthermore, simulated systems need to be characterized by their processor, cache and memory features. Therefore, we surveyed a wide range of commercial computer systems and categorized them into different hardware classes represented by the “H/W class” column in Table 1. The footnote shows the H/W class values and their class description. For systems of each H/W class, we collected values of nine features; three processor features, CPU speed, instruction-set-architecture (ISA), cores, three-level cache sizes, and three memory features, type, access speed and size. For example, class 1 systems were built using systems features of AMD Ryzen and EPYC, class 5 systems were built using Apple systems features, class 6 systems were built based on Intel Core and so on. Table 1 shows feature values we used to build gem5 simulated systems.

## 3 Why to use Gem5 Simulator?

Several research works have utilized various simulators to simulate real systems. For example, [2] and [3] uses SimpleScalar, [5] uses MultiSim, [4] uses BookSim. However, due to the advantages of gem5 many recent researchers [1, 8–14] have utilized gem5 simulator. We primarily use the gem5 simulator due to its following advantages:

**Table 1** Computer systems built in Gem5 simulator

H/W class	ISA	CPU speed GHz	Cores	Mem type	Mem access MHz	L1-L3 cache size	Cnt
1	x86	2–3.5	2–18	DDR4	2400–2666	32kB-64MB	50
2	x86	2.8–4.7	1–8	DDR3	1600–1866	16kB-8MB	60
3	ARM	1.7–2	4,8	DDR4	1866	32kb-8MB	15
4	ARM	1–2.7	2–8	LPDDR2	400–1866	4kB-3MB	70
5	ARM	1.1–2.34	1–4	LPDDR3	1600–1866	32kB-4MB	35
6	x86	1.3–3.5	2,4	DDR3	1600	32kB–8MB	60
7	x86	1.7–3.5	2–18	DDR4	1866–266	32kB–16MB	95
8	x86	1.3–3.5	2,4	LPDDR3	1600–2133	32kB–8MB	90

\*Memory Size range 1–8GB

H/W Class and its associated class of systems: (1) AMD Ryzen and Epyc. (2) AMD Bulldozer and Piledriver. (3) AMD Opteron (4) Qualcomm Snapdragon. (5) Apple. (6) Intel Core i7, i5 and i3 with DDR3 DRAM. (7) Intel Core i9, i7, i5 and i3 with DDR4 DRAM. (8) Intel Core i7, i5 and i3 with LPDDR3 DRAM

- The gem5 simulator has the system-call emulation mode to simulate systems rapidly. Therefore, we use system-call emulation mode instead of full-system mode to build simulation-based systems in the gem5 simulator.
- The gem5 simulator supports processors with six different instruction-set-architectures (ISAs), including ARM and x86.
- The gem5 simulator supports various processor types, including the out-of-order (OoO) widely used in real computer systems.
- The gem5 simulator can be used in conjunction with McPAT to collect power consumption for applications executed on Gem5 simulated systems.

#### 4 Challenges During Construction of Simulated Systems in Gem5

Due to the support of several instruction set architectures (ISAs) in gem5, we built 120 ARM-based systems and 355 x86-based systems in gem5, with a total of 475 systems as shown by the “cnt” column of Table 1. However, we faced several challenges while using other hardware features, such as types of memory, three levels of cache and so on, for building systems in the gem5 simulator. The subsections below discuss

**Table 2** Supported memory modules in Gem5 simulator

Mem type	Mem access MHz	Based on datasheet	Added
DDR3	1600, 2100	Micron MT41J512M8	No
DDR4	2400	Micron MT40A512M16	No
LPDDR2	1066	Micron MT42L128M32D1	No
LPDDR3	1600	Micron EDF8132A1MC	No
GDDR5	4000	SK Hynix H5GQ1H24AFR	No
HMC	2500		No
WideIO	200		No
HMB	1000		No
DDR3	1066, 1333, 1866	Micron MT41J512M8	Yes
DDR4	1866, 2133, 2666	Micron MT40A512M16	Yes
LPDDR2	933, 800, 667, 533, 400, 333	Micron MT42L128M32D1	Yes
LPDDR3	1866	Micron EDFA164A1PB	Yes
LPDDR3	2133	Micron EDFA232A1MA	Yes

each challenge and the solution we have applied to resolve it. We also discuss the limitations of gem5 simulated systems compared to the physical systems.

**4.1 Challenge 1: Gem5 Memory Support**

The gem5 simulator supports several memory modules with a diversity of memory types and access speeds indicated in Table 2 with Added column value as No. However, the challenge is that the gem5 simulator does not support all the required memory modules with different memory types and access speeds to build simulation systems in the gem5 simulator that represent the surveyed commercial computer systems. To overcome this challenge, we reviewed the gem5 source code available with open access and made changes to add all the required memory modules. Table 2 with the Added column as Yes shows all the memory modules with respective memory types and access speeds added in the gem5 simulator.

The code below shows an example of the changes made to DRAMCtrl.py python code in gem5 source to add two of the memory modules. It is evident from the code example that we needed to use specification values from the datasheet of memory provided by the manufacturer to add memory modules in gem5. The “Based On Datasheet” column in Table 2 indicates the datasheets we have used to collect the specification values.

```

577 class DDR3_1866_x64(DDR3_1600_x64):
578     # 933 MHz
579     tCK = '1.07ns'
580
581     # 8 beats across an x64 interface translates to 4 clocks @ 933 MHz
582     tBURST = '4.28ns'
583
584     # DDR3-1866 13-13-13
585     tRCD = '13.91ns'
586     tCL = '13.91ns'
587     tRP = '13.91ns'
588     tRAS = '34ns'
589     tRRD = '5ns'
590     tXAW = '27ns'
591
592     # Current values from datasheet Die Rev E,J
593     IDD0 = '62mA'
594     IDD2N = '35mA'
595     IDD3N = '41mA'
596     IDD4W = '141mA'
597     IDD4R = '174mA'
598     IDD5 = '242mA'
599     IDD3P1 = '41mA'
600     IDD2P1 = '37mA'
601     IDD6 = '20mA'
602     VDD = '1.5V'

```

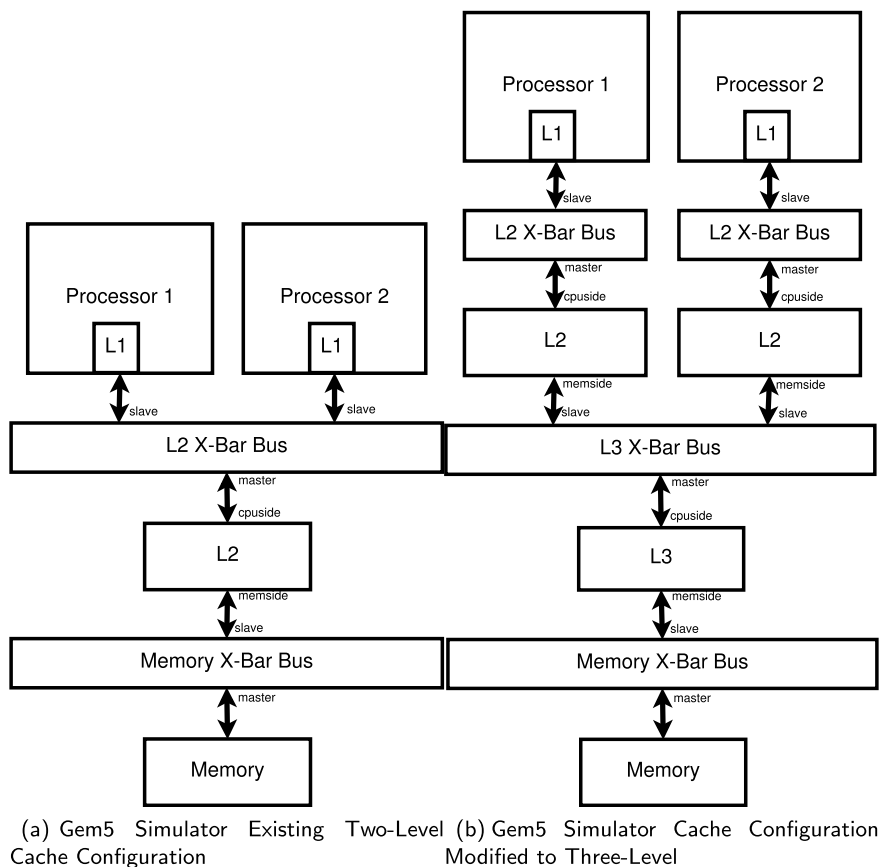
```

949 class LPDDR2_933_x32(LPDDR2_S4_1066_x32):
950     # 466.5 MHz
951     tCK = '2.144ns'
952
953     # Irrespective of speed grade, tWIR is 7.5 ns
954     tWIR = '7.5ns'
955
956     # Default same rank rd-to-wr bus turnaround to 2 CK, @466.5 MHz = 2.144 ns
957     tRTW = '4.288ns'
958
959     # Default different rank bus delay to 2 CK, @466.5 MHz = 2.144 ns
960     tCS = '4.288ns'
961
962     # Current values from datasheet
963     IDD4W2 = '185mA' # 190 - 5
964     IDD4R2 = '194mA' # 220 - 194 = 26

```

## 4.2 Challenge 2: Gem5 Cache Support

The gem5 simulator supports a two-level cache hierarchy for systems building, as shown in Fig. 1a. In contrast, commonly used commercial systems today have a three-level cache, a second challenge we faced in building simulated systems in



**Fig. 1** Gem5 simulator system cache configuration

gem5. To cope with this challenge, we analyzed the source code that implemented the cache structure in gem5 and identified two problems that require resolution.

First, the current implementation of gem5 supports only two levels of cache (L1 and L2); hence, we made code changes to add the third level cache (L3). The new L3 cache, a last-level cache in the new cache configuration, will be shared between all the processors, making it a system-level cache. However, in the current implementation, the L2 cache is a last-level cache; therefore, L2 is a system-level cache shared by all processors, which was a second problem. Hence, we added an option to have the L2 cache shared by all processors to make it a system-level cache or individual L2 cache for each processor. Figure 1b shows the three-level cache configuration that we implemented in gem5.

To implement the three-level cache configuration, we first needed to support system-level (shared) L2 cache and individual processor (non-shared) L2 cache modes to ensure that backward compatibility with the two-level cache configuration. We used an option `l2cache_sharedbycpu` with true or false values depending on

whether the two-level configuration required a shared L2 cache or a non-shared L2 cache in a three-level configuration. The code listing below articulates the changes made to `CacheConfig.py` to check for the `l2cache_sharedbycpu` option and create either system-level shared L2 cache or non-shared L2 cache for each processor.

```

114 elif options.l2cache:
115 # AKM— Added Shared L2 Cache
116 # if l2cache_sharedbycpu true then l2 cache is as cpu level and not system
    wide
117 if not options.l2cache_sharedbycpu:
118 # AKM— Added Shared L2 Cache
119 # Provide a clock for the L2 and the L1-to-L2 bus here as they
120 # are not connected using addTwoLevelCacheHierarchy. Use the
121 # same clock as the CPUs.
122 system.l2 = l2_cache_class(clk_domain=system.cpu_clk_domain,
123                             size=options.l2_size,
124                             assoc=options.l2_assoc)
125
126 system.tol2bus = L2XBar(clk_domain = system.cpu_clk_domain)
127 system.l2.cpu_side = system.tol2bus.master
128 system.l2.mem_side = system.membus.slave

164 # AKM— Added Shared L2 Cache
165 # if l2cache_sharedbycpu true then l2 cache is as cpu level and not system wide
166 if options.l2cache_sharedbycpu:
167     l2 = L2Cache(clk_domain=system.cpu_clk_domain,
168                  size=options.l2_size,
169                  assoc=options.l2_assoc)
170
171     l2.writeback_clean = True
172     system.cpu[i].addTwoLevelCacheHierarchy(icache, dcache, l2,
173                                              iwalkcache, dwalkcache)
174 else:
175 # AKM— Added Shared L2 Cache

```

To add L3 cache to complete the three-level cache configuration, we first added an option for L3 cache in `Options.py`.

```

101 # AKM— Added L3 Cache
102 parser.add_option("--l2cache-sharedbycpu", action="store_true") # —l2cache
    option must be true to use this option
103 parser.add_option("--l3cache", action="store_true")
104 # AKM— Added L3 Cache

```

We then added L3 cache in `Caches.py` and cross bar switch (Xbar) in `XBar.py` that will connect L3 cache to the memory.

```

79 # AKM— Added L3 Cache
80 class L3Cache(Cache):
81     assoc = 12
82     tag_latency = 36
83     data_latency = 36
84     response_latency = 36
85     mshrs = 36

```





```
108         system.l2.writeback_clean = True
109         system.tol2bus = L2XBar(clk_domain = system.cpu_clk_domain)
110         system.l2.cpu_side = system.tol2bus.master
111         system.l2.mem_side = system.tol3bus.slave
112     elif options.l2cache:
113 #         if options.l2cache:
114 # AKM- Added L3 Cache
```

4.3 Challenge 3: Gem5 Processor Support

The gem5 simulator supports an out-of-order (OoO) processor; however, its implementation is based on five-stage pipeline Alpha 21264 processor [15] with features shown in Table 3. In contrast, processors used in today’s computer systems have pipelines with varying stages with advanced features such as vector processors. It is an unimaginable task to make the gem5 OoO processor support all the features of several processors of today, which is a considerable challenge. Therefore, we have constructed all the gem5 simulation-based systems using the only available OoO processor, which is a limitation of our approach.

Table 3 Gem5 simulator O3CPU model features

Feature	Value
Pipeline stages	5 Fetch,Decode,Rename, Issue/Execute/Writeback, Commit
Branch predictor	Tournament (used)
Number of reorder buffer	1
Number of reorder buffer entries	192
Number of load queue entries	32
Number of store queue entries	32
Number of physical integer registers	256
Number of physical float registers	256
Functional Units	IntALU-6,IntMultDiv-3 FPALU-4,FPMultDiv-2 RdWrPort - 4

## 5 Experimental Details

As per the detail in Sect. 4, we first modified the source code of the Gem5 simulator to implement solutions to resolve challenges. Utilizing the modified version of the Gem5 simulator source code, we built 475 computer systems considering hardware features of the real-world systems listed in Table 1. We executed each application listed in Sect. 5.1 on each of the 475 Gem5 simulated systems. We collected hardware feature values from Gem5 simulated systems and runtimes from Gem5 execution logs for each execution. We compared the data from simulated systems against the data from the physical systems.

### 5.1 Applications Selection As Workloads

The categorization of applications according to their compute and data access patterns are shown in [16, 17], categorizing them as compute-bound, memory-bound or compute-plus-memory-bound. We have selected three applications having different compute and data access patterns. A memory-bound application matrix multiplication from linear algebra, a compute-bound application to calculate the value of PI using monte carlo and quicksort, a compute-plus-memory-bound application. Each application was written in C language using known implementation or taken from standard benchmarks such as MiBench [18]. We generated executable binaries of all applications on the same host computer with Linux 16.04 LTS operating system, Intel Core i5-6500 3.2 GHz processor, DDR3 1600 MHz memory. We utilized the GCC compiler to generate binaries for x86 systems and the GCC cross compiler to generate binaries for ARM-based systems. The purpose of using the same host systems is to eliminate the effect of different software environments such as operating systems, compiler optimization levels.

We have considered several problem sizes for each application. For example, we use matrix sizes of 50–200 in increments of 50 for matrix multiplication, loop iterations of 100000, 500000 and 1000000 to calculate PI for monte carlo, and quicksort sorts 1000–6000 words in an increment of 1000. The execution of these applications on all 475 Gem5 simulated systems on the same host computer took seven days, six days and one and half days continuous runs for matrix multiplication, monte carlo and quicksort.

### 5.2 Results and Observations from Gem5 Simulated Systems

To analyze the results from the execution of three applications with various problem sizes on 475 Gem5 systems, we plotted hardware feature versus runtime in Figs. 2, 3 and 4. We have the following observations from the plots:

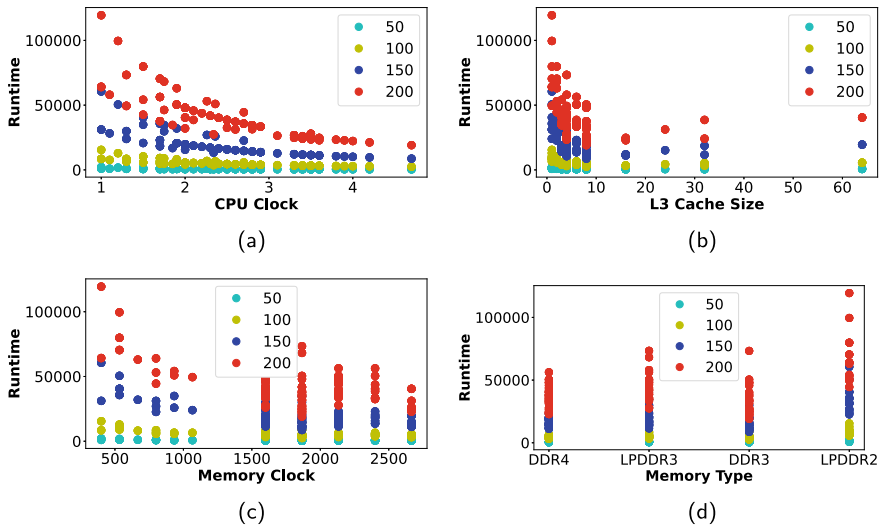


Fig. 2 Gem5 simulated systems - matrix multiplication hardware features versus runtime

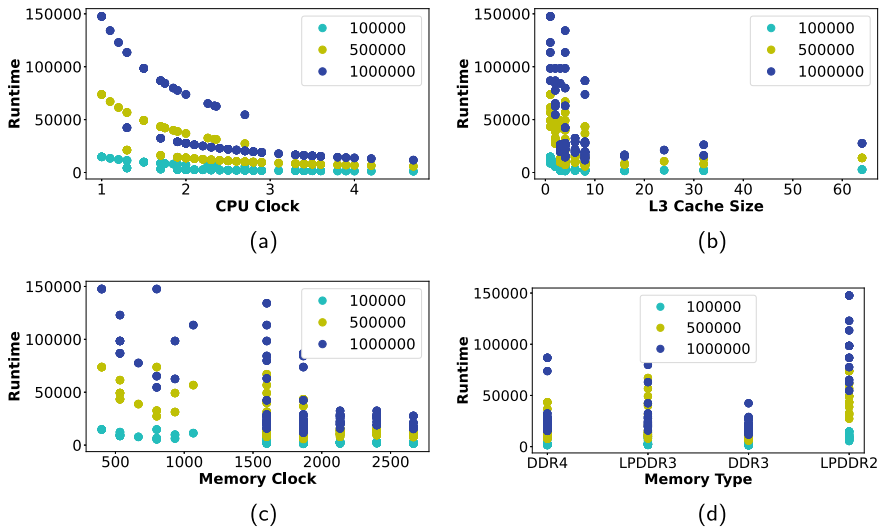
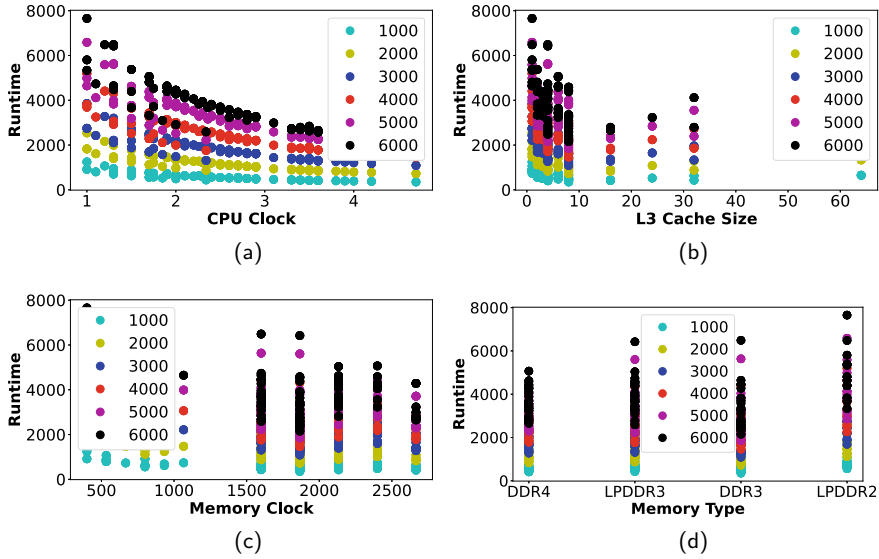


Fig. 3 Gem5 simulated systems - Monte Carlo hardware features versus runtime



**Fig. 4** Gem5 simulated systems - quickSort hardware features versus runtime

- The runtime reduces with an increase in CPU clock for each problem size, i.e., runtime is inversely proportional to the CPU clock.
- A similar trend is also observed in the case of L3 cache size and memory clock.
- Slower memory types LPDDR2 and LPDDR3 require higher runtimes than faster memory types DDR3 and DDR4. Similarly, LPDDR2 takes higher runtime compared to LPDDR3.
- We observe that in both simulated systems and physical systems memory clock greater than 1600 MHz significantly reduces the runtimes.

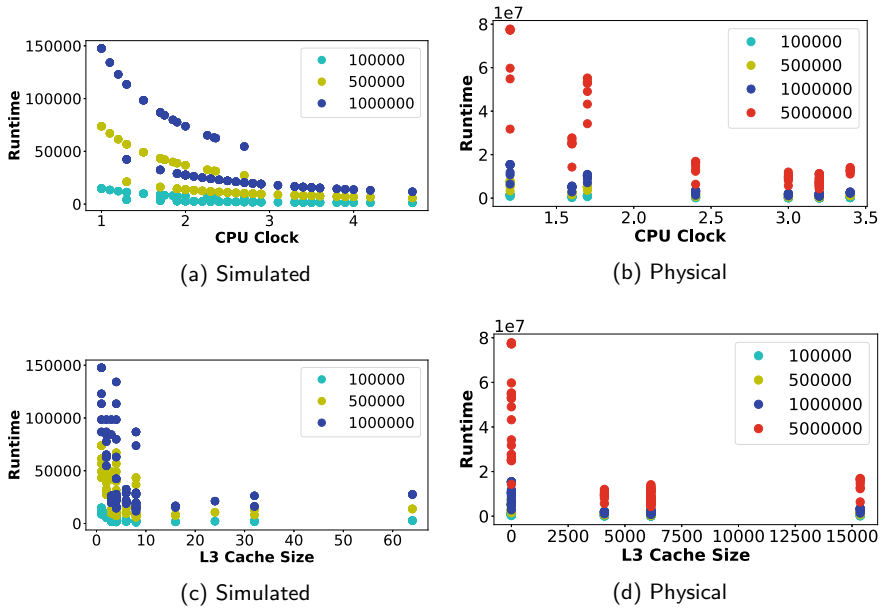
### 5.3 Comparing Simulated Systems Data with Physical Systems

We need to ensure that the results from Gem5 simulated systems are analogous to that of physical systems. We, therefore, executed the same three applications on eight physical systems with hardware features listed in Table 4. We collected hardware features using the dmidecode utility on x86-based systems and from the datasheet of ARM-based systems. We also collected runtimes from application logs for each application execution on these physical systems. We plotted the hardware feature values with respect to runtimes collected from these physical systems and compared them with simulated systems in Figs. 5 and 6. We make the following observations:

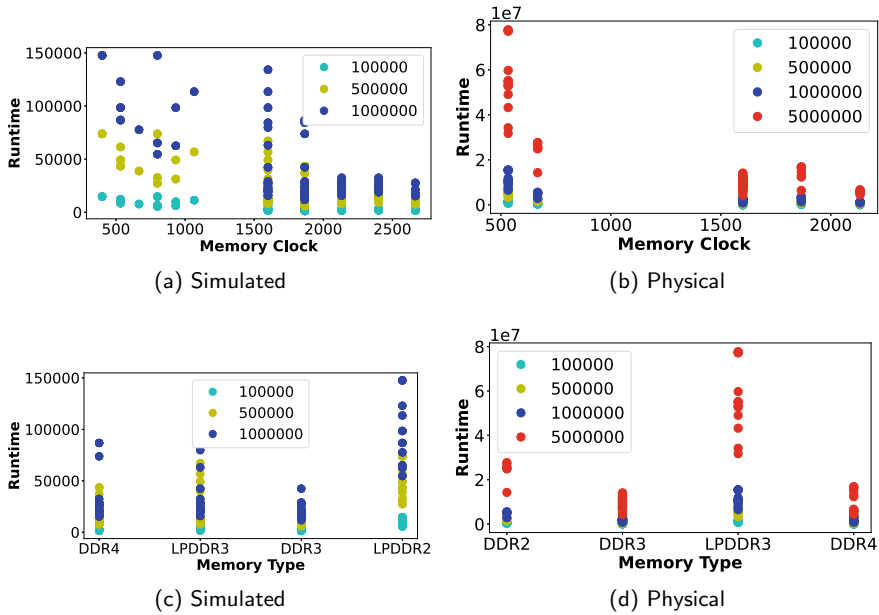
**Table 4** Physical computer systems

Sr	ISA	CPU speed GHz	Cores	Mem type	Mem access MHz	Mem size GB	L1–L3 cache size
1	x86	1.6	2	DDR2	667	2	32–512 kB
2	x86	3.2	4	DDR3	1600	4	32 kB–6 MB
3	ARM	1.2	4	LPDDR3	533	1	8–512 kB
4	ARM	1.7	2	LPDDR3	533	2	4 kB–2 MB
5	x86	3.2	4	DDR3	1600	4	32 kB–6 MB
6	x86	2.4	12	DDR4	1866	16	32 kB–15 MB
7	x86	3.2	4	DDR4	2133	4	32 kB–6 MB
8	x86	3.2	4	DDR3	1600	4	32 kB–6 MB
9	x86	3.4	4	DDR3	1600	4	32 kB–6 MB

Configuration taken from following models  
(1) Intel Core 2 Duo. (2) Intel Core i54460. (3) Qualcomm ARM Cortex A53. (4) Qualcomm snapdragon 600. (5) Intel Core i56500. (6) Intel Xeon E52620. (7) Intel Core i56500. (8) Intel Core i53470. (9) Intel Core i53470



**Fig. 5** Gem5 simulated systems versus physical systems - Monte Carlo hardware features versus runtime



**Fig. 6** Gem5 simulated systems versus physical systems - Monte Carlo memory features versus runtime

- Figure 5b depicts that runtime is inversely proportional to CPU clock even for physical systems as the case with for simulated systems in Fig. 5a.
- We observe that the x86-based physical system with a 1.6 GHz CPU clock has lower runtime compared to the ARM-based system with a CPU clock of 1.7 GHz. This is because an x86-based system with a 1.6 GHz CPU clock has DDR2 667MHz memory which is faster than LPDDR3 533 MHz memory used in ARM-based systems with a 1.7 GHz CPU clock.
- For physical systems, slower memory types LPDDR3 require higher runtimes than faster memory types DDR2, DDR3 and DDR4. This is similar to slower memory types LPDDR2 and LPDDR3 memory taking higher runtimes than faster memory types DDR3 and DDR4.

## 6 Conclusion

In this paper, we have identified the challenges of building systems in Gem5 simulator emulation mode with real-world systems hardware features. We have addressed these challenges by modifying the Gem5 source code to provide solutions. We have shown that with these modifications in the Gem5 simulator, we can emulate systems with features close to real-world systems. The results show that the relationships between

the hardware feature values with respect to runtimes in simulated systems match with physical systems. In the future, we plan to use systems built in Gem5 for performance prediction modeling.

## References

1. Butko A, Garibotti R, Ost L, Sassatelli G (2012) In ReCoSoC 2012–7th International Workshop on Reconfigurable and Communication-Centric Systems-on-Chip, Proceedings (University of York, York. <https://doi.org/10.1109/ReCoSoC.2012.6322869>
2. Li B, Peng L, Ramadass B (2009) J Syst Architect 55(10–12):457. <https://doi.org/10.1016/j.sysarc.2009.09.004>
3. Ozisikyilmaz B, Memik G, Choudhary A (2008) In proceedings of the international conference on parallel processing , pp 495–502. 10.1109/ICPP.2008.36
4. Kumar A, Talawar B (2018) In 2018 Eleventh international conference on contemporary computing (IC3) : 2-4 August 2018. Jaypee Institute of Information Technology, Noida, India (IEEE, Noida, India)
5. Malazgirt GA, Yurdakul A (2017) J Syst Architec 72:3. <https://doi.org/10.1016/j.sysarc.2016.07.004>
6. Binkert N, Beckmann B, Black G, Reinhardt SK, Saidi A, Basu A, Hestness J, Hower DR, Krishna T, Sardashti S, Sen R, Sewell K, Shoaib M, Vaish N, Hill MD, Wood DA (2011) ACM SIGARCH Comput Architec News 9(2):1. 10.1145/2024716.2024718. <https://dl.acm.org/doi/10.1145/2024716.2024718>
7. Cano-Cano J, Andújar FJ, Alfaro FJ, Sánchez JL (2019) J Parallel Distrib Comput 133:124. <https://doi.org/10.1016/j.jpdc.2019.06.013>
8. Reddy BK, Walker MJ, Balsamo D, Diestelhorst S, Al-Hashimi BM, Merrett GV (2017) In 2017 27th international symposium on power and timing modeling, optimization and simulation, PATMOS 2017, vol 2017-Janua. Institute of Electrical and Electronics Engineers Inc., pp 1–8. 10.1109/PATMOS.2017.8106988
9. Butko A, Bruguier F, Gamatié A, Sassatelli G, Novo D, Torres L, Robert M (2016) In proceedings - IEEE 10th international symposium on embedded multicore/many-core systems-on-chip, MCSoc 2016. Institute of Electrical and Electronics Engineers Inc., pp 201–208. 10.1109/MC-SoC.2016.20
10. Ma J, Yan G, Han Y, Li X (2016) IEEE Trans Comput 65(2):367. <https://doi.org/10.1109/TC.2015.2419655>
11. Walker M, Bischoff S, Diestelhorst S, Merrett G, Al-Hashimi B (2018) In: Proceedings - 2018 ieee international symposium on performance analysis of systems and software, ISPASS 2018. Institute of Electrical and Electronics Engineers Inc., pp 44–53. 10.1109/ISPASS.2018.00013
12. Odajima T, Kodama Y, Sato M (2018) In: 21st IEEE symposium on low-power and high-speed chips and systems, COOL Chips 2018 - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp 1–3. 10.1109/CoolChips.2018.8373083
13. Yao G, Yun H, Wu ZP, Pellizzoni R, Caccamo M, Sha L (2016) IEEE Trans Comput 65(2):601. <https://doi.org/10.1109/TC.2015.2425874>
14. Wu H, Liu F, Lee RB (2017) IEEE Comput Architec Lett 16(1):14. <https://doi.org/10.1109/LCA.2016.2597818>
15. Gem5 (2012) Gem5: O3CPU. <http://gem5.org/O3CPU>
16. Asanovic K, Bodik R, Christopher B, Joseph C, Gebis J, Husbands P, Keutzer K, Patterson DA, Lester W, John P, Samuel S, Williams W, Yelick KA (2006) The landscape of parallel computing research: a view from Berkeley. Tech Rep. <http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html>



17. Asanovic K, Bodik R, Demmel J, Keaveny T, Keutzer K, Kubiawicz J, Morgan N, Patterson D, Sen K, Wawrzynek J, Wessel D, Yelick K (2009) Commun ACM 52(10):56. <https://doi.org/10.1145/1562764.1562783>
18. Guthaus MR, Ringenberg JS, Ernst D, Austin TM, Mudge T, Brown RB (2001) In: 2001 IEEE international workshop on workload characterization, WWC 2001. Institute of Electrical and Electronics Engineers Inc., pp 3–14. 10.1109/WWC.2001.990739

# Construction Method of Operational Concept Model Based on Architecture Framework



Jing An, Lei Zhang, Miaoting Zeng, and Xu Han

**Abstract** The operational concept is the proposed solutions to operational problems under the specific space–time conditions in the future by judging the established operational concept and analyzing the essential laws of operational elements, such as operational conditions and adversaries. Its elements are numerous and complex. In order to achieve a clear description of the elements, structure, and relationship of the operational concept, this paper, based on the analysis and comparison of the architecture framework, takes DoDAF2.0 as a basic architecture framework, proposes a construction method of the operational concept model based on the architecture framework, realizes the modeling of the static elements and dynamic processes of the operational concept.

**Keywords** DoDAF2.0 · Operational concept · Architecture model

## 1 Introduction

The development process of operational concept needs to go through a cycle of iteration from idea to action, from strategic analysis to detailed design, from subjective conception to objective description, from general principles to specific guidance, from theoretical research to transformation and application, from rough to detailed, and constantly deepen. This process is not a simple logical reasoning process, but a

---

J. An (✉) · L. Zhang · M. Zeng · X. Han  
National Defense University, Beijing 100091, China  
e-mail: [anj21\\_2000@sina.com](mailto:anj21_2000@sina.com)

L. Zhang  
e-mail: [zxcv123o@163.com](mailto:zxcv123o@163.com)

M. Zeng  
e-mail: [zmt1001@163.com](mailto:zmt1001@163.com)

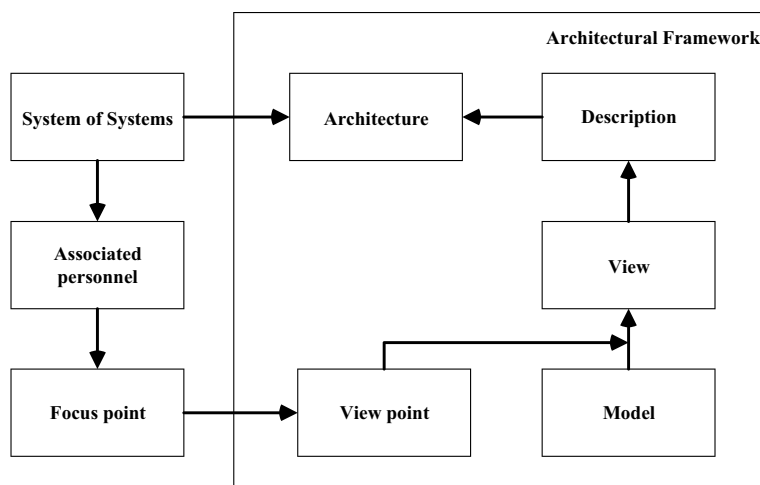
X. Han  
e-mail: [hanxu@sina.com](mailto:hanxu@sina.com)

complex scientific system project requiring long-term cooperation [1]. For example, DARPA's Strategic Technology Office (STO) put forward the "Mosaic Warfare" in August 2017. By September 2019, with the cooperation of multiple departments, it had only completed the concept connotation, theoretical analysis, concept formulation, and preliminary evaluation by using seminars, academic conferences, war games, limited target experiments, spiral events, focused simulation experiments, and focused practical exercises [2, 3]. If we simply rely on the macro-abstract combat concept description, it is not only difficult to effectively clarify its element composition and mutual relationship, but also impossible to carry out subsequent combat concept design, development, analysis, and continuous iterative update. Therefore, it is necessary to use scientific and systematic engineering methods, based on different perspectives, to take into account micro-actions and macro-effects, to carry out structural decomposition and transformation of the system, as well as framing and formalizing description of operational concepts with low structure, strong abstraction, and high complexity, and to build an integrated and consistent operational concept model that can reflect the characteristics and requirements of operational concepts from different perspectives, so as to realize the standardized, explicit and unambiguous description of all elements and interactions of the operational concept, and better support the development of the operational concept. Based on DoDAF2.0 basic architecture framework, this paper proposes the construction method of operational conceptual model.

## 2 Selection of DoDAF2.0 Architecture Framework

In the field of military, there are many architecture frameworks, such as UML military modeling, DoDAF2.0 architecture framework (Department of Defense Architecture Framework), MoDAF, Coalition Battle Management Language (CBM), Military Scenario Definition Language (MSDL).

This paper selects DoDAF2.0 as the basic architecture framework for modeling, mainly based on the following considerations: Firstly, DoDAF2.0 is the most widely used framework to guide the construction and description of architecture models, with a mature multi-view model system. Fifty-two models based on eight viewpoints (see DoDAF2.0 official literature [4, 5]) can basically meet the multi-perspective requirements of combat concept modeling. The second is CV in DoDAF2.0 capability perspective, which can cooperate with other perspectives to better describe the combat capability of executing specific combat operations under specific conditions to obtain the expected combat effect in terms of capability concept, classification, phase, dependency, capability organization mapping, capability operational action mapping, capability service mapping, etc., basically meeting the requirements of the focus capability requirements of the combat concept model. The third is that DoDAF2.0 abstracts, classifies, and organizes the complex elements within the operational concept system according to their relevance and consistency by using models, viewpoints, and views, as shown in Fig. 1, to achieve a structured and standardized



**Fig. 1** Architecture framework

description of the operational concept and ensure the consistency and coherence of the architecture data in the whole life cycle of operational concept design, development, analysis, etc.

### 3 Analysis of Typical Elements of Operational Concept and Construction of Relationship Map

As the basic architecture framework of operational concept modeling, DoDAF2.0 has certain advantages, but it also fails to fully meet the requirements of operational concept architecture modeling due to the following aspects: Firstly, DoDAF is originated from information system C<sup>4</sup>ISR, focusing more on information transmission and processing, and it has low support for the description of relationships between operational elements such as combat units and combat tasks, except for information flow [6, 7]; second, although DoDAF2.0 has added a capability view model, it is more of a top-level design model, focusing on the analysis of capability requirements, deployment planning, etc. It does not support the capability indicators and the relationship between capability indicators for specific operations of the operational concept. Therefore, based on the analysis of operational concept elements and the construction of the relationship map of operational concept elements, we tailored the applicability of the existing framework of DoDAF2.0 and optimized and adjusted its view model to match the modeling requirements and provide support for the construction of the operational concept model. 67.

### ***3.1 Analysis of Typical Elements of Operational Concept***

The development process of “threat based” operational concept is usually based on the research of operational environment and combat opponents, considering the impact and constraints of the battlefield environment, analyzing the elements such as operational forces, operational activities, and operational concept for the implementation of operational missions and tasks, clearly proposing the key capabilities required to support the operational concept, describing the requirements, and analyzing the capability dependency, capability gap, etc. Therefore, the operational concept usually contains three types of elements.

The first type is the description of operational problems, which is the description of the operational background, objectives, tasks, environments, opponents, time settings, conflict scenarios, etc., of the operational concept under the specific space–time conditions in the future.

The second type is the solutions to specific operational problems, including the guiding ideology of operational concepts, operational principles, operational forces, operational activities, and operational concepts.

The third type is operational capability requirements, that is, the key capabilities and their indicator requirements required to transform landing solutions to achieve operational effects, as well as the gap analysis between current capabilities and required capabilities.

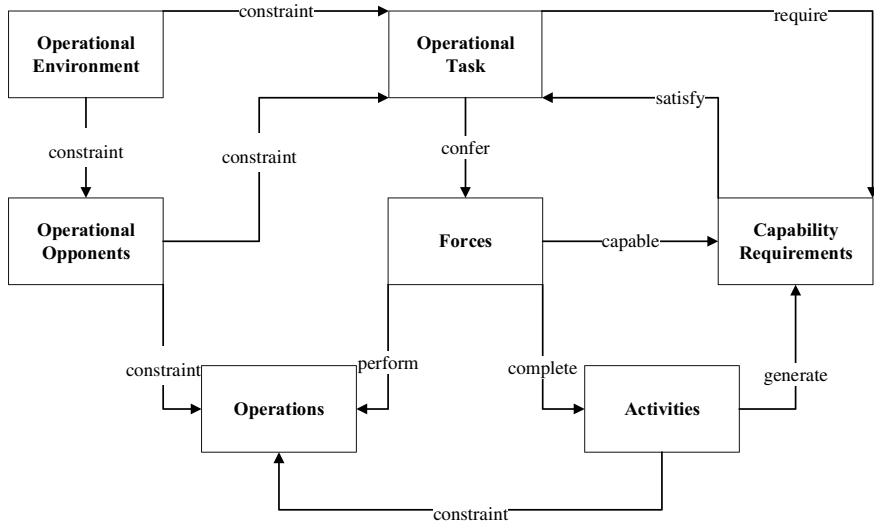
### ***3.2 Construction of the Relationship Map of Typical Elements of Operational Concept***

Centering on the generation of combat capability, the above three types of elements are sorted out, and the core typical elements describing the operational concept are extracted, mainly including the environment, tasks, opponents, forces, operations, activities, and capability requirements. By analyzing the association relationship of each element, the relationship map is constructed as shown in Fig. 2.

Based on the relationship map of the typical elements of the operational concept, the DoDAF2.0 architecture framework is selected to build the operational concept architecture model.

## **4 Construction of Operational Concept Architecture Model**

Based on the map of typical elements of the operational concept, focusing on the “capability requirements analysis” of the established target, the architecture model of the operational concept is built with the Capability View (CV) as the core and the



**Fig. 2** Typical elements' relationship graph of operational concept

model under the All View (AV), Operational View (OV), and System View (SV) as the basic model.

**4.1 Construction of Profile Information Model (AV-1)**

Describe the purpose, scope, background of the operational concept architecture model, form a summary information model AV-1, and output it in structured text.

**4.2 Construction of Capability Concept Model (CV-1)**

According to the operational concept theory, describe the operational concept, propose the decomposition of the operational purpose and capability, and form the capability concept model CV-1, which is mainly described in the form of graphics or text as shown in Fig. 3.

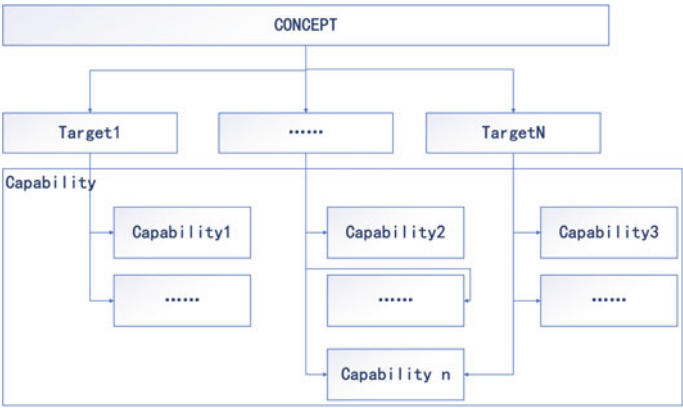


Fig. 3 Capability concept model CV-1

**4.3 Construction of Top-Level Operational Concept Map (OV-1)**

The operational environment, operational tasks, operational opponents are described in top-level operational concept map OV-1. Due to the comprehensiveness of the map, various forms of expression can be adopted, such as graphics, text, video.

**4.4 Modeling of Operational Forces (OV-4, SV-5b)**

The operational forces’ model is described, including:

- (1) The formation and interaction of forces, and form an organizational chart OV-4, as shown in Fig. 4.
- (2) The equipment system of systems and operations’ tracking matrix SV-5b is formed by using the two-dimensional matrix, taking the weapon system and operations as the input and the mapping relationship as the output, as shown

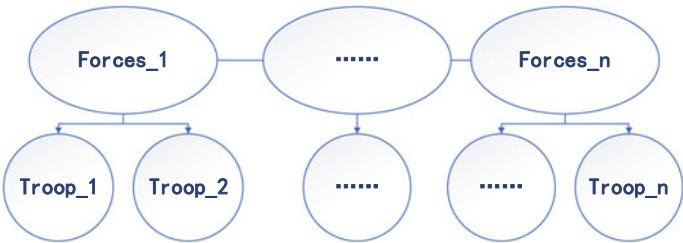


Fig. 4 Organization chart OV-4

**Table 1** Equipment–operations’ tracking matrix SV-5b

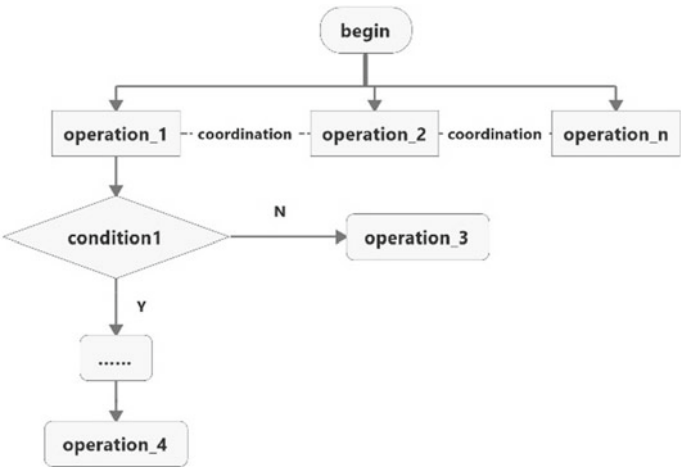
		Operational action				
		Action 1	Action 2	.....	.....	Action n
Equipment	Equipment 1	✓		✓		
	Equipment 2	✓			✓	✓
	.....		✓			
	Equipment n				✓	✓

in Table 1. This table describes the mapping relationship between weapon equipment system and operations, and it provides analysis basis for confirming whether key operations have been supported by reasonable weapons and equipment.

4.5 Modeling of Operations (OV-5a, OV-5b)

The main operations are described, including:

- (1) Based on operational functions, operational processes, operational objectives, operational means, etc., the operations are decomposed from top to bottom, layer by layer, forming an operational decomposition tree OV-5a, which is usually displayed in a tree structure.
- (2) The interaction characteristics between operations, including collaboration, cycle iteration, selection, judgment, etc., are described. A combat action model OV-5b is formed, usually using graphics to show, as shown in Fig. 5.



**Fig. 5** Operational model OV-5b



4.6 Modeling of Operational Activities (OV-6a, OV-6c)

Operational activities are described, that is, add time and sequence characteristics on the basis of the static structure, including:

- (1) The combat action and combat force are combined to describe the process and rules of activities and form the rule model OV-6a. These rules can be expressed in words or visualized in diagrams, as shown in Fig. 6.
- (2) The sequence diagram is used to represent the coordination relationship between different clusters and the sequence of operations, forming the event tracking model OV-6c, as shown in Fig. 7.

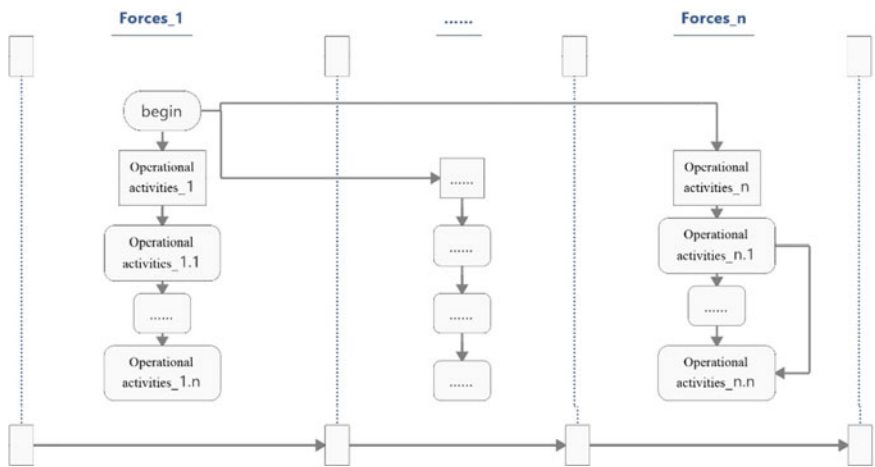


Fig. 6 Operational rule model OV-6a

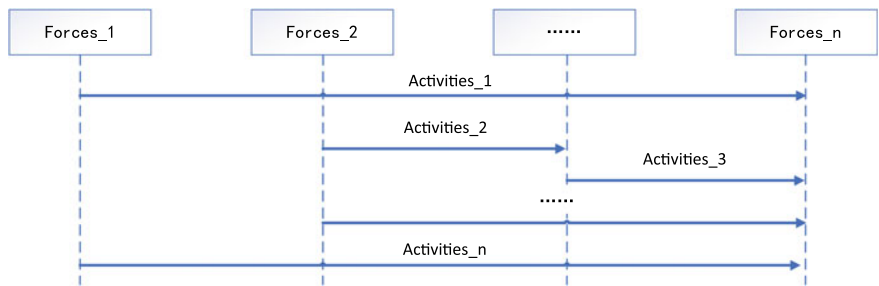


Fig. 7 Operational events' tracking model OV-6c

4.7 Modeling of Operational Capability Requirements (CV-6, CV-2, CV-4, CV-3)

According to the description of operations and activities, the capability requirements are proposed, including:

- (1) The mapping between combat operations and capabilities. The capability operations' mapping model CV-6 is formed using the mapping matrix table to express it, as shown in Table 2. Row represents capability and column represents operation [8]. The marks in the table indicate whether the implementation of combat operations supports corresponding capability requirements; OV-6a and OV-6c are the basis for scenario preparation and scheme design of simulation experiments.
- (2) The ability is classified and described in detail to form the ability classification model CV-2, which defines the ability demand indicators, the relationship between indicators and the ability measurement, and provides guidance and basis for building the ability analysis indicator system, which can be expressed in text, table, or graph, as shown in Fig. 8. The measurement of capabilities can be described in the form of tables.
- (3) The dependencies between capabilities are described to form the capability dependency model CV-4, which is usually described by graphs and represented by lines, as shown in Fig. 9.

Table 2 Capability–operation mapping model CV-6

		Operational action				
		Action 1	Action 2	Action 3	.....	Action n
Capability requirements	Capability 1	✓	✓	✓	✓	✓
	Capability 2	✓	✓		✓	✓
	.....	✓	✓		✓	
	Capability n				✓	

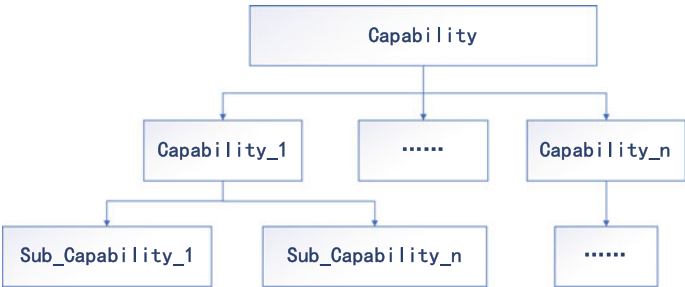


Fig. 8 Capability classification model CV-2

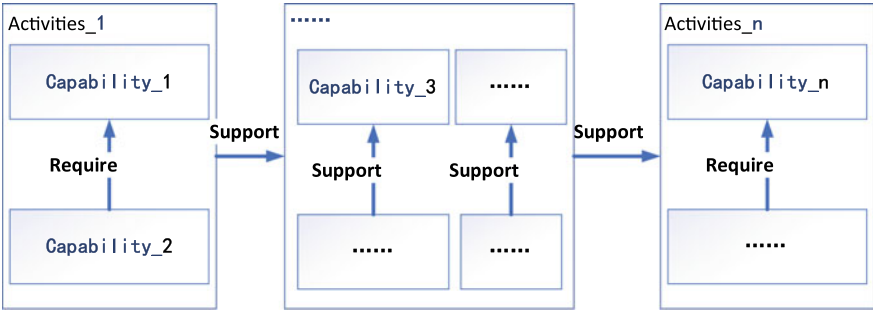


Fig. 9 Capability dependency model CV-4

5 Conclusion

Aiming at the modeling problem of operational concept architecture, based on the analysis of typical elements of operational concept, this paper adopts the engineering and systematic development concept and relies on the standardized and normalized model framework to construct a multi-perspective architecture model of operational concept with capability requirements as the core. This method can ensure the effectiveness, continuity, and compatibility of the model in the whole life cycle of the operational concept development and provide support for the design, development, and implementation of the operational concept.

References

1. An J, Liu W, Gu Z, Xu Weipeng (2022) Overview of capability requirements analysis methods for operational concept development. *Mil Oper Res Eval* 37 (3): 75–80
2. Guohong D (2020) The characteristics of the development of us army operational concept and its implications. *Natl Defense Sci Technol* 4:56–61
3. Chen S, Sun P, Li Daxi (2019) Analysis of the new concept of operations. Xi'an: Xi'an University of Electronic Science and Technology Press, pp 15–16
4. DoD architecture frame working group. DOD Architecture framework version 2.0, vol 1: Definitions and Guidelines. United States: Department of Defense
5. DoD architecture frame working group. DOD architecture framework version 2.0, vol 2. Product Descriptions. United States: Department of Defense
6. Sarkees M, Schafer P (2000) The correlates of war data on war: an update to 1997. *Confl Manag Peace Sci* 18(1):123–144
7. Hieb MR, Blalock J (2001) Data alignment between army C4I databases and army simulations. Abtechnologiesin Cal Alexandriava
8. Zhang Y, Qisheng G (2021) Operational concept design method of ground unmanned combat system based on DODAF. *Fire Command Control* 46(5):52–63

# Representation Learning with Attention for Spatial Reuse Optimization in Dense WLANs



Stephen Azeez and Shagufta Henna

**Abstract** IEEE802.11ax is designed to support self-configuration and adaptation functionality in dense deployment to enhance dynamic network conditions. Presence of variable transmission range is one of the major bottlenecks to network performance. In the absence of proper power management, co-existing IEEE 802.11ax access points cause co-channel interference, degrading throughput. Therefore, it is essential to consider the impact of the variable transmit power of neighboring nodes to optimize the network performance. This work proposes an affinityGNN-attention mechanism to capture neighborhood transmit power to generate an expressive network representation. Experiments results show that the attention module integration improves the prediction accuracy and robustness of the baseline affinityGNN model.

**Keywords** IEEE802.11 · Network management · Coexistence interference

## 1 Introduction

With the exponential growth of wireless networks, wireless access point self-management and adaptive adjusting capabilities have become increasingly important. Wireless network performance optimization is a critical requirement for intelligent wireless local area networks (WLANs) deployment. Nonetheless, the WLAN's tremendous deployment efficiency is also jeopardizing its potential development. Users are becoming more demanding, and the density of networks and clients is rising, thus the existing state-of-the-art in wireless technology is likely to soon fall short of supporting the ultra-dense deployment of WLAN access points (APs) and stations (STAs). Although, next-generation standard IEEE802.11ax uses a new PHY layer technology called Orthogonal Frequency Division Multiple Access (OFDMA)

---

S. Azeez (✉) · S. Henna

Department of Computing, Atlantic Technological University, Donegal, Ireland  
e-mail: [L00162428@atu.ie](mailto:L00162428@atu.ie)

S. Henna

e-mail: [shagufta.henna@atu.ie](mailto:shagufta.henna@atu.ie)

to enhance the performance and scalability of ultra-dense networks for a variety of transmission needs. Yet, when the number of APs and legacy nodes increases, the ultra-dense throughput of IEEE 802.11ax degrades.

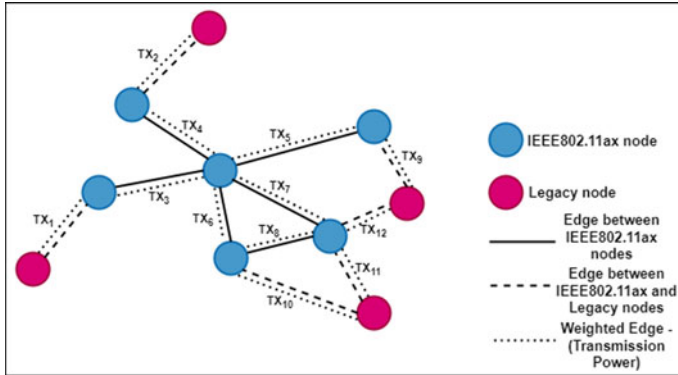
One of the key network resources in wireless communication is transmission power. High transmission power influences AP's power usage, creating interference between nodes operating simultaneous frequency. In contrast, a low transmission power will adversely affect the receiver's signal-to-noise ratio. To ensure successful transmission, power level of APs must be high enough while still low to avoid interfering with neighboring APs. In order to reduce network interference, high efficiency (HE) devices aim simultaneous transmissions to improve the overall network throughput. Recent years have witnessed an increase in research in transmission power control and dynamic sensitivity thresholds creating asymmetric links. These asymmetric links can create various types of interferences with different sensitivity and power control levels [1]. Variable transmission range results in interference among nodes under dense deployment scenario resulting in poor performance. The main goal of IEEE 802.11ax is to dramatically enhance user experience in deployment circumstances with high density.

Numerous efforts have been made to evaluate network throughput in the presence of variable transmission power to minimize interference to improve network throughput [2, 3]. These approaches leverage transmission power to exploit parallel transmissions based on the network's topological features [4–6]. As deep learning and AI techniques continue to advance, deep neural networks have gained significant popularity to solve and address network challenges [7–10]. Graph Attention QNetwork (GAQ) is introduced in [11] to evaluate the effect of the neighboring node using the attention framework. In another work in [12], a model named LA-ResNet is proposed that is based on spatio-temporal analysis and predicts wireless network traffic using an attention mechanism. A graph attention framework called ST-GAT [13] captures the spatial relationships using the LSTM network for extracting temporal domain elements. The suggested approach, in comparison to earlier relevant studies, is capable of capturing spatial relationships of traffic networks.

The reviews of existing literature on deep learning approaches for interference and network management work well under symmetrical links, however, these schemes fail to address the issue of coexistence under asymmetrical. To address this, this paper proposes graph-based deep learning approach called as attention-based affinityGNN using attention mechanism. The proposed approach outperforms graph-based deep learning approach without attention, i.e., affinityGNN.

## 2 Graph Model

The graph theory principles and measures were used to design the weighted wireless network. Edges are undirected in this scenario as they reflect the bidirectional relationship between APs. To define and model the wireless network's interacting relationships, an undirected graph  $G = (V, E, \omega)$  is employed. The vertex set for



**Fig. 1** An illustration of the graph model  $G$

the  $n$ -th AP vertices is  $V = \{v_1, v_2, \dots, v_N\}$ . The edge array  $E \in (V \times V)$ , where  $(v_i, v_j)$  means that node  $i$  and node  $j$  interact and  $\omega$  is the array of edge weights.

Under variable transmission power, the goal is to assign weights to the interference edges in the channel assignment where each assigned weight indicates variable transmission power. To compute edge weights in measure terms of distance  $d$ , the received signal strength indicator (RSSI) measurements (in dBm) must be converted into a metric of distance linking the access points  $i$  and  $j$  as given:

$$TX_{P(i,j)} = 10^{-\text{RSSI}_{i,j}/10} \quad (1)$$

Where  $TX_{P(i,j)}$  denotes importance between node ' $i$ ' features to node ' $j$ ' in the graph  $G$  as shown in Fig. 1.

### 3 Proposed Model

The proposed model employs an attention-based aggregating mechanism on affinityGNN to capture relational importance features with their attention signals on a static IEEE802.11ax communication network. AffinityGNN model with an attention mechanism can provide access points-specific attention weights on its interactive features. The graph structure with its weights is incorporated into attention mechanisms through affinityGNN embeddings that leverage the node2vec approach to learn the vertex representations. The affinityGNN embedding only provides static representations of IEEE802.11ax network interaction that do not reflect the dynamics IEEE802.11ax among access points in the network. Iteratively using features of every node for affinity calculation, attention-affinityGNN learns the hidden features of each node using self-attention. A standard convolution in direct-affinityGNN and skip-affinityGNN contains the standardized gated fusion aggregation of the features

of adjacent nodes in Eqs. (2) and (3). where  $M = D^{\frac{1}{2}} A' D^{\frac{1}{2}}$  and  $M_S = D^{\frac{1}{2}} A'_S D^{\frac{1}{2}}$ .

$$D^{l+1} = \sigma \sum (M D^{(l)} W^{(l)}_0, M_S S^{(l)} W^{(l)}) \quad (2)$$

$$S^{l+1} = \sigma \sum (M_S S^{(l)} W^{(l)}_S, M D^{(l+1)} W^{(l)}) \quad (3)$$

The aggregation function in Eqs. (2) and (3) propagates information between nodes and updates the hidden state of nodes to output the final embedding as the affinityGNN. In the graph convolution, attention-affinityGNN substitutes the above convolution operation with an attention mechanism. To depict the mode at which features of each node at  $l$ th layer are updated to those of  $l$ th + 1 layer, the model integrates the graph attentional layer's constituting component. A collection of node features in both direct  $D^l \in \mathbb{R}^F$  and skip  $S^l \in \mathbb{R}^F$  affinity is fed into an attention layer, with  $F$  denoting the number of features from each node. A pooled weight matrix  $\psi \in \mathbb{R}^{F \times F}$  is utilized to project the input toward other feature space of  $F^*$  dimension. Then, to measure/scores the relational importance between the nodes  $i$  and  $j$  if there is an edge between the node points, we define them as inputs to the attention layer to capture the attention coefficient score  $A_{i,j}$  in each encoded state for direct affinity ( $D$ ) and skip affinity ( $S$ ) is given in Eqs. (4) and (5), respectively.

$$A_{i,j}^D = \text{Att}(\psi D_i^{(l)}, \psi D_j^{(l)}) \quad (4)$$

$$A_{i,j}^S = \text{Att}(\psi S_i^{(l)}, \psi S_j^{(l)}) \quad (5)$$

such that  $\text{Att} : \mathbb{R}^F \times \mathbb{R}^{F^*} \Rightarrow \mathbb{R}$  is the attention layer, and  $A_{i,j}$  denotes the resulting computed attention correlation coefficient. The topological structure of the graph  $G$  in Eq. (4),  $A_{i,j}^S$  in equation 5 tends to solve this limitation by computing the attention coefficients of 2-hop neighbors. The attention scores is employed by utilizing the softmax function to normalize the attention coefficients using weight matrix  $\psi \in \mathbb{R}^{F \times F^*}$ .

$$\theta_{i,j} = \text{softmax}(A_{i,j}) \quad (6)$$

$$\theta_{i,j}^{D,S} = \frac{\exp(A_{i,j}^{D,S})}{\sum_{K \in \mathbb{N}(i)} \exp(A_{ik}^{D,S})} \quad (7)$$

where  $\theta_{i,j}$  is the attention score indicating the importance of  $i$  to  $j$  in direct affinity and skip affinity.

**Algorithm 1** affinityGNN-attention

---

```

1: INPUT: AFFINITYGRAPH  $G = (V, E)$ 
2:  $A \leftarrow$  Adjacency Matrix
3:  $\hat{A} \leftarrow \text{Normalization}(A)$ 
4:  $D \leftarrow \text{Node2vec}(\hat{A})$ 
5:  $D^{(0)} \leftarrow DW^{(0)}$ 
6: For each attention layer  $l, l = 1, \dots$  in  $(i, j) \in G$  do
7:    $D^{(l)} \leftarrow D^{(l)} W^{(l)}$ 
8:   COMPUTE ATTENTION VECTOR
9:    $A_{i,j}^D \leftarrow \text{Att}(\psi D_i^{(l)}, \psi D_j^{(l)})$ 
10:   $\theta_{i,j} \leftarrow \text{softmax}(A_{i,j}^D)$ 
11:  LEARN INFORMATIVE FEATURE REPRESENTATIONS
12:   $D^{(l+1)} \leftarrow D^{(l)} \theta_{i,j}$ 
13: End
14: COMPUTE OUTPUT
15:  $O \leftarrow \text{softmax}(D^{(\ell+1)}, W^{(\ell+1)})$ 
16: return  $O$ 

```

---

$$\tilde{D} = \theta^D D, \quad \tilde{S} = \omega \theta^S S \quad (8)$$

The fusion of  $\tilde{D}$  with  $\tilde{S}$  in the presence of trainable kernel coefficient  $k$  as given:

$$\gamma = \text{FusionGate}(\tilde{D}, \tilde{S})^k \quad (9)$$

We apply a further fully connected point-wise feedforward network,  $P_w$  to capture nonlinearity after the respective layers in the encoded state, matching the underlying structure of attention-affinityGNN.  $P_w$  with two linear transformations and an activation function ReLU is given in.

$$P_w = \sigma(\gamma W + \beta) \times W' + \beta' \quad (10)$$

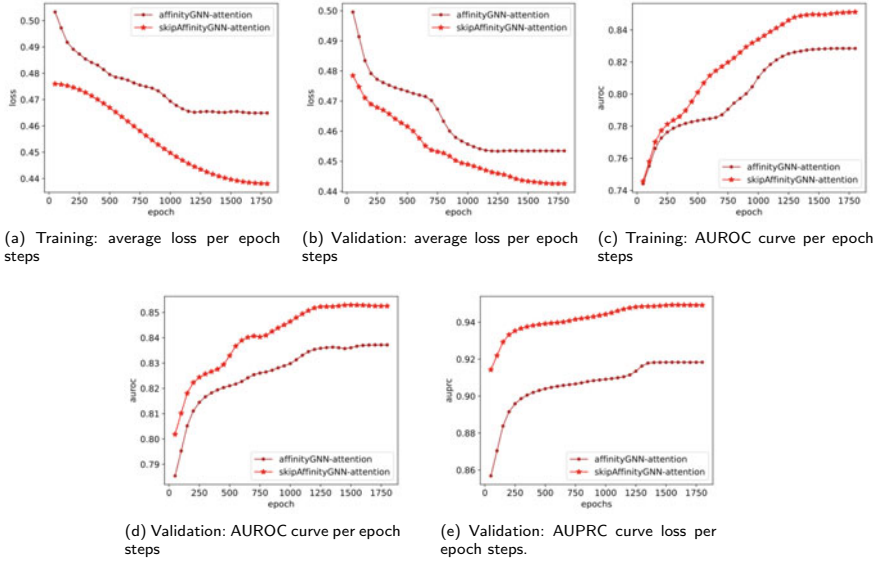
Here,  $W, W'$  are the propagated weights,  $\beta, \beta'$  denoting binary indicator bias and  $\gamma$  being the resulting attention matrix. Attention mechanism for the direct-affinityGNN and skip-affinityGNN-attention are given in Algorithms 1 and 2.

$$Z = \text{softmax}(W^* P_w + b^*) \quad (11)$$

$$\mathcal{L} = \sum T_{i,j} \log Z_{ij} + (1 - T_{i,j}) \log(1 - z_{pq}) \quad (12)$$

where  $T_{ij}$  is an indicator variable.





**Fig. 2** Experimental result showing the learning curves of affinityGNN-attention and skip-affinityGNN-attention

---

### Algorithm 2 skip-affinityGNN-attention Algorithm

---

```

1: INPUT: AFFINITYGRAPH  $G = (V, E)$ 
2:  $A \leftarrow$  Adjacency matrix of network AffinityGraph  $G$ 
3:  $A_D \leftarrow \text{Normalize}(A)$ 
4:  $A_S \leftarrow \text{Normalize}(AA^T)$ 
5:  $D \leftarrow \text{Node2Vec}(A_D), S \leftarrow \text{Node2Vec}(A_S)$ 
6: For each attention layer  $l, l = 1, \dots$  in  $(i, j) \in G$  do
7:    $D^{(l)} \leftarrow D^{(l)} W^{(l)}$ 
8:    $S^{(l)} \leftarrow S^{(l)} W^{(l)}$ 
9:   COMPUTE ATTENTION VECTOR
10:   $A_{i,j}^D \leftarrow \text{Att}(\psi D_i^{(l)}, \psi D_j^{(l)})$ 
11:   $A_{i,j}^S \leftarrow \text{Att}(\psi S_i^{(l)}, \psi S_j^{(l)})$ 
12:   $\theta_{D,S} \leftarrow \text{softmax}(A_{D,S})$ 
13:  LEARN INFORMATIVE NODE FEATURE REPRESENTATIONS
14:   $\hat{D} \leftarrow D^{(l)} \theta_{D,S}^D$ 
15:   $\hat{S} \leftarrow S^{(l)} \theta_{D,S}^S$ 
16:   $\gamma \leftarrow \text{FusionGate}(\hat{S}, \hat{D})_k$ 
17: End
18: POINTWISE FEEDFORWARD
19:  $P_w = \sigma(\gamma W + \beta) \times W' + \beta'$ 
20:  $z = \text{softmax}(W^* P_w + b^*)$ 
21: return  $z$ 

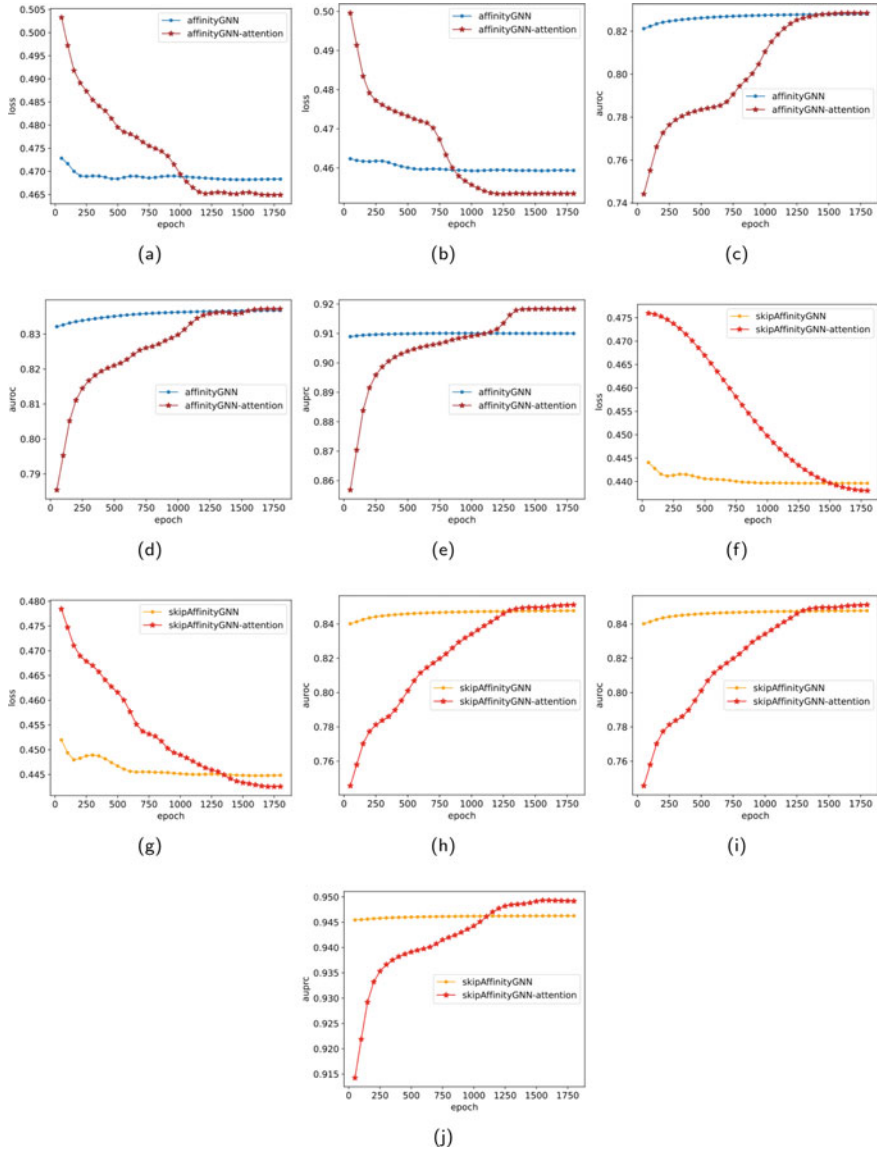
```

---

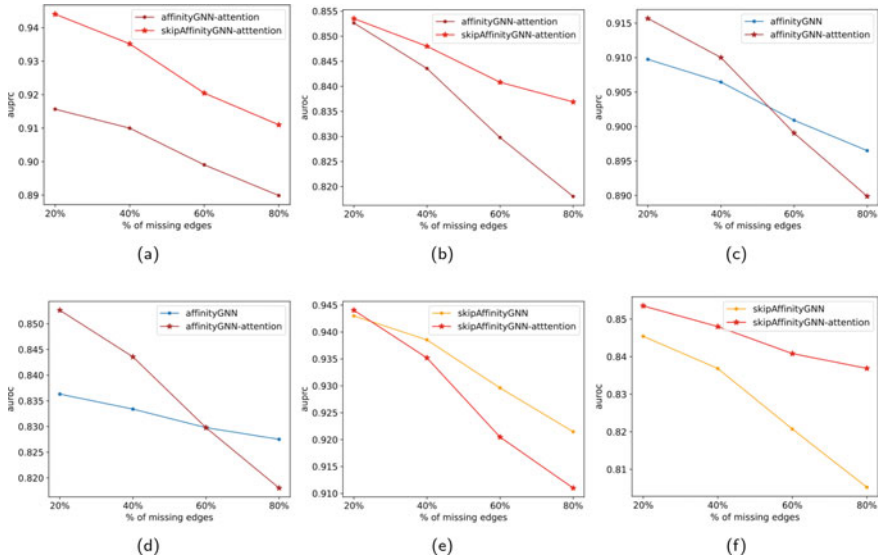
## 4 Result Analysis

In this setting, we first compare performance of affinityGNN-attention on a weighted IEEE 802.11ax-compliant network under 5 GHz band with skip-affinityGNN-attention model. We then validate the effectiveness and efficiency of affinityGNN-attention by comparing it to the baseline affinityGNN model. The learning rate is set to  $5e^{-4}$ ; the minibatch size of 256; the Adam optimizer. In our experiment, the total number of epochs is 1800 for each of the 50 steps. That is, within 3000 steps, 36 epochs will be executed. Figure 3 shows the learning curves for our proposed approach which depicts the performance level of skipAffintyGNN-attention over the just affinityGNN-attention for which across all the evaluation metrics, skip-affinityGNN-attention outperforms affinityGNN-attention. Figure 3a and b show lower loss values for both training and validation losses for skip-affinityGNN w.r.t to epochs compared to affinityGNN-attention model, resulting in higher training and validation accuracies according to Fig. 3c, d and e. The affinityGNN-attention curve reached the convergence zone at about 950–1100 epochs across all learning curves. In contrast, the converging speed of skip-affinityGNN-attention is much slower, taking much longer steps and time to reach the converging zone at about 1150–1250 epochs. For the skip-affinityGNN-attention model, the losses curve keeps a certain distance above the AffintyGNN-attention losses explaining the larger difference between validation and training accuracies reported for this model. Referring to Fig. 3e, the learning curve of the skip-affinityGNN-attention method converges to about 94.5% AUPRC accuracy of its validation while the curve of affinityGNN-attention increases up to 91.8%, that is, skip-affinityGNN-attention is 2.9% higher than affinityGNN-attention over the AUPRC metric.

Furthermore, a direct correlation of the proposed affinityGNN-attention model with the baseline affinityGNN model is another important criterion for evaluating the model in evaluation of the quality of the performance outcomes. As illustrated in Fig. 2, the affinityGNN model converges substantially quicker than the affinityGNN-attention model. As such, affinityGNN's convergence zone is reached before 250 steps, whereas the affinityGNN-attention loss curve is not converged till 1200 epochs, and there's a noticeable downward trend between zero and 1200 epochs. AffinityGNN-attention and affinityGNN validation PRC-AU curves converge at approximately 91.7% and 90.8%, respectively. Thus, the affinityGNN-attention method achieves greater improvements on the performance metrics than the baseline affinityGNN method. Consequently, skip-affinityGNN curve is consistently more stable early on and retains its quick convergence features, the skip-affinityGNN-attention curve continues to improve over time. In terms of the PRC-AU and ROC-AU trends depicted in the curve, the skip-affinityGNN-attention model exhibit the highest performance among the others. According to the PRC-AU curve displayed in Figs. 2e and 3e, our proposed models are successful in prediction classification with accuracy above 91% point.



**Fig. 3** Results i affinityGNN-attention versus affinityGNN of **a** training: average loss; **b** validation: average loss; **c** training: AUROC; **d** validation: AUROC; **e** validation: AUPRC. **ii** Skip-affinityGNN-attention versus skip-affinityGNN of **f** training: average loss; **g** validation: average loss; **h** training: AUROC; **i** validation: AUROC; **j** validation: AUPRC

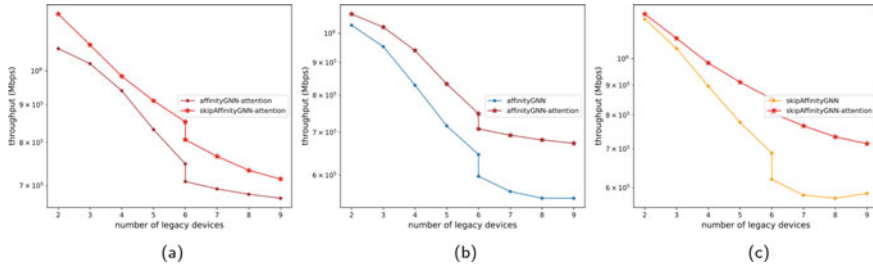


**Fig. 4** Experimental results showing the predictive robustness performance evaluation in the presence of incomplete network

### 4.1 Impact of Missing Edges

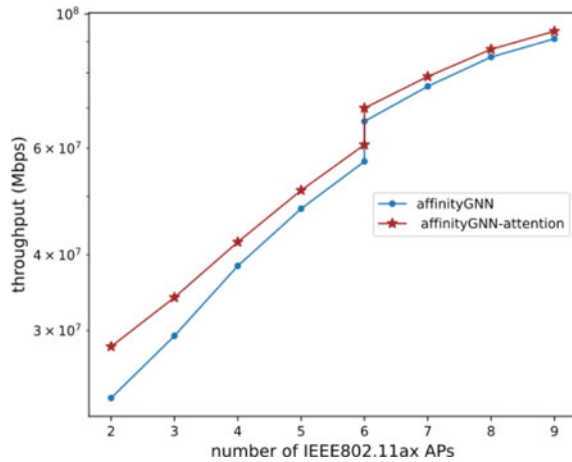
With the wireless network dynamics taken into account, we estimated the proposed model's performance as it neutralizes the missing edges during the network evolution. Our hypothesis is that the proposed model is able to handle the dynamics of wireless networks.

In the experiments, we randomly excluded 20, 40, 60, and 80% of the network edges to evaluate each algorithm on its classification prediction performance ability of the overall network's missing edges. We quantify the model prediction by using AUROC and AUPRC compared to the baseline approaches. Figure 4 shows the results of our proposed approach as skip-affinityGNN-attention gives a significant boost in performance in capturing true communicating network structure compared to affinityGNN-attention in both AUROC and AUPRC score in Fig. 4a, b and c, respectively. In contrast, skip-affinityGNN shows better performance in AUROC but a less score in AUPRC. In other respects, affinityGNN-attention performs particularly well when edges are incomplete over affinityGNN as shown in Fig. 4d, e and f.



**Fig. 5** Experimental results showing the throughput estimation in presence of legacy devices

**Fig. 6** Throughput versus number of connecting IEEE802.11ax APs



## 4.2 Impact on Throughput

The results presented in Fig. 5 demonstrate that, as the number of legacy devices increases, the overall performance continuously degrades. As a result, of an increase in the diameter of the network, IEEE802.11ax compatible APs require higher transmission power. To compare the efficiency of the proposed method for networks of various sizes, it is concluded that the proposed attention method achieves better throughput than the baseline method (Fig. 6)..

## 5 Conclusions

The effect of variable transmission power in heterogeneous networks coexistence is relatively proportional to the node and network existence as the appropriate transmission power rate is dependent on the interaction range. In order to achieve maximum network throughput, it is therefore necessary to find a balance between transmission

rate and interference levels effected by transmitted power. The proposed attention mechanism for affinityGNN captures the relative importance of each edge in terms of contribution to total network throughput.

## References

1. Afaqui MS, Garcia-Villegas E, Lopez-Aguilera E, Smith G, Camps D (Mar 2015) Evaluation of dynamic sensitivity control algorithm for IEEE 802.11ax. In: Proceedings IEEE wireless communications and networking conference (WCNC), pp 1060–1065
2. Ifedayo AO, Dlodlo ME (2015) Variable transmission power control in wireless Ad-Hoc networks. In: AFRICON 2015, pp 1–5
3. Shih KP, Chen YD, Chang CC (2011) A physical/virtual carriersense-based power control mac protocol for collision avoidance in wireless ad hoc networks. *IEEE Trans Parallel Distrib Syst* 22(2):193–207
4. Afaqui MS, Brown S, Farrell R (2018) Uplink performance optimization of ultra dense Wi-Fi networks using AP-managed TPC. *Wireless Days (WD)*. <https://doi.org/10.1109/wd.2018.8361703>
5. Gray S, Vadde V (May 2001) Throughput and loss packet performnee of DCF with variable transmit power. *IEEE 802.11* 0-11227
6. Khorov E, Kiryanov A, Lyakhov A, Bianchi G (2018) A tutorial on IEEE 802.11ax high efficiency WLANs. *IEEE Commun Surv Tutor*. <https://doi.org/10.1109/comst.2018.2871099>
7. Schmidt M, Block D, Meier U (2017) Wireless interference identification with convolutional neural networks. In: *IEEE 15th International conference on industrial informatics (INDIN)*. <https://doi.org/10.1109/indin.2017.8104767>
8. Grunau S, Block D, Meier U (2018) Multi-label wireless interference classification with convolutional neural networks. In: *IEEE 16th International conference on industrial informatics (INDIN)*, pp 187–192. <https://doi.org/10.1109/INDIN.2018.8471956>,
9. Junfei Y, Jingwen L, Bing S, Yuming J (2018) Barrage Jamming detection and classification based on convolutional neural network for synthetic aperture radar. In: *IGARSS 2018-2018 IEEE international geoscience and remote sensing symposium*, pp 4583–4586. <https://doi.org/10.1109/IGARSS.2018.8519373>.
10. Shen Y, Shi Y, Zhang J, Letaief KB (2019) A graph neural network approach for scalable wireless power control. In: *IEEE Globecom workshops (GC Wkshps)*. pp 1–6
11. Jin Y, Vannella F, Bouton M, Jeong J, Hakim EA (2021) A graph attention learning approach to antenna tilt optimization. *arXiv preprint arXiv:2112.14843*
12. Li M, Yuewen W, Zhaowen W, Huiying Z (2020) A deep learning method based on an attention mechanism for wireless network traffic prediction. *Ad Hoc Netw* 102258. <https://doi.org/10.1016/j.adhoc.2020.102258>,
13. Kong X, Xing W, Wei X, Bao P, Zhang J, Lu W (2020) STGAT: spatial-temporal graph attention networks for traffic flow forecasting. *IEEE Access* 8:134363–134372. <https://doi.org/10.1109/ACCESS.2020.3011186>

# A New Ultralightweight Authentication Protocol for IoTs: MFRAP



Umar Mujahid  and Binh Tran

**Abstract** Internet of Things (IoT) are a system of interconnected computing devices which have unique identification and ability to communicate with the backend cloud without human intervention. In IoT's systems, the identification of these computing devices (Edge Technology layer) incorporates Radio Frequency IDentification (RFID) technology and uses wireless channels to get connected to the gateway layer. Since the communication takes place wirelessly using channels making it vulnerable to many types of interceptors, therefore it poses many security, privacy, and traceability challenges. To ensure the secure communication between the computing devices and the readers/gateway, a new cryptographic field (for low-cost computing devices) was introduced: Ultralightweight Cryptography. The Ultralightweight encryption schemes mainly involve authentication protocols which allow the devices to mutually authenticate each other before the transmission of the data. These authentication protocols are termed as Ultralightweight Mutual Authentication Protocols (UMAPs) which mainly involve bitwise logical operations and some extremely lightweight non-triangular functions. However, the majority of these UMAPs are susceptible to many security and privacy attacks. This paper proposes a new UMAP: MFRAP which involves simple bitwise logical operations and performs authentication between the tags and the readers. We have used formal (structural) cryptanalysis methods to validate security of MFRAP, and the formal verification of the protocol is performed using BAN logic. The MFRAP protocol clearly fills the research gap of UMAPs and will can be used as a stepping stone in new development of UMAPs.

**Keywords** IoTs · RFID · UMAP · MFRAP · Privacy and security

---

U. Mujahid (✉) · B. Tran  
Georgia Gwinnett College, Lawrenceville, GA 30004, USA  
e-mail: [ukhokhar@ggc.edu](mailto:ukhokhar@ggc.edu)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023  
X.-S. Yang et al. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 694,  
[https://doi.org/10.1007/978-981-99-3091-3\\_78](https://doi.org/10.1007/978-981-99-3091-3_78)

961

# 1 Introduction

Internet of Things (IoTs) are an emerging technology in this modern era. The applications of the IoTs cover almost all the aspects of life, e.g., supply chain, home security, SCADA systems, weather forecasting, vehicle maintenance, safety systems, etc. The IoTs systems is the network of interconnected small computational devices which can communicate with each other and backend cloud autonomously. All these small devices are provided a unique ID through which these devices can be identified and differentiated with other devices. Almost, all of the IoTs systems incorporate RFID for device identification and authentication.

The RFID systems are mostly comprised three components: tag, reader, and the backend cloud. The tag is a small electronic chip which can be attached to the object/device and is used for identification. The reader performs a scanning operation and reads the contents of the all tags, entering its vicinity. The backend cloud stores detailed information about all the tags in the system. The backend cloud does share some of the information about the tags which is most likely to be scanned by the reader or will most likely to enter within the specific reader’s range. Communications between the reader and the backend cloud are secure, since it may use Virtual Private Network (VPN) or wires for connection. However, the communication between the reader and the tag is wireless which is open for all types of attackers, can be intercepted, and is prone to Man In the Middle Attack (MIMA). Figure 1 depicts the RFID system overview.

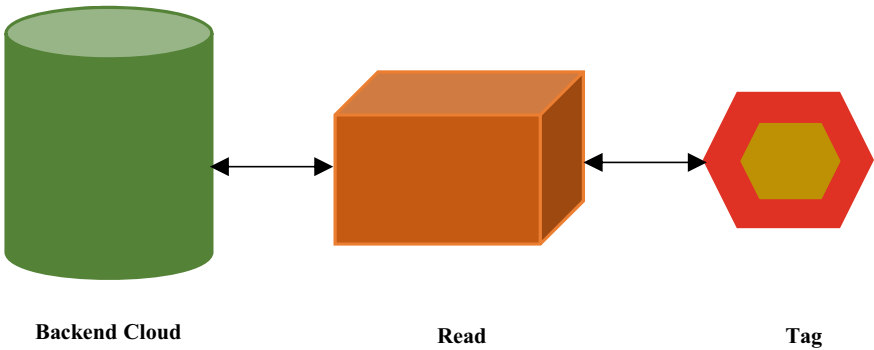


Fig. 1 Components of RFID system



## 2 Related Works

The concept of extremely lightweight authentication protocols for RFID and IoTs was introduced by Peris-Lopez et al. [1–3]. Pedro was the first researcher who used bitwise logical operators and designed a mutual authentication protocol for inexpensive systems. However, the term Ultralightweight was firstly used by Chien [4] when Chien designed his own UMAP [4] with non-triangular primitive. The detailed journey of UMAP's development and their cryptanalysis is presented as follows.

In 2006, Peris-Lopez et al. [1–3] proposed three UMAPs: Lightweight Mutual Authentication Protocol (LMAP), Extremely Lightweight Mutual Authentication Protocol (EMAP), and Minimalist Mutual Authentication Protocol (M2AP). All three UMAPs implemented bitwise logical operators such as OR, AND, XOR, and Modulo 2 addition and therefore can be implemented within 4 K GE (EPCC1G2 tags). The authors have performed the randomness test of the protocol messages to ensure the freshness and anonymity of the exchanged messages between the reader and the tag. Additionally, basic security analysis of the protocols was performed to justify the security claims of the protocols. In 2007, Li and Wang [5] identified two active attacks on LMAP: desynchronization and the full disclosure attacks. Since, in LMAP, the tag does not store the additional copies of the keys and IDS, the attackers exploit this inherent weakness and make the tags and reader store different copies of the victim tag which causes them desynchronized. In a full disclosure attack, the perpetrator firstly intercepts the messages of a legit session and then impersonates the reader. The attacker modifies the  $i$ th bit of the captured messages which directly flipped the concealed random numbers. After receiving communication from the tag, the attacker further eliminates the random numbers by solving the protocol messages (previous and fresh) and try to find the key and the tag ID. This attack model can be further extended to cryptanalyze the EMAP and M2AP. Later on, many other attacks on these protocols were also reported, and main vulnerability in which all these attackers were exploiting was utilization of only the binary operations in protocol messages designing.

In 2007, Chien [4] improved the design of these UMAPs and introduced a new UMAP: Strong Authentication and Strong Integrity (SASI). Besides the bitwise logical operators, the SASI protocol involves a non-triangular function "Rot" (Rotation) in the protocol messages design. The SASI protocol opened the new horizon in UMAP development and the protocol proved to more stable than its contending protocols. In 2008, Peris-Lopez et al. [6] exploited the poor construction of the protocol messages and reported full disclosure attack. The SASI protocol requires that the tag stores the previous and the current values of the ID, keys, and IDS, while the reader and the backend cloud only store the current values. In 2011, Sun et al. [7] exploited this weakness and identified two desynchronization attacks.

Later on, many other UMAPs, e.g., GOASMMER [8], DAVID-PARSAD [9], RAPP [10], RCIA [11], SASI+ [12], etc., were proposed which involved numerous non-triangular functions in protocol messages design. However, all of these protocols have been reported to be vulnerable against various security attacks [13–16].

In 2017, Mujahid et al. [17] introduced a new non-triangular primitive: Psuedo-Kasami code. The newly invented primitive was inspired from the Kasami sequence which enhanced the diffusion properties of the protocol messages. The protocol has been formally analyzed using well-structured cryptanalysis models, e.g., RLC, RDC, and Tango attacks. However, Saffkhani and Bagheri [15] identified weaknesses of KMAP structure and reported desynchronization attack on the protocol.

In 2022, Shariq et al. [18] proposed a UMAP: ESRAS. The ESRAS involves a new Ultralightweight function; Rank. The protocol was fully formally analyzed and verified using structural cryptanalysis and Scyther Tool (automatic security protocol verification tool). Recently, Servati et al. [19] found weaknesses of ESRAS and have shown that the recently proposed UMAP (ESRAS) does not protect the full disclosure attack and attacker can easily reveal the security ID of the tag.

On the major vulnerabilities of the existing UMAPs is that the freshness of the sessions entirely depends on the reader [20]. The reader generates the random numbers and then wraps it in the protocol messages in which tag needs to unwrap to authenticate the reader. To improve the security and to avoid the existing full disclosure attacks, both the parties (readers and the tags) should be able to generate the nonce.

### 3 A New Ultralightweight Mutual Authentication Protocol: MFRAP

This section presents our novel UMAP: Miniaturized Fully Random Authentication Protocol (MFRAP). The MFRAP involves existing bitwise logical operators: bit shuffling (clock function) and Exclusive OR (XOR), to design the protocol messages. However, in MFRAP, we have introduced a novel concept where besides the reader, the tag is also capable of generating the random numbers. We have proposed an extremely lightweight random number generator which does not increase the overall cost of the tag and it still compliance to EPCC1G2 tags. The randomness tests of the proposed random number generator have been performed using ENT, DIEHARD, and NIST test suites. The mathematical expression and the hardware design of the Miniaturized Random Number Generator (MRNG) is described as follows:

$$n_i = \text{Cl}(n_{i-1}, m),$$

where

$$m = C \oplus n_{i-1}.$$

$C$  is the constant value.

The computation of the clock function (Cl) involves two steps:

**Step 1: Rearrange bitstreams:**

$$n_{i-1} = a_0a_1a_2 \dots a_{l-1}a_l$$

$$m = m_1m_2m_3 \dots m_{l-1}m_l$$

Rearrange  $n_{i-1}$  and  $m$ :

$$a_0a_la_1a_2a_3a_{l-1} \dots$$

$$m_1m_2m_3m_lm_4m_5 \dots$$

**Step 2: XOR rearranged bitstreams:**

$$a_0a_la_1a_2a_3a_{l-1} \dots$$

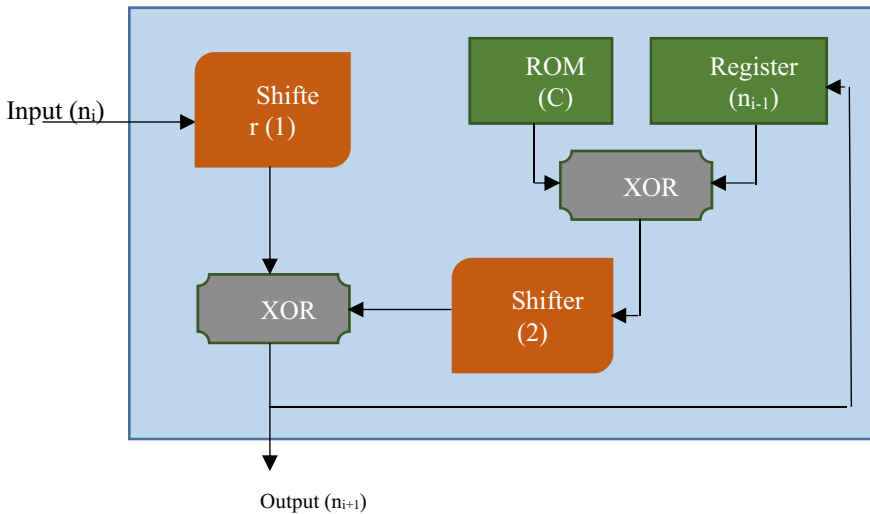
$$m_1m_2m_3m_lm_4m_5 \dots$$

$$\oplus$$

$$x_1x_2x_3x_4x_5x_6 \dots$$

Figure 2 depicts the hardware schematic of MRNG.

We have generated a 300 MB file of random numbers and used ENT, DIEHARD, and NIST randomness test suites to validate the statistical properties of MRNG. The short result of the randomness test is presented in Table 1.



**Fig. 2** Hardware schematic of MRNG

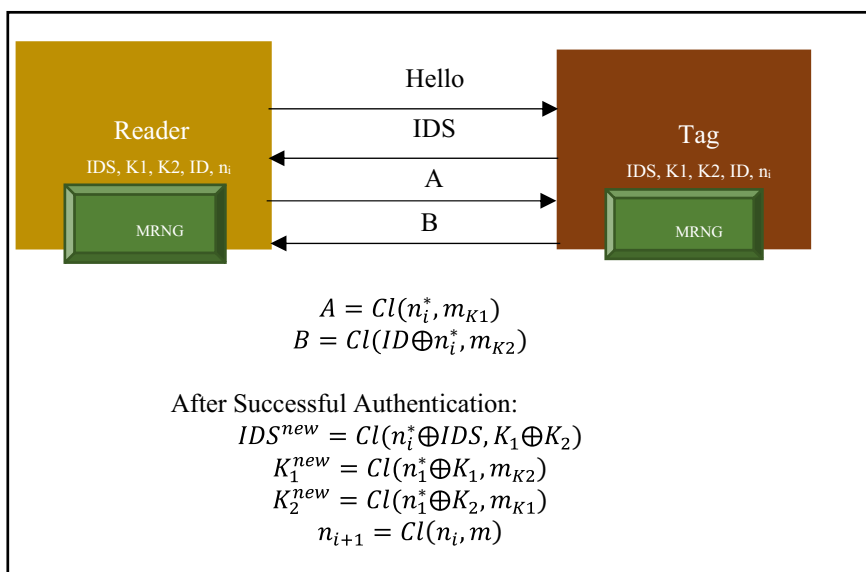
**Table 1** Randomness test results of MRNG

Parameters	Random number
Entropy (bits/bytes)	7.99997
Arithmetic mean	127.3896
Monte Carlo pi estimation	3.1418
Diehard (overall $p$ -value)	0.3376
NIST	Passed

The MFRAP protocol involves the novel clock function in its messages design and offers extremely good security with no additional communication cost. Figure 3 presents the working of MFRAP.

After the tag enters in the reader's vicinity:

1. The reader sends a "hello" message to the tag.
2. After receiving the "hello", the tag responds with IDS.
3. The reader computes the message "A" (which involves random number (known to tag) and pre-shared secret Key ( $K_1$ )).
4. The tag also calculates a local copy of message "A"; if both values match, the reader is authenticated and the tag computes and sends message "B". The message "B" includes the secret ID of the tag. After sending the message "B", the tag updates its IDS, both secret keys ( $K_1$ ,  $K_2$ ), and random number ( $n_i$ ).

**Fig. 3** MFRAP protocol

5. Once message “ $B$ ” is received, the reader also calculates the local copy of message “ $B$ ”, and if both values coincide, then tag also got authenticated and the reader will also update its variables for the specific tag.

Both the tag and the reader also store the previous values of keys and IDS; however, each time it will generate a new random number (unique for each session). If the session did not get completed, then the tag and reader will use the previous keys and IDS (with fresh random number) to calculate the public messages.

## 4 Security Analysis of MFRAP

In this section, we will perform the security analysis of the newly proposed MFRAP to validate our security claims. The security analysis model validates confidentiality, integrity, and avoidance to desynchronization and full disclosure attacks. The detailed description is presented as follows:

- (a) **Confidentiality:** The MFRAP offers extremely good confidentiality of the protocol messages. The publicly disclosed messages are encrypted with the secret Pseudo Random Number (PRN) and the pre-shared keys. It would be impossible for an attacker to retrieve the secret keys and the ID without knowing the random number. Moreover, the newly proposed non-triangular primitive clock function increases the computational complexity to  $O(2^n \times 2^C \times 2^{K1} \times 2^{K2})$  which is theoretically impossible to break within a practical time frame.
- (b) **Integrity:** Since, the MFRAP involves pre-shared secrets and random number (in messages design) which provides optimal integrity of the messages. For example, if an eavesdropper modifies the message “ $A$ ”, then legitimate tag will come up with a different value of “ $A$ ”, and therefore, the tag will be unable to authenticate the message  $A$  and turn down the communication session.
- (c) **Desynchronization attack:** To understand how MFRAP avoids the desynchronization attacks, let us consider the desynchronization attack scenario inspired from Sun et al. [7] attack model.

Assume that, an attacker joins the legitimate session. The attacker firstly intercepts the IDS and message “ $A$ ” of the session 1. When the legitimate tag responds with the message  $B$ , the attacker blocks the message “ $B$ ”, so the tag updates its variables while the reader will keep using the previous pseudonym IDS and keys. Now, in the session 2, the reader again sends “hello” message to initiate the communication session with the tag. The tag responds with IDS; if this IDS is IDS (old), then the reader uses previous variables with fresh random number to compute the message  $A$ , which will easily be authenticated by the tag. Therefore, if an attacker blocks any of the legit public messages, it will not affect the legitimate reader nor the tag. The presence of the fresh random number for each new session avoids all types of the desynchronization attacks.

- (d) **Structural Cryptanalysis:** In this section, we will discuss how MFRAP avoids the full disclosure attacks using well-defined structural cryptanalysis:
- **Tango Attack:** In the Tango attack model [16], the attacker eavesdrops number of sessions and collects the publicly exchanged messages for several sessions. Then, attacker creates the Good Approximation (GA) equation and tries to guess the concealed secrets. However, in MFRAP, since each session involves a fresh random number, therefore, there will be not relationship/interdependency between the messages of the different sessions. The GA equations will also be not able to retrieve the concealed secrets.
  - **RLC and RDC:** Ahmedian et al. [14] presented two formal structural cryptanalysis: Recursive Linear Cryptanalysis (RLC) and Recursive Differential Cryptanalysis (RDC).

The RLC mainly exploits the inherent weak diffusion properties of  $T$ -function and tries to form a linear matrix of known and unknown variables, then solve the matrix to find the unknown parameters. However, RLC will not be applicable to MFRAP since, we have extensively used non- $T$ -function (Clock) in our messages design which prevents the linear matrix computation for attackers.

The RDC is an active attack model, which tries to restrict the legitimate parties for more communication sessions and tries to create linear equation (differences) of various public session. However, the inclusion of random number in each session makes it impossible to create linear equations of the variables.

## 5 Conclusions

In this paper, we have presented a new UMAP: MFRAP which involves a new non-triangular function (Clock) in its messages design to avoid many of the existing attacks. In MFRAP, we introduced a concept of random number generation at the tag side which even makes it more robust against all the existing desynchronization and the full disclosure attacks. We believe that MFRAP will provide stability and longevity to the UMAP family and can act as a bridge for future improvement of the UMAPs. In future, we will perform the hardware implementation of the MFRAP using ASIC to conform its compatibility with EPCC1CG2 tags.






## References

1. Peris-Lopez P, Hernandez-Castro J, et al (2006) LMAP: a real lightweight mutual authentication protocol for low-cost RFID tags. In: Proceedings of the 2nd workshop on RFID security, Austria, pp 100–112
2. Peris-Lopez P, Cesar Hernandez J, et al (2006) EMAP: an efficient mutual-authentication protocol for low-cost RFID tags. In: Proceedings of the 1st international workshop on information security (OTM-2006), France, pp 352–361

3. Peris-Lopez P, Hernandez-Castro JC, Tapiador JME, Ribagorda A (2006) M2AP: a minimalist mutual authentication protocol for low cost RFID tags. In: International conference on ubiquitous intelligence and computing, Wuhan China, pp 912–923
4. Chien HY (2007) SASI: a new ultralightweight RFID authentication protocol providing strong authentication and strong integrity. *IEEE Trans Depend Sec Comput* 4(4):337–340
5. Li T, Wang G (2007) Security analysis of two ultra-lightweight RFID authentication protocols. In: International information security conference, South Africa, pp 109–120
6. Peris-Lopez P, Hernandez-Castro JC, Tapiador JME, van der Lubbe JCA (2009) Security flaws in a recent ultralightweight RFID protocol. *arXiv preprint (Technical Report) No. 0910.2115*
7. Sun HM, Ting WC, Wang KH (2011) On the security of Chien's ultralightweight RFID authentication protocol. *IEEE Trans Depend Sec Comput* 8(2):315–317
8. Peris-Lopez P, et al (2009) Advances in ultralightweight cryptography for low-cost RFID tags: gossamer protocol. In: Proceedings of the 9th international workshop on information security applications, Korea, pp 56–68
9. David M, Prasad NR (2009) Providing strong security and high privacy in low-cost RFID networks. In: International conference on security and privacy in mobile information and communication systems, Italy, pp 172–179
10. Tian Y, Chen G, Li J (2012) A new ultralightweight RFID authentication protocol with permutation. *IEEE Commun Lett* 16(5):702–705
11. Mujahid U, Najam-ul-Islam M (2015) Ali Shami M (2015) RCIA: a new ultralightweight RFID authentication protocol using recursive hash. *Int J Distr Sens Netw* 642180:8
12. Mujahid U, Najam-ul-Islam M, Jafri AR, Ali Shami M (2016) A new ultralightweight RFID mutual authentication protocol; SASI using recursive hash. *Int J Distr Sens Netw* 12(2):8971. <https://doi.org/10.1155/2016/9648971>
13. Ahmadian Z, Salmasizadeh M, Aref MR (2013) Desynchronization attack on RAPP ultralightweight authentication protocol. *Inform Process Lett* 113(7):205–209
14. Ahmadian Z, Salmasizadeh M et al (2013) Recursive linear and differential cryptanalysis of ultralightweight authentication protocols. *IEEE Trans Inform Forens Sec* 8(7):1140–1151
15. Safkhani M, Bagheri N (2016) Generalized desynchronization attack on UMAP: application to RCIA, KMAP, SLAP and SASI+ protocols. *Cryptology*
16. Hernandez-Castro JC, et al (2010) Cryptanalysis of the David-Prasad RFID ultralightweight authentication protocol. In: Workshop on RFID security and privacy, Turkey, pp 22–34
17. Mujahid U, Najam-ul-Islam M (2017) KMAP: a new ultralightweight RFID authentication protocol for passive low-cost tags. *Wirel Person Commun* 94:725–744
18. Shariq M et al (2022) ESRAS: an efficient and secure ultra-lightweight RFID authentication scheme for low-cost tags. *Comput Netw* 217:109360
19. Servati MR et al (2022) Cryptanalysis of two recent ultra-lightweight authentication protocols. *Mathematics* 10:4611. <https://doi.org/10.3390/math10234611>
20. Mujahid U, Najam-ul-islam M (2014) Ultralightweight cryptography for passive RFID systems. *Int J Commun Netw Inform Sec* 6(3):173–181

# Development and Implementation of a Scalable and Replicable Industrial Environment at Low Cost to Control an Industrial Process



Serpa-Andrade Luis , Mata-Quevedo Paul ,  
Guerrero-Vasquez Fernando , Garcia-Velez Roberto ,  
and Gonzalez-Gonzalez Sandro 

**Abstract** Automation systems allow the processes to be efficient; however, they are expensive, and an automation process is proposed at low cost, through the use of open-source tools considering all the variables necessary for its optimal operation, and whose results are intended to be replicable and scalable in different processes of an industrial plant. Of course, it contains the design, development, and implementation of three human–machine interfaces that facilitate the control and monitoring of each of the variables and states of the process. These interfaces are segmented as follows. The first one proposes the use of a chatbot that, on the one hand, allows remote control instructions to be sent to the process, while, on the other hand, it can receive information on the status of the process, either by request, or automatically, considering cases of alarms. The second interface is related to the use of a voice assistant allowing to connect with the user bidirectionally with the process through voice commands using natural language. The third interface is related to the use of augmented reality techniques to access process information. In general, the project seeks to present an alternative solution for the implementation of automated industrial processes, focusing its focus, on the one hand, on reducing costs without altering the characteristics of an industrial automatization, while on the other hand, it seeks to incorporate new interfaces, that seek to reduce operator–machine interaction times resulting in greater efficiency and autonomy of the process.

---

S.-A. Luis (✉) · G.-V. Fernando · G.-V. Roberto  
Universidad Politécnica Salesiana, Grupo de Investigacion en Hardware Embebido Aplicado  
GIHEA, 010102 Cuenca, Ecuador  
e-mail: [lserpa@ups.edu.ec](mailto:lserpa@ups.edu.ec)

G.-V. Roberto  
e-mail: [rgarcia@ups.edu.ec](mailto:rgarcia@ups.edu.ec)

M.-Q. Paul  
Universidad Católica de Cuenca, 030101 Azogues, Ecuador

G.-G. Sandro  
Independent Researcher, 030102 Azogues, Ecuador



**Keywords** Industrial process · Low-cost tools · Human–machine interface · Chatbot · Commands voice

## 1 Introduction

The industry has evolved from the first phase in 1960, where the process machinery was controlled by the operators, to the present, when there is teleoperation of complete companies. The programmable logic programmers are the functional devices for the management of machinery, and at first, they were only understood as programmable controllers and controlled the central process, being the step for the evolution of the industry. PLCs went a little further, and they were in charge of giving logic to the processes, that is, being able to control through sequences.

The automaton, being the evolution of the PLC, is more integral and is still used in some factories, previously pioneers in the evolution. First of all, it solves the main connection problem between the control devices and greatly reduces the rigidity of the process.

Before the third industrial revolution, the control structures of industrial systems were associated with wired electromechanical devices. Those early control systems consisted entirely of relays, requiring extensive wiring and the inevitable expense associated with their inflexibility. These circuits cannot be rearranged to accommodate other applications in the same field. Even the slightest change in the number or nature of the inputs and outputs requires a completely new circuit that can communicate correctly [1].

Over time, this electrical concept can cause breakdowns and make correct maintenance difficult because there is no standardized system for locating errors or ways of acting to solve them. A trained technician is the best solution to get these relay control systems up and running. These features are associated with high costs.

With the advent of PLCs, the versatility of control circuits is enormous. Capable of storing sequencing, timing, counting, arithmetic, data manipulation, and communication instructions to control machinery and industrial environments, these PLCs have revolutionized management strategy and mass production lines.

Many of the important characteristics that are now common to electronic equipment in factories, being these reliable and robust, generating a reduction in maintenance costs by generating programmable logic processes, from their sensor inputs to the control of their actuators [2].

Industry 4.0 is addressing solutions that converge the physical and digital domains in a cleaner way, without the complications of crude systems. To this is added the wireless networks that allow the development of the Industrial Internet of Things (IIoT), contribute to lower costs and promote fast and efficient data management through “quasi-intelligent” machines, creating industrial cyber-physical systems [3].

The field of low-cost automation is a necessary focus in the middle, a strategy that allows implementing the process engineering of our graduate engineers in order to

have an updated industry. Low-cost automation can be achieved with the ingenuity and use of relatively cheap hardware, without affecting performance [4].

Plant control decentralization methodologies should be proposed in low-consumption intelligent systems that are interconnected working as a whole and generating instant results by functional stages, generating efficiency in the review and management personnel of the industrial environment.

## 2 State of the Art

The control of industrial automation processes, in its different lines, is generally carried out through the use of screens, touch or control through buttons. Globally, the control of an industrial process seeks for the operator to have control of it in a safe and timely manner, even in the event of events unrelated to the nature of the industrial field. All these considerations have allowed industrial automation to bring with it a series of benefits both in production times and in operating costs. On the other hand, new digital monitoring solutions integrating hardware and software allow determining and configuring states and variables in real time. In this sense, a brief review of the work is carried out in this area.

The basic PLCs allow the generation of instructions, sequences, timing, manipulation, and communication to control machines and processes, the fundamental characteristic is the robustness of the equipment, the implemented logic replaced traditional electromechanical systems, and now, new scopes have been implemented to achieve interconnectivity [5].

Wireless networks and the industrial internet of things allow more economical and powerful processes, processing data through accessible and intelligent machines in the industrial field in cohesion with management and control data, taking into account that high-end PLCs have these peripherals, but the low-end or cheap ones do not have these accesses [6].

Automation takes a different approach when it calls itself low cost, it becomes an engineering solution, a use of immediate environmental resources to generate a device capable of being efficient, robust and comparable with others on the market, if we review the characteristics of automata. At a high level, we would have to imitate some functionalities under an open-source software that actually exists and has already been worked on by the current research community in the area [7].

Solving the need for low cost in the automation of industrial processes, the use of intelligent devices, interlocutors for the decentralization of industrial environments is the best option, allowing the flow of knowledge of the industrial plant through different communication routes to the supervisors, and this is being important data for future maintenance and timely operation of the same [8].

### 3 Proposal

It is proposed to generate an embedded system that allows a low-cost device to have additional connectivity and plug-in characteristics to be applied in the industry.

We start from the premise that the brain is capable of being programmable and its input and output interface has the corresponding protections to work in an industrial environment.

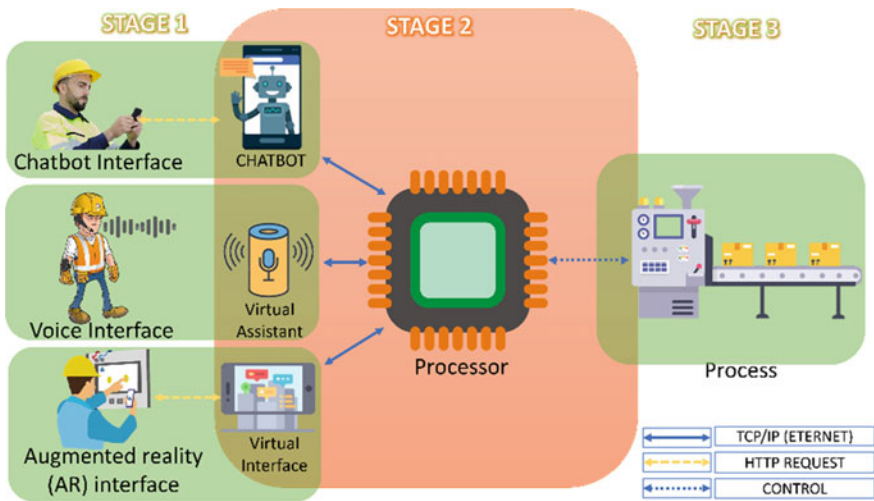
It supports voice command applications, chat command recognition, and an interface to review the main variables, previously chosen, with augmented reality (see Fig. 1).

The developed environment is multiplatform and is compatible to work in any PLC that has an Ethernet connection port available and the Modbus TCP/IP protocol.

#### 3.1 Chatbot Interface

The literature has been reviewed allowing us to observe that there are systems that take on the challenge of working with chatbots, one of the proposals and premises of a modern industry. The improvements are immediately observable with the use in an environment of preparation for the use of machinery executed by new employees for example, generating to the company greater profits and better understanding of the machine language of the environment or production process [9].

Our objective is to generate a database of variables with which we will work for queries from Telegram Messenger using @BotFather, managing to interact with a



**Fig. 1** Figure system proposal

specific script, responding to queries about the status of the machines and processes, allowing optimization of resources in regarding the entry of menus for a quick consultation of the operators or the supervision personnel.

Node-Red is used as an interlocutor between the software and the hardware or embedded system to be implemented, and this platform is being in charge of training the GPIOs to be programmed, whether analog or digital [10].

### **3.2 Voice Interface**

The automation paradigms and their advances are put into play when a virtual assistant is implemented, for example, it can be done, studies show it, being strictly necessary to satisfy the demands of the operators but above all generating robust systems [11].

In our case, the creation of “skills” for the virtual assistant is not limited to a specific syntax as long as these are related to the variable or actuator to be controlled, so that it is common domain for different operators.

Although the use of virtual assistants is oriented toward auditory-uncontaminated environments, they were used in the industrial environment, with the consideration that, according to validation tests, they can be used for higher levels within the automation pyramid, in auditory quiet environments; however, this does not limit that they can be used at lower levels in conjunction with other human-machine interfaces [12].

### **3.3 Augmented Reality**

Although the state-of-the-art review suggests that the industry is not yet a robust platform for AR implementation, [13] argues that it could not be effectively introduced into the industry to support tasks such as maintenance, training, user manuals, collaborative design, rapid visualization of product changes, and modifications. It is intended to implement a parameter monitoring solution in a first stage through an intelligent device, be it a cell phone, tablet, or similar.

## **4 Conclusions**

In Ecuador, unfortunately, very little is invested in the generation of digital skills and competencies. There is a very large gap created by the insufficient number of technological professionals, the vast majority of industrial automation is comprised commercial brand programmable controllers recognized, and this leads to the high

cost of automating a plant, which is why this project proposes the use of a low-cost programmable automaton that is capable of supporting the implementation of chatbot-type virtual assistants and control by voice commands, in addition to offer information in real time through augmented reality. This is intended to provide an alternative option for automation, with low cost and replicable to various types of industrial plants.

The control of industrial automation processes, in its different lines, is generally carried out through the use of screens, touch or control through buttons. Globally, the control of an industrial process seeks for the operator to have control of it in a safe and timely manner, even in the event of events unrelated to the nature of the industrial field. All these considerations have allowed industrial automation to bring with it a series of benefits both in production times and in operating costs. On the other hand, new digital monitoring solutions integrating hardware and software allow determining and configuring states and variables in real time. In this sense, a brief review of the work carried out in this area is carried out.

## References

1. Taalbi J (2019) Origins and pathways of innovation in the third industrial revolution. *Ind Corp Change* 28(5):1125–1148
2. Da Fonseca RHH, Pinto FR (2019) The importance of the programmable logic controller “PLC” in the industry in the automation process
3. Mellado J, Núñez F (2022) Design of an IoT-PLC: a containerized programmable logical controller for the industry 4.0. *J Ind Inform Integr* 25:100250
4. Aghenta LO, Iqbal MT (2019) Low-cost, open source IoT-based SCADA system design using thinger: IO and ESP32 thing. *Electronics* 8(8):822
5. Vishnu Easwaran E, Tigadi N, Chipkar A, Akshai M, Kushalkar R, Moudgalya KM, Zoit A, Alves T (2018) Proceedings, IECON 2018—44th annual conference of the IEEE industrial electronics society: Omni Shoreham Hotel, Washington DC, USA, 20–23 October, 2018, ISBN 9781509066841
6. Vieira G, Barbosa J, Leitão P, Sakurada L (2020) Proceedings, 2020 IEEE international conference on industrial technology: buenos aires institute of technology (ITBA), Buenos Aires, Argentina, 26–28 February, 2020, ISBN 9781728157542
7. Vieira G, Leitão P, Barbosa J, Ulisses M, Bragança G (2018) Low-cost industrial controller based on the Raspberry Pi platform. Bragança
8. Krupa P, Limon D, Alamo T (2018) Implementation of model predictive controllers in programmable logic controllers using IEC 61131-3 standard. 16. Aufl., ISBN 9783952426999
9. Casillo M et al (2021) A chatbot for training employees in industry 4.0. The international research and innovation forum. Springer, Cham, pp 397–409
10. Gonzalez-Gonzalez S, Serpa-Andrade L (2022) Development of a virtual assistant chatbot based on artificial intelligence to control and supervise a process of 4 tanks which are interconnected. In: Ahram T, Kalra J, Karwowski W (eds) Artificial intelligence and social computing. AHFE international conference. AHFE open access, vol 28. AHFE International, New York. <https://doi.org/10.54941/ahfe1001464>
11. Gärtler M, Schmidt B (2021) Practical challenges of virtual assistants and voice interfaces in industrial applications. In: Proceedings of the 54th Hawaii international conference on system sciences, p 4063

12. Arias J, Serpa-Andrade L (2022) Implementation of an IoT-based environment to control an industrial process by voice commands using a virtual assistant. In: Mrugalska B (eds) Production management and process control. AHFE international conference. AHFE open access, vol 36. AHFE International, New York. <https://doi.org/10.54941/ahfe1001629>
13. Santi GM et al (2021) Augmented reality in industry 40 and future innovation programs. *Technologies* 9(2):33

# Graph Embedding of Chronic Myeloid Leukaemia K562 Cells Gene Network Reveals a Hyperbolic Latent Geometry



Paola Lecca, Angela Re, Giulia Lombardi, Roberta Valeria Latorre, and Claudio Sorio

**Abstract** Recent research emphasises the significance of identifying the latent geometry—also known as the geometry underlying a complex network—which is determined by the manifold class, curvature, and dimension. Geometry’s explanation of the organisational principles of complex systems and the potential that network nodes might be categorised according to their distances in this geometry are what confers geometry its significance. In this study, we analysed the network of genes resulting differentially expressed as a consequence of the modulation of protein tyrosine phosphatase receptor type G (PTPRG) in chronic myeloid leukaemia (CML) cell model K562 transcriptome. We found that the latent geometry of this network is hyperbolic, and that clustering according to the angular coordinates of its nodes in hyperbolic space classifies genes by functional pathway classes.

---

P. Lecca (✉)

Faculty of Computer Science, Smart Data Factory Laboratory, Free University of Bozen-Bolzano, Bolzano, Italy  
e-mail: [Paola.Lecca@unibz.it](mailto:Paola.Lecca@unibz.it)

Member of National Group for Mathematical Analysis, Probability and their Applications, Francesco Severi National Institute of High Mathematics, Rome, Italy

A. Re

Department of Applied Science and Technology, Politecnico di Torino, Turin, Italy  
e-mail: [angela.re@polito.it](mailto:angela.re@polito.it)

G. Lombardi

Department of Mathematics, University of Trento, Trento, Italy  
e-mail: [giulia.lombardi@unitn.it](mailto:giulia.lombardi@unitn.it)

R. V. Latorre · C. Sorio

Department of Medicine, University of Verona, Verona, Italy  
e-mail: [robertavaleria.latorre@univr.it](mailto:robertavaleria.latorre@univr.it)

C. Sorio

e-mail: [claudio.sorio@univr.it](mailto:claudio.sorio@univr.it)

**Keywords** Graph embedding · Computational geometry · Hyperbolic geometry · Biological networks · Chronic myeloid leukaemia · Gene network

## 1 Introduction

Systems of interacting entities are frequently represented as networks with nodes and edges embedded in space, and networks are often modelled as graphs. The Internet, social, IT, and neural networks are only a few instances where space is crucial because topology alone does not fully encapsulate the totality of information. Thus, characterising and understanding the structure and development of networks is crucial for many diverse fields, such as urbanisation, society, IT networks, biological networks, ecological networks, and epidemiology [4, 14, 19, 20]. The cost of edge length on networks is a significant effect of space, and this cost has a significant impact on the topology of these networks.

A network's latent geometry and its formation and evolution are intertwined [16, 21]. Indeed, a network is born and evolves in response to a combination of two principles, the *popularity* principle and the *similarity* principle. Preferential attachment, as commonly alleged to explain scaling establishment in developing networks, is based on the notion that 'popularity is desirable'. If new connections are created preferentially to more well-liked nodes, as seen in many real networks, the distribution of the number of connections that nodes possess will follow power laws. Preferential attachment can result from a variety of mechanisms based on fitness of nodes, ranking of nodes, optimization, duplication, or random walks [15], and it has been explicitly demonstrated in some real-world networks.

Related nodes are more likely to link, even if they are not popular. In the social sciences, this tendency is referred to as *homophily* and has been observed in many real networks. In addition to well-known Websites like Google or Facebook, someone creating a new homepage on the Internet could link to lesser-known Websites that are related to her/his own hobbies [15]. These observations suggest that network formation is the result of the balanced action of two principles: the popularity principle and the similarity principle [15]. Imagine a network being formed through the progressive addition of nodes. If nodes enter the network one at a time, the node entry time is simply the index of the entrance order  $s = 1, 2, \dots, N$ . The node entry time is the most direct indicator of popularity. If all else is equal, older nodes have a higher chance of becoming attractive and attracting connections. Now, let us imagine placing at random nodes on a circle. In this experiment, circle represents the most basic similarity space. Angular distances between nodes in this space simulate their differences in terms of similarity.

Creating new connections that optimise the product between the entrance order and angular distance of nodes is a method for modelling a balance between popularity and similarity. This is the model. At first, the network is empty. At time  $s > 1$ , new node  $s$  is placed at position  $\theta_s$  on the circle. The node  $s$  connects to a subset of existing nodes  $t$  (with  $t < s$ ) consisting of the  $m$  nodes with the smallest  $m$  values of



the product  $s\theta_{ts}$ , where  $\theta_{ts}$  is the angular distance between node  $t$  and node  $s$ , and  $m$  is the half of the average node degree according to [15]. This network growth model can be translated geometrically as follows. All nodes reside on a plane, rather than a circle, with their polar coordinates being  $(r_s, \theta_s)$ , which are obtained by converting their birth time  $s$  to their radial coordinate  $r_s$  using the formula  $r_s = \ln s$ . Then, it emerges that new nodes simply connect to the nearest  $m$  nodes on the plane, with distances that are not Euclidean but rather hyperbolic [7, 15]. In hyperbolic space, the distance of two nodes of polar coordinates  $(r_{s_1}, \theta_{s_1})$  and  $(r_{s_2}, \theta_{s_2})$  is

$$d_{12} = r_{s_1} + r_{s_2} + \ln \left( \frac{\theta_{12}}{2} \right) = \ln \left( s_1 s_2 \frac{\theta_{12}}{2} \right) \quad (1)$$

where  $\theta_{12}$  is the convex angle subtended by the arc of the circumference bounded by the positions of the two nodes on the circumference. According to this calculation, nodes classified as similar due to the fact that their distance is small are also nodes whose angular coordinates do not differ too much from each other. *The above explains why in a hyperbolic disk the radial coordinates of the nodes indicate the network degree heterogeneity, and the angular coordinates their similarity.* Furthermore, formula (1) says that the hyperbolic distance is a way to combine the radial popularity and angular similarity into a single metric.

In a network with hyperbolic latent geometry, heterogeneous degree distribution and heightened clustering naturally appear as straightforward reflections of the negative curvature and metric property of this geometry. *Vice versa*, a network has a hyperbolic geometry if it has a metric structure and a heterogeneous degree distribution [9].

Although the theory contained in this paper can be applied to the study of networks of any type that exhibit hyperbolic latent geometry, we focus here on the study of a biological network of considerable interest for the medical applications that its knowledge may entail: the gene network of differentially expressed genes in chronic myeloid leukaemia that has been investigated by some authors of this paper in the previous study [13]. Biological networks frequently display hierarchical tree-like organisation at various scales of biological organisation, and in this paper, we first show that this network also has a hyperbolic latent geometry. We then show that clustering according to the angular coordinate of nodes embedded in a hyperbolic space is equivalent to clustering according to the class of functional pathways.

The classification of the nodes of a network according to angular coordinates, which in hyperbolic geometry is equivalent to the classification according to similarity, is a particularly important tool for understanding the structure of a network and its modularity when there is no a priori information on the function of the nodes and/or their properties. For this reason, the analysis we present in this article we think may be of great interest and utility in applications in the biological field (e.g. to infer new interactions), where the complexity of the systems studied and the difficulties of collecting complete experimental data often compromise the reliability of statistical and mathematical analyses. The validation of this conjecture requires

further experiments and analyses, which are the future research directions opened up by this work.

## 2 Hyperbolic Geometry

Here, we introduce some fundamental concepts of hyperbolic geometry that are useful for understanding the methods and analyses described in this work.

A hyperbolic space is a non-Euclidean geometry with a negative constant sectional curvature, commonly known as Lobachevsky-Bolyai-Gauss geometry. This geometry satisfies all of Euclid's postulates except for the parallel postulate, which is changed to read there exist numerous more indefinitely extending straight lines that pass through  $P$  but do not intersect  $L$  for every infinite straight line  $L$  and any point  $P$  that is not on it.

In hyperbolic geometry, triangles with the same angles have the same areas, and their total angles are less than 180 degrees. Additionally, not every triangle has the same angle sum (cf. the AAA similarity theorem for triangles in bi-dimensional Euclidean space). Hyperbolic geometry does not have any triangles that are similar. The spheres in Lorentzian four-space are the most well-known illustration of a hyperbolic space. The metric of the Lorentz space

$$ds^2 = -dx_0^2 + dx_1^2 + dx_2^2,$$

in  $\mathbb{R}^3$  is associated to  $F((x_0, x_1, x_2)) := -x_0^2 + x_1^2 + x_2^2$ . The norm of a vector  $\mathbf{x}$  is  $(F(\mathbf{x}))^{1/2}$ . Consider the set

$$H = \{\mathbf{x} : F(\mathbf{x}) = -1\}$$

that is the sphere of unitary radius centred at the origin. According to the sign of  $x_0$ , there are two components to the hyperbolic shape  $H$ : the upper sheet  $H^+$  where  $x_0 > 0$ , and the lower sheet  $H^-$  where  $x_0 < 0$ . The distance in  $H^+$  between two points  $\mathbf{x}, \mathbf{y}$  is

$$d(\mathbf{x}, \mathbf{y}) = \operatorname{arccosh}(-F(\mathbf{x}, \mathbf{y})).$$

A circle with radius  $\rho$  has a circumference of hyperbolic length

$$\int_0^{2\pi} \frac{2r}{1-r^2} d\theta = \frac{4\pi r}{1-r^2}$$

with  $r = \tanh\left(\frac{\rho}{2}\right)$ . As  $1-r^2 = \frac{1}{\cosh^2\left(\frac{\rho}{2}\right)}$ ,

$$\text{Circumference} = 4\pi \sinh\left(\frac{\rho}{2}\right) \cosh\left(\frac{\rho}{2}\right) = 2\pi \sinh(\rho).$$

The hyperbolic area of the circle is

$$\int_{t=0}^r \int_0^{2\pi} \frac{4t}{(1-t^2)^2} dt d\theta = 8\pi \left[ (2(1-t^2))^{-1} \right]_0^r = \frac{4\pi r^2}{1-r^2}. \quad (2)$$

So

$$\text{Area} = 4\pi \sinh^2\left(\frac{\rho}{2}\right), \quad (3)$$

and therefore, the volume of a hyperbolic sphere is

$$\text{Volume} = 2\pi (\sinh(\rho) - \rho). \quad (4)$$

A representation of  $n$ -dimensional hyperbolic geometry is the Poincaré ball in  $n$ -dimensions  $\mathbb{B}_{\mathbb{R}}^n$  ( $\mathbb{B}_{\mathbb{R}}^n = \{\mathbf{x} \mid \|\mathbf{x}\|^2 < 1\}$ ) in which the diameters of circles or the arcs of circles with end perpendicular to the boundary of the ball serve as the lines. In two dimensions, the Poincaré space is an open disc. The distance between two points  $z_1, z_2 \in \mathbb{C}$  in this space is

$$H_2(z_1, z_2) = \log(|1 - \overline{z_1}z_2| + |z_2 - z_1|) - \log(|1 - \overline{z_1}z_2| - |z_2 - z_1|). \quad (5)$$

and the distance of a point  $z$  from the origin is

$$H_2(0, z) = \log(1 + |z|) - \log(1 - |z|). \quad (6)$$

### 3 Data Collection and Gene Network Construction

In this study on human K562 chronic myeloid leukaemia clones, we consider the genes resulting differentially expressed between the untreated group (empty vector and inactive mutant domain D1028A) and the treatment group expressing full-length PTPRG [17]. The RNAs hybridization on Agilent whole human genome oligo microarray and the statistical methods implementing the differential expression analysis are described in Lombardi et al. [13]. Interactions amongst genes were inspected by accessing Pathway Commons' archives [1, 3]. In the case study under investigation, we focussed on genetic interactions repositories with the aim of building an interaction network for differentially expressed genes identified by the previous analyses. More in detail, we accessed the PathwayCommons.All.hgnc repository, which contains all the interactions retrieved by the multiple data sources accessed by Pathway Commons [2]. Therefore, we queried the table to identify interactions for the genes of interest. The investigation provided a match of 335 genes out of 384 and

thereby produced the required input data to pursue interactions network analyses. The network thus reconstructed turned out to have 7286 nodes and 18722 edges and to be a connected graph.

## 4 Graph Embedding

In order to establish the geometry of the network, we immersed the network in three metric spaces: Euclidean, hyperbolic, and spherical and calculated the distortion (usually called *embedding stress*) caused by each of these embeddings. The embedding with the smallest distortion value corresponds to the embedding in the optimal latent geometry for the network. Lecca [11], and Lecca et al. [12] developed mathematical models of embeddings in the three spaces. In the next subsection, we summarise what is reported in [12].

We implemented a Python code for graph embedding that is publicly available in Lecca et al. [12].

### 4.1 Embedding in Euclidean Space

The matrix  $U = [u_1 u_2 \dots u_m]$  of the graph Laplacian eigenvectors provides the embedding in a Euclidean space of dimension  $m$  ( $U$  which is a  $m \times m$  matrix). According to increasing values of the respective eigenvalues, the eigenvectors are arranged in ascending order in a matrix, whose  $i$ -th row defines the coordinates of the node  $v_i$  in Euclidean space.

### 4.2 Embedding in Constant Curvature Manifolds

We calculated the spectral decomposition of the matrix

$$C_{ij}^{U,k} = \cos(\sqrt{k}d_{ij}). \quad (7)$$

where  $\mathbb{U}$  is a space of dimension  $n$ ,  $d$  is a function such that  $d : U \times U \rightarrow \mathbb{R}^+$ ,  $D = \{d_{ij}\} = \{d(u_i, u_j)\}$  denotes the node-to-node distance matrix, and  $k$  the curvature of the space.  $C^{U,k}$  is a  $n \times n$  matrix.

We also implemented the criteria for determining the possibility of an isometric embedding. In this regard, Blumenthal [6] and Schoenberg [18] are credited with a theorem defining the requirements for the embedding to be isometric in a space with constant curvature. This Theorem reported and commented in details in Lecca et al. [12] says that if  $k < 0$ , the space defined by  $U$  can be isometrically embedded in  $\mathbb{H}_k^m$  if and only if the number of positive, negative, and zero eigenvalues of  $C^{U,k}$  is 1,  $p$  and

$n - p - 1$  respectively, where  $qp \leq m$ . In the case of  $k < 0$ , the space defined by  $U$  can be isometrically embedded in  $\mathbb{H}_k^m$  if and only if the number of positive, negative, and zero eigenvalues of  $C^{U,k}$  is  $p$ ,  $0$ , and  $n - p$ , respectively, with  $p \leq m + 1$ . The most current proof of this theorem may be found in [5]. In accordance with Begelfor et al. [5], if the curvature is negative, the coordinates of node  $i$  are given by

$$v_i = \frac{1}{\sqrt{1 - ||w_i||^2}} U_m \sqrt{-\Sigma_m} \quad (8)$$

where  $\Sigma_m$  is the diagonal matrix of the  $m$  most negative eigenvalue of  $C^{U,k}$  and  $(w_1 \ w_2 \ \dots \ w_n)^T = U_m \sqrt{-\Sigma_m}$ .

If the curvature is positive, the  $i$ -th coordinates are

$$v_i = \frac{1}{||w_i||} U \sqrt{\Sigma}. \quad (9)$$

The  $C$  eigenvalue decomposition and the selection of the dominating  $m$  eigenvectors are used if the conditions for an isometric embedding are not met.

### 4.3 Embedding Stress

Unlike in [12], in this study, the embedding stress is defined as

$$\text{Stress} = \frac{1}{\Xi} \sqrt{\sum_{ij} (d_{ij} - d_{ij}^*)^2}. \quad (10)$$

In formula (10), the graph's node count is  $\Xi$ , and the edge weight is  $d_{ij}$ ,  $d_{ij}^*$  is the edge's weight in the space in which the graph has been embedded.

### 4.4 Assignment of Edge Weights to Unweighted Network

Real networks have a significant variability in the capacity and intensity of the connections along with a complicated topological structure. In the cases in which there are no experimental data quantifying the intensity of a connection (i.e. the edge weight), edges are usually represented as binary states, and their weights are either 1 or 0. This is the case for the network considered in this study. Although the embedding procedure can also be applied to unweighted networks (in this case  $d_{ij} = 1$  if node  $i$  and node  $j$  are connected by an edge, and 0 otherwise), it is recommended to apply embedding methods to weighted networks (whenever weights are available)

in order to give the algorithm as a complete and realistic description of the network as possible.

In this analysis, we assign to the edge  $i-j$  a weight calculated as the average of the clustering coefficients of the node  $i$  and node  $j$ . Indeed, the clustering coefficient  $C_i$  of a node  $i$  is the proportion of pairs of neighbours of  $i$  that are connected by an edge. In formulas, if  $m_i$  is the number of pairs of neighbours of  $i$  that are connected, and  $k_i$  is the degree of node  $i$ , the clustering coefficient of node  $i$  is

$$C_i = \frac{2m_i}{k_i(k_i - 1)}, \quad (11)$$

where  $k_i(k_i - 1)/2$  is the number of possible pairs of neighbours of  $i$ . Then,  $C_i$  is the probability that a pair of neighbours of  $i$  are linked. Therefore, if no experimental measure is available for  $d_{ij}$ , we estimate it as

$$d_{ij} = \frac{C_i + C_j}{2}. \quad (12)$$

## 5 Results

We assigned edge weights to the CML gene networks according to the procedure in Sect. 4.4, and we embedded it in Euclidean and hyperbolic spaces (Poincaré model) according to the procedures in Sect. 3. The best embedding is the hyperbolic one, as it has a stress lower than the stress of Euclidean embedding. This result is expected since the graph of the network is connected and consequently can be well approximated by a tree [10, 11]. Using formula (10) with  $\Xi = 7286$ , we found that

$$\text{Stress}_H = (0.01353399 \pm 10^{-9}), \text{ and } \text{Stress}_E = (0.06912648 \pm 10^{-9})$$

where  $\text{Stress}_H$  and  $\text{Stress}_E$  denote the hyperbolic and Euclidean embedding stress, respectively. The stress of Euclidean embedding is five times greater than the stress of hyperbolic embedding.

We then proceeded to cluster the nodes according to the angular coordinate on the Poincaré disk, i.e. the input to the clustering procedure is the array of nodes' angular coordinates. We used R's NbCLust function [8] with the 'McQuitty' aggregation method and Euclidean distance as a measure of dissimilarity. NbCLust automatically determines the optimal number of clusters in a range predefined by the user and taking into account 30 validity indices. In the McQuitty method, the distance between clusters  $CL_i$  and  $CL_j$  is the mean of the between cluster dissimilarities:  $\text{Diss}_{ij} = (\text{Diss}_{ik} + \text{Diss}_{il})/2$  where cluster  $CL_j$  is formed from the aggregation of clusters

**Table 1** Clusters according to the angular coordinate in 2D hyperbolic space

Graph	$x$	$y$	Area	Size
Whole network	$-2.140\text{E}-05$	$-4.070\text{E}-06$	141.9	7286
Cluster 1	$2.10\text{E}-03$	$-4.00\text{E}-04$	24.61	2476
Cluster 2	$-3.60\text{E}-03$	$7.00\text{E}-03$	9.897	1476
Cluster 3	$2.90\text{E}-03$	$1.00\text{E}-04$	1.022	2600
Cluster 4	$-0.011$	$-0.01$	3.4910	734

Baricenters, area, and nr. of nodes (size) of the clusters and whole network

$CL_k$  and  $CL_l$ . We obtained that the optimal number of clusters of the nodes’ angular coordinate values is 4. The coordinates of the centres of gravity of the clusters, clusters’ areas, and the number of nodes contained in each cluster are shown in Table 1.

The area has been calculated as the sum of areas of circles centred on node, where circle radii have been determined according to a simplified circle packing scheme. The area of the circle has been calculated on the hyperbolic space with formula (3). It provides a measure of distribution and part-to-whole relationships in the data. The results in Table 1 state that the cluster most distinct from the others is Cluster 4.

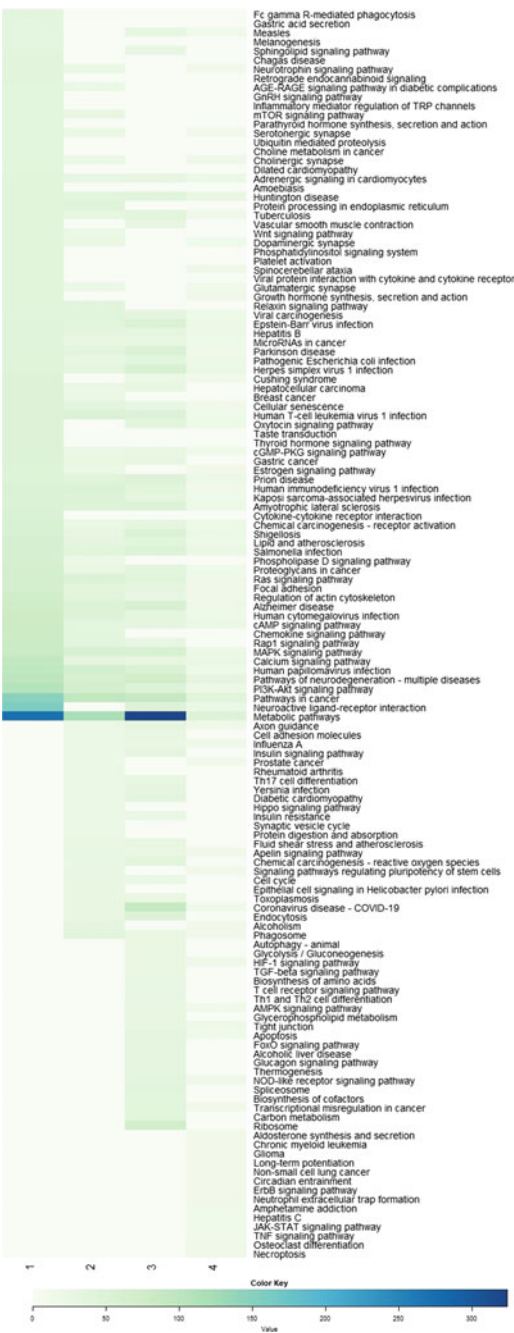
Of the nodes in each cluster, we performed functional analysis. The heatmap in Fig. 1 shows that in fact the aggregation according to angular coordinate mirrors *to a certain extent* the aggregation according to pathway/functional process. We would like to point out here that functional analysis classifies a gene into macro-sectors that in turn can include multiple and specific functions. This fact has to be taken into account in evaluating the clustering result we see in Fig. 1 and in recognising its limitations. In spite of this, we nevertheless note an aggregation by functional pathways and by the proportion with which these pathways are found within each cluster.

We found that the most frequent pathway in each cluster is ‘Metabolic pathway’. In order to understand whether, despite this characteristic shared by all four clusters and despite different sizes and areas shown by the clusters, there were significant differences in the proportion of ‘Metabolic pathways’ in each cluster, we performed a proportion test, whose results are shown in Table 2.

The test confirms a statistically significant difference between the proportion of ‘Metabolic pathways’ between clusters 1, 2, 3, and the cluster 4, whilst accepting the null hypothesis regarding the proportion of metabolic pathways in the mutual comparisons between clusters 1, 2, and 3.

Finally, we focussed on five genes identified as differentially expressed between the PTPRG modulated gene group and in the control group in the previous work [13]: RARG ( $p$ -value = 0.001), CD36 ( $p$ -value = 0.1), MECP2 ( $p$ -value= 0.01), TFAP2C ( $p$ -value = 0.01), TRPS1 ( $p$ -value = 0.001), SMAD1 ( $p$ -value = 0.0001). We found that RARG, CD36, and MECP2 belong to Clusters 1, 1, and 3, respectively, whilst TFAP2C, TRPS1, and SMAD1 all belong to Cluster 4. We note that the genes with marked statistical significance of the difference in expression are found in Cluster 4.

**Fig. 1** Heatmap showing the clustering according to functional pathways/processes as obtained from the clustering by angular coordinates of the genes on Poincaré disk. On the x-axis, the numeric ID of the clusters is shown





**Table 2** Proportion test on the proportion of the most frequent function/pathway in the four clusters found according to the angular coordinate

Samples (A, B)	Z statistics	$p_A$	$p_B$	$n_A$	$n_B$	p-value
Cluster 1, Cluster 4	2.82E+00	3.22E−02	2.09E−02	8.48E+03	2.25E+03	4.86E−03
Cluster 2, Cluster 4	9.84E−01	2.46E−02	2.09E−02	4.91E+03	2.25E+03	3.25E−01
Cluster 3, Cluster 4	5.89E+00	4.99E−02	2.09E−02	6.49E+03	2.25E+03	3.75E−09
Cluster 1, Cluster 2	2.50E+00	3.22E−02	2.46E−02	8.48E+03	4.91E+03	1.25E−02
Cluster 1, Cluster 3	−5.48	0.032	0.050	8477	6494	1
Cluster 2, Cluster 3	−6.90	0.025	0.050	4911	6494	1

‘Metabolic pathways’ is the most frequent item in all four clusters, but there is a significant statistical difference between the proportion of ‘Metabolic pathways’ only between cluster 4 and the remaining three. No statistical significance about the difference has been found by comparing Clusters 1 with Cluster 3 and Cluster 2 with Cluster 3

In view of this result and more generally of the results of this analysis, we propose the following conjecture: the angular coordinate is a generalisation of the concept of similarity and can account for various aspects of it, such as belonging to the same functional pathway and response to stimuli. This second aspect would prove to be of considerable importance for all analyses concerning the dynamics of a network and its interactions with its environment.

5.1 Remarks

The limitations and advantages of applying the method of classifying nodes according to the angular coordinate in a hyperbolic space should be identified in the light of the definition of the concept of similarity. Mathematically, the similarity of nodes in a graph is defined as an equivalence class defined by an equivalence relation on the topological structure of the node-induced graph. The applicability of the clustering method according to angular coordinate is highly dependent on the equivalence class with which the physical concept of similarity is abstracted. From the physical or biological point of view, similarity between nodes is expressed by various physical parameters whose nature depends on the type of network under consideration (e.g. chemical affinity between reagents in a biochemic network, degree of co-expression in a gene network, connection probability in a social network, etc.). An understanding of similarity that cannot be abstracted with a definition of equivalence corresponding to structural similarity may not be suitable to be revealed by clustering by angular

coordinate. Provided the abstraction suitability is met, the proposed method can be used for time-dependent networks and also on incomplete networks (within a certain level of incompleteness of nodes and arcs).

## 6 Conclusions

In this study, we have shown that the network of differentially expressed genes between the modulation condition of PTPRG, and the control condition has latent hyperbolic geometry and that the clustering according to angular coordinates of the nodes (genes) on a Poincaré space reflects the clustering of the genes according to function/process. We have also found that the genes for which the statistical test decision is sharper (p-values below  $10^{-3}$ ) belong to the most isolated cluster.

This result suggests the conjecture that in networks with hyperbolic geometry, there is a bi-univocal correspondence between the gene's functional macro-sector, the intensity of the response to network perturbations (obtained in this case by modulating PTPRG), and the range of values of the angular position coordinate. The importance of this conjecture lies in the fact that in networks with a hyperbolic geometry, and of whose nodes there is little knowledge in terms of function and response to stimuli, clustering according to the angular coordinate can provide a first important clue as to the existence of functional modules and the reactivity of the nodes.

**Author Contribution and Acknowledgements** Paola Lecca conceptualised the research, developed the mathematical graph embedding models, and implemented them in Python and R. Angela Re review the existing literature, tested the software, analysed the outputs, and dealt with data visualisation. Giulia Lombardi dealt with the reconstruction of the gene network, and the functional analysis of genes grouped according to angular coordinates in hyperbolic space. Roberta V. Latorre and Claudio Sorio led the experimental activity for raw data collection and preparation. All authors contributed to writing and revising the paper.

## References

1. BioPAX Homepage (2022). <http://www.biopax.org/>. Accessed: 01 Dec 2022
2. Pathway Commons data sources archive (2022). <https://www.pathwaycommons.org/archives/PC2/v13/datasources.txt>. Accessed: 01 Dec 2022
3. Pathway Commons Homepage (2022). <http://www.pathwaycommons.org/>. Accessed: 12-01 Dec 2022
4. Alanis-Lobato G, Mier P, Andrade-Navarro M (2018) The latent geometry of the human protein interaction network. *Bioinformatics* 34(16):2826–2834 (Apr 2018). <https://doi.org/10.1093/bioinformatics/bty206>
5. Begelfor E, Werman M (2005) Learning curved manifolds the world is not always flat or learning curved manifolds (2005). <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.4022>

6. Blumenthal LM (1970) Theory and applications of distance geometry. Chelsea Publishing Company, 2 edn. (Jan 1970)
7. Bonahon F (2009) Low-dimensional geometry. American Mathematical Society, Providence, RI (Jul, Student mathematical library)
8. Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) Nbclust: an r package for determining the relevant number of clusters in a data set. *J Stat Softw* 61(6):1–36 (2014). <https://www.jstatsoft.org/index.php/jss/article/view/v061i06>
9. Krioukov D, Papadopoulos F, Kitsak M, Vahdat A, Boguñá M (2010) Hyperbolic geometry of complex networks. *Phys Rev E* 82(3): (Sep 2010). <https://doi.org/10.1103/physreve.82.036106>
10. Kurkofka J, Melcher R, Pitz M (2011) Approximating infinite graphs by normal trees. *J Comb Theory, Ser B* 148:173–183 (May 2021). <https://doi.org/10.1016/j.jctb.2020.12.007>
11. Lecca P (2023) Uncovering the geometry of protein interaction network: the case of SARS-CoV-2 protein interactome. In: Proceeding of the 11th international conference on mathematical modeling in physical sciences, 5–8 Sept 2022. AIP Conference Proceedings, In Press (2023)
12. Lecca P, Re A (2022) Checking for non-euclidean latent geometry of biological networks. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE (Dec 2022). <https://doi.org/10.1109/bibm55620.2022.9995274>
13. Lombardi G, Latorre RV, Mosca A, Calvanese D, Tomasello L, Boni C, Ferracin M, Negrini M, Dewik NA, Yassin M, Ismail MA, Carpentieri B, Sorio C, Lecca P (2022) Gene expression landscape of chronic myeloid leukemia k562 cells overexpressing the tumor suppressor gene PTPRG. *Int J Mol Sci* 23(17):9899 (Aug 2022). <https://doi.org/10.3390/ijms23179899>
14. Michielan R, Litvak N, Stegehuis C (2022) Detecting hyperbolic geometry in networks: why triangles are not enough. *Phys Rev E* 106:054303 (Nov 2022). <https://link.aps.org/doi/10.1103/PhysRevE.106.054303>
15. Papadopoulos F, Kitsak M, Serrano MÁ, Boguñá M, Krioukov D (2012) Popularity versus similarity in growing networks. *Nature* 489(7417):537–540 (Sep 2012). <https://doi.org/10.1038/nature11459>
16. Papadopoulos F, Krioukov D, Boguna M, Vahdat A (2010) Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In: 2010 Proceedings IEEE INFOCOM. IEEE (Mar 2010). <https://doi.org/10.1109/infcom.2010.5462131>
17. Peruta MD, Martinelli G, Moratti E, Pintani D, Vezzalini M, Mafficini A, Grafone T, Iacobucci I, Soverini S, Murineddu M, Vinante F, Tecchio C, Piras G, Gabbas A, Monne M, Sorio C (2010) Protein tyrosine phosphatase receptor type  $\gamma$  is a functional tumor suppressor gene specifically downregulated in chronic myeloid leukemia. *Cancer Res* 70(21), 8896–8906 (Oct 2010). <https://doi.org/10.1158/0008-5472.can-10-0258>
18. Schoenberg IJ (1935) Remarks to Maurice Frechet’s article “sur la definition axiomatique d’une classe d’espace distances vectoriellement applicable sur l’espace de hilbert. *Annal Math* 36(3):724 (Jul 1935). <https://doi.org/10.2307/1968654>
19. Sharpee T (2022) Hyperbolic geometry in biological systems. *FASEB J* 36(S1) (May 2022). <https://doi.org/10.1096/fasebj.2022.36.s1.0i221>
20. Zhou Y, Sharpee TO (2021) Hyperbolic geometry of gene expression. *iScience* 24(3):102225 (2021). <https://www.sciencedirect.com/science/article/pii/S2589004221001930>
21. Zuev K, Boguñá M, Bianconi G, Krioukov D (2015) Emergence of soft communities from geometric preferential attachment. *Sci Rep* 5(1) (Apr 2015). <https://doi.org/10.1038/srep09421>

# Trend of M-Health Research in the Self-management of Chronic Illness: Bibliometric Analysis



Ade Komariah and Erna Rochmawati 

**Abstract** This article uses a bibliometric analysis method to explore the development and trend of research that investigates the use of m-health in self-management of patients with chronic illness. A literature search using Scopus was conducted. Keywords related to m-health, self-management, and chronic illness were used. Synonym mobile application was used. For chronic illness cancer, hypertension and COPD were used when searching the literature. Boolean phrase AND or OR was used to combine the keywords. Any studies published from 2011 to 2022 were included. An analysis conducted on 336 original articles and reviews revealed that most studies about m-health were published from 2019 to 2021. The countries that produce the most articles in this field are the United States (US), the United Kingdom (UK), and are followed by Australia. Network analysis based on keyword cooccurrence reveals three main keyword types, namely self-care, mobile application, and self-management. Based on studies conducted, it is shown that m-health can help improve self-management in person with chronic illness that can lead to patient's quality of life improvement.

**Keywords** M-health · Self-management · Chronic illness · Bibliometric analysis · Network analysis

---

A. Komariah

Master in Nursing Program, Universitas Muhammadiyah Yogyakarta, Tamantirto Kasihan Bantul, Indonesia

E. Rochmawati (✉)

Universitas Muhammadiyah Yogyakarta, Kasman Singodimejo Postgraduate Building, Level 2, Jl Brawijaya, Tamantirto Kasihan Bantul, Indonesia

e-mail: [erna.rochmawati@umy.ac.id](mailto:erna.rochmawati@umy.ac.id)

## 1 Introduction

The development of scientific and technological innovations in health services has caused real changes in health epidemiology that the increasing need for technology in the health sector encourages the development of innovations in health services [1]. Apart from the rapid development of technology, changes have also occurred in the trend of diseases, namely infectious diseases have decreased substantially and changed to non-communicable diseases with a wide variety of comorbidities, both physical and psychological [2]. The severity of chronic condition can be directly related to the behavior of a particular patient [3]. Chronic disease can be prevented from triggering comorbidities and can be managed through early detection and changing lifestyle including regular exercise, healthy diet, and adherence to therapy [4].

To live optimally with chronic illness, patients are required to have capability to independently manage the symptoms and conditions. In the management of chronic diseases, self-management is a recommended strategy [5]. Patient's self-management refers to ability in managing symptoms and lifestyle that can lead to an improvement in patient's health [6–8]. Self-management includes disease awareness, improved self-efficacy, and motivation [8]. Patients are better equipped to control their illness and condition when they have a greater understanding of the methods and abilities of self-management. Consequently, patient's self-management enables to assist patients in better health management and behavior alteration to enhance health status [9, 10].

Improving self-management as one of the most effective health intervention strategies continues to shift chronic illness treatment away from hospitalization. According to the previous studies, self-management among persons with chronic illness can promote health behavior, reduce the need of health services, and enhance patients' quality of life [11]. There are technological advancements that allow patients with chronic conditions to be monitored even when they are not in the hospital, including the ability to monitor patients via a mobile health application, generally known as m-health [12]. M-health is defined as the devices to assist practice by medical and public health. The devices include smart phone and other personal monitoring devices and digital assistants [13].

By enhancing the data available to clinicians, including patients in self-management, and minimizing comorbidities, m-health can be applied effectively to enhance patient's outcomes. The m-health program facilitates the monitoring and intervention of chronic illnesses via remote access [12]. The integration of cellular technology into illness services is one of the potential methods for providing efficient, cost effective, and patient-centered chronic care [1]. Smartphone technology enables preventive interventions and continuous monitoring with the aid of these devices.

There is a growing study conducted on the development and application of m-health for patient's self-management, however, limited evidence shows the trend of available research. Therefore, the review aimed to explore the development and trend

of studies that investigates the adoption of m-health in relation to self-management of chronic illness.

## 2 Methods

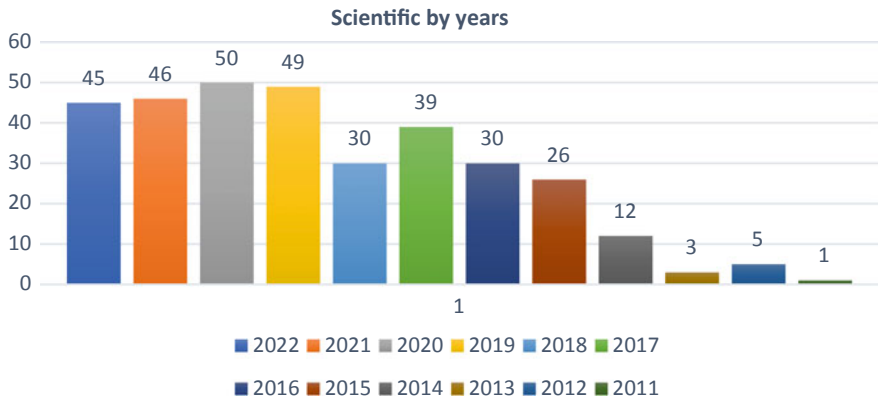
Scopus database was used to search available research. Keywords including m-health; self-management; and chronic illness, and their synonyms were used during the literature search. The searching statement using Boolean AND or OR is as follows: “m-health” OR “mobile-application”, AND “chronic illness using” OR “cancer” OR “hypertension” OR “Chronic Obstructive Pulmonary Disease”. The inclusion criteria were: (i) published between 2011 and 2022; (ii) written in English; (iii) research articles.

Data analysis used VOS viewer software to map the most prevalent and associated terms. All data is imported into VOS viewer v.1.6.18, which is commonly used to study and illustrate the relationship between year, author, country, and article terms. Publications are carefully categorized and evaluated by publication year, nation, journal, study field, author affiliation, and journal title. In addition, the frequency of keywords extracted from papers was evaluated and afterward incorporated into the network analysis of research development on m-health with chronic disease self-managers.

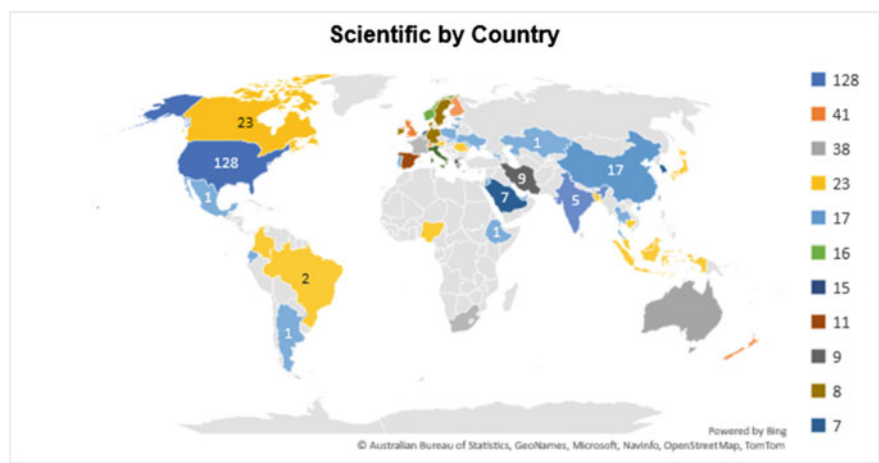
## 3 Results and Discussion

A total of 336 papers were published that focus on the application of mobile health and self-management in patients with chronic conditions. The database contains the most pertinent and recently released data from 2011 to 2022, with the maximum number of three publications occurring in 2019 ( $n = 49$ ), 2020 ( $n = 50$ ), and 2021 ( $n = 46$ ). In 2019, 2020, and 2021, researchers used technology (m-health) in the health sector to make it easier for health workers to maintain the stability of patients' conditions, so as not to be exposed to the virus during a pandemic that occurs around the world, when patients with chronic diseases are extremely susceptible to exposure to the virus. The evolution of m-health and self-management research in patients with chronic conditions by years can be observed in Fig 1, and the distribution of research on of m-health and self-management in patients with chronic conditions by countries can be seen in Fig 2.

Of the 60 countries that contributed in development of research on m-health and self-management among patients with chronic illness, it was explained that the countries that contributed actively in the development of research included the United States (US), United Kingdom (UK), Australia, Canada, China, Norway, Netherlands, Spain, Greece, Iran, Denmark, Germany, Ireland, Sweden, Saudi Arabia, and South Korea as shown in (Fig. 3). Based on the results of Biblioshiny, there are three



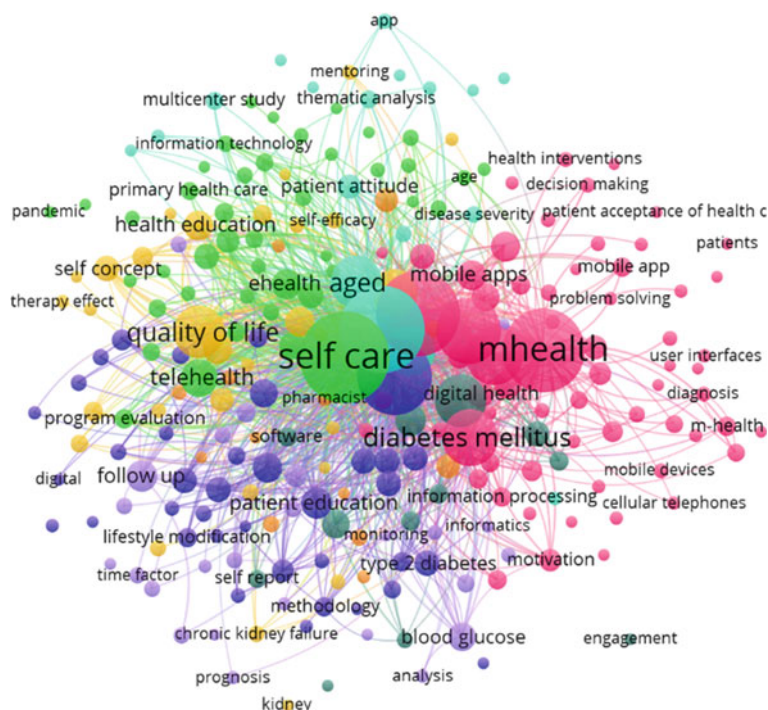
**Fig. 1** Investigation of e-health and self-management to chronic illness by years



**Fig. 2** Investigation of e-health and self-management to chronic illness by country

countries that contribute the most research, namely the United States (US) ( $n = 128$ ), the United Kingdom (UK) ( $n = 41$ ), and Australia ( $n = 38$ ).

Top five most global and cited publications (2019–2020–2021). This report analyzes global quotations from previously conducted studies. Table 1 displays the top five most referenced publications worldwide in the 2019. The number of paper citations retrieved from the database is a global source. There are three articles that have received the most citations in 2019; the most cited article was on the use of mobile health form older adult patients [14]. The second most cited study focused on smartphone applications and its use in the self-management of patients with type 2 diabetes mellitus, cited by 60 studies [15]. The third studies conducted



by Márquez Contreras et al. (2019) discussions related to specific hypertension was the third most cited article, cited by 40 publications.

Table 2 displays the five most cited publications in 2020, with three articles being most frequently cited: Liu et al. study (2020) that focused on the effectiveness of intervention using mobile app assisted to improve health outcomes in patients with type 2 diabetes mellitus (DM) and/or hypertension and a systematic review and meta-analysis of randomized controlled trials as many as 33 ( $n = 33$ ). In addition, research

Authors	Source title	Year	Cited
Wildenbos et al. (2019)	International journal of medical informatics	2019	73
Jeffrey et al. (2019)	Diabetology and metabolic syndrome	2019	60
Márquez Contreras et al. (2019)	Current medical research and opinion	2019	40
Scott et al. (2019)	Australian health review	2019	38
Wang et al. (2019)	BMC endocrine disorders	2019	35

Authors	Source title	Year	Cited
Wildenbos et al. (2019)	International journal of medical informatics	2019	73
Jeffrey et al. (2019)	Diabetology and metabolic syndrome	2019	60
Márquez Contreras et al. (2019)	Current medical research and opinion	2019	40
Scott et al. (2019)	Australian health review	2019	38
Wang et al. (2019)	BMC endocrine disorders	2019	35



on smartphone-based technology in type 2 DM management conducted by Doupis et al. has been cited 31 publications [16]. Smartphone-based technology in the DM management has 31 citations, followed by research conducted by Sittig et al. 2020 that has 28 citations [17].

Table 3 displays a large number of publications cited in 2021, with three dominant articles: the first article investigated the use of machine learning using interactive e-app to monitor patients with diabetes ( $n = 23$ ), the second most cited article was a systematic review that synthesized the use of smartphone technologies in disease monitoring ( $n = 14$ ), and the third most article was focused on the digital interventions for patients with chronic obstructive pulmonary (COPD) ( $n = 9$ ). There are nine examples of digital interventions for the therapy of chronic obstructive pulmonary disease.

Table 4 and Fig. 4 show the frequency of keywords use by authors who predominantly use the keywords “self-care” ( $n = 2,607$ ), “mobile application” ( $n = 2,267$ ), “self-management” ( $n = 1,884$ ), “Telemedicine” ( $n = 1,794$ ), and “m-health” ( $n = 1428$ ), which are in accordance with the purpose of writing to determine the effect of combining m-health with self-management in patients with chronic illness, which is in accord [1].

Based on the analysis carried out, it shows that m-health is a technology by applying elements of health services that can be easily accessed by patients with chronic diseases as an effort to self-protect. It is also shown that patients can develop themselves without external influences; in other words, patients will be responsible for their own health [18]. In contrast, patients with chronic diseases will attempt to maintain healthcare that is carried out autonomously by addressing the demands of

**Table 2** Top five cited publications (2020)

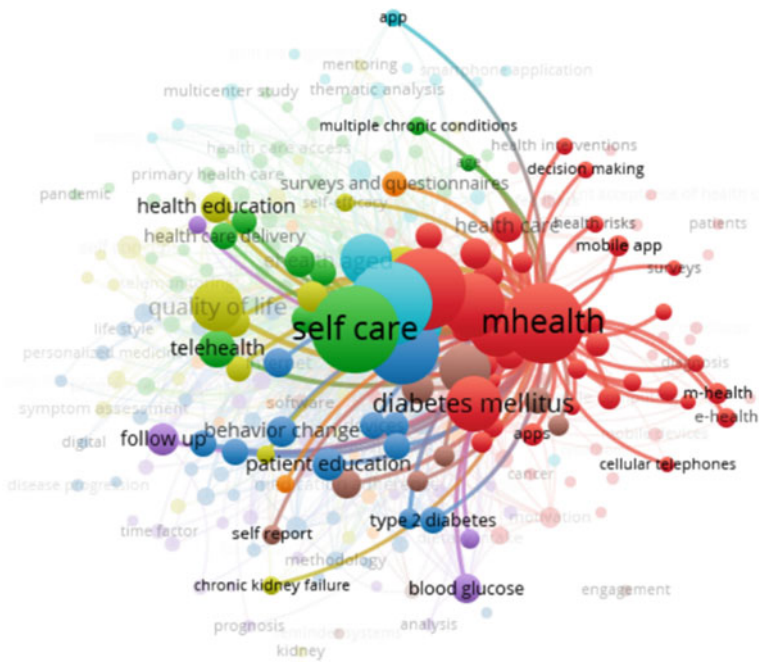
Authors	Source title	Year	Cited
Liu et al. (2020)	Jmir Mhealth and Uhealth	2020	33
Doupis Georgios et al. (2020)	Diabetes Ther	2020	31
Sittig Scot et al. (2020)	Jmir Mhealth and Uhealth	2020	28
Barbosa et al. (2020)	Journal of chronic obstructive pulmonary disease	2020	25
Duan et al. (2020)	Jmir Mhealth and Uhealth	2020	21

**Table 3** Top five cited publications (2021)

References	Source title	Year	Cited
Alazzam et al. (2021)	Computational Intelligence and Neuroscience	2021	23
Moses et al. (2021)	Healthcare	2021	14
Janjua et al. (2021)	Digital Interventions for the Management	2021	9
Wang et al. (2021)	Clinical Rehabilitation	2021	6
(Racioppi et al. 2021)	Transplantation and Cellular Therapy	2021	5

**Table 4** The ten most frequent authors’ keywords

Ranking	Author’s keywords	Frequency	Ranking	Author’s keywords	Frequency
1°	Self-care	2607	11°	Mobile phone	775
2°	Mobile application	2267	12°	Hypertension	676
3°	Self-management	1884	13°	Telehealth	519
4°	Telemedicine	1794	14°	Patient education	373
5°	m-health	1428	15°	Internet	352
6°	Chronic disease	1144	16°	Follow up	351
7°	Diabetes mellitus	1011	17°	Patient care	332
8°	Quality of life	923	18°	Text messaging	305
9°	Mobile health	875	19°	Health education	287
10°	Smartphone	792	20°	Medical technology	274



**Fig. 4** Relation of m-health and self-management with chronic illness

knowledge, medication adherence, and self-awareness to maintain health in order to generate a high quality of life for patients with chronic diseases [19].

This finding is in accordance with the research conducted by Jeffrey et al. [15], that focuses on the use of m-health applications in the self-management of patients with

type 2 DM. Also, the results of the study revealed that the use of (m-health) successfully assists patients with chronic illness and that m-health applications can increase self-management in patients with chronic diseases, as well as their ability, knowledge, and self-awareness of the significance of health. The application of m-health is good for use in patients with chronic diseases because it can provide knowledge that m-health is a healthcare tool, information about the severity of the disease, and technological literacy. The lack of application in m-health in the studies was found by Moses et al. (2021). Further, the previous review found that lack of application was related to absence of features, such as complications of diabetes, including decreased function in organ and hypoglycemic episodes [20]. In addition, tracking features including monitoring blood glucose levels with dynamic trends and tips to manage comorbidities and lifestyle changing including nutritional features were also still lacking [15].

Similar research that focused on the use of m-health for older adult patients, explained that adult patients with health disorders can be controlled by the use of m-health [14], but there is more attention to older patients in running of m-health with the need for supervision, which is more to evaluate the technology that is already available, due to physical ability barriers. The development of the US module framework with the Think Aloud module can facilitate the identification of obstacles and difficulties in the healthcare program for older patients, allowing for their optimal resolution [14].

Marquez Contreras et al. conducted research with discussions related specifically on the use of smartphone application to improve medication adherence in hypertension [21]. Their study states that application (m-health) is an effective way to improve treatment adherence and self-management with chronic patients, there are changes in compliance in patients who experience hypertension disease that are improving with the application (m-health) used to manage health woes. The average daily compliance percentage is between 80 and 100%, and control with daily compliance is 93.15% and 86.3%, and 70.66% and 62.6% with a time span of six and twelve months, respectively (p. 05). The availability of m-health improves compliance of patients with hypertension, cancer, and type 2 DM.

## 4 Conclusion

According to the analysis conducted, a mobile health smartphone application is a type of telehealth or tele-rehabilitation, which is an essential instrument in the field of health promotion. The advantages of mobile health or m-health can help individuals with chronic conditions enhance their self-management. Self-management is an important feature of coping with chronic diseases, and the existence of self-management has a favorable impact on patients' quality of life, exercise capacity, and hospitalizations due to patient complaints.

Over time, the care of chronic diseases can result in a plethora of major complications and substantial morbidity and mortality worldwide. The primary objective

of m-health is to raise awareness through self-management by highlighting difficult-to-obtain information or expertise, such as training information and the difficulty of maintaining lifestyle alterations. In addition, m-health can improve access to specialized health care. The evolution of mobile technology has spawned a large number of health-related m-health applications that aim to improve patients' self-management skills in chronic illness, to facilitate communication between patients and healthcare professionals and to also improve patient adherence to care, which focuses primarily on self-management, lifestyle modification, and treatment adherence motivation.

## References

1. Agnihothri S, Cui L, Delasay M, Rajan B (2020) The value of mHealth for managing chronic conditions, (in eng). *Health Care Manag Sci* 23(2):185–202. <https://doi.org/10.1007/s10729-018-9458-2>
2. Catarino M, Charepe Z, Festas C (2021) Promotion of self-management of Chronic disease in children and teenagers: scoping review, (in eng). *Healthcare (Basel)* 9(12), 27 Nov 2021. <https://doi.org/10.3390/healthcare9121642>
3. Stampe K, Kishik S, Müller SD (2021) Mobile health in Chronic disease management and patient empowerment: exploratory qualitative investigation into patient-physician consultations, (in eng). *J Med Internet Res* 23(6):e26991, 15 Jun 2021. <https://doi.org/10.2196/26991>
4. Hamine S, Gerth-Guyette E, Faulx D, Green BB, Ginsburg AS (2015) Impact of mHealth chronic disease management on treatment adherence and patient outcomes: a systematic review, (in eng). *J Med Internet Res* 17(2):e52, 24 Feb 2015. <https://doi.org/10.2196/jmir.3951>
5. Cheng ASK, Liu X, Ng PHF, Kwok CTT, Zeng Y, Feuerstein M (2020) Breast cancer application protocol: a randomised controlled trial to evaluate a self-management app for breast cancer survivors, (in eng). *BMJ Open* 10(7):e034655, 5 Jul 2020. <https://doi.org/10.1136/bmjopen-2019-034655>
6. Allegrante JP, Wells MT, Peterson JC (2019) Interventions to support behavioral self-management of Chronic diseases, (in eng). *Annu Rev Public Health* 40:127–146. <https://doi.org/10.1146/annurev-publhealth-040218-044008>
7. Chan SW (2021) Chronic disease management, self-efficacy and quality of life, (in eng). *J Nurs Res* 29(1):e129, 1 Feb 2021. <https://doi.org/10.1097/jnr.0000000000000422>
8. Ha Dinh TT, Bonner A, Clark R, Ramsbotham J, Hines S (Jan 2016) The effectiveness of the teach-back method on adherence and self-management in health education for people with chronic disease: a systematic review, (in eng). *JBIM Database Syst Rev Impl Rep* 14(1):210–47. <https://doi.org/10.11124/jbisrir-2016-2296>
9. Kang YN et al. (2019) Does a mobile app improve patients' knowledge of stroke risk factors and health-related quality of life in patients with stroke? a randomized controlled trial, (in eng). *BMC Med Inf Decis Mak* 19(1):282, 21 Dec 2019. <https://doi.org/10.1186/s12911-019-1000-z>
10. Stamenova V et al. (2020) Technology-enabled self-management of chronic obstructive pulmonary disease with or without asynchronous remote monitoring: randomized controlled trial (in eng). *J Med Internet Res* 22(7):e18598, 30 Jul 2020. <https://doi.org/10.2196/18598>
11. Sockolow PS, Buck HG, Shadmi E (2021) An integrative review of chronic illness mHealth self-care interventions: mapping technology features to patient outcomes, (in eng). *Health Inf J* 27(3):14604582211043914, Jul–Sep 2021. <https://doi.org/10.1177/14604582211043914>
12. del Río-Lanza A-B, Suárez-Vázquez A, Suárez-Álvarez L, Iglesias-Argüelles V (2020) Mobile health (mhealth): facilitators and barriers of the intention of use in patients with chronic illnesses. *J Commun Healthcare* 13(2):138–146, 04 Jan 2020. <https://doi.org/10.1080/17538068.2020.1777513>

13. Hallberg D, Salimi N (2020) Qualitative and quantitative analysis of definitions of e-Health and m-Health, (in eng). *Healthc Inform Res* 26(2):119–128. <https://doi.org/10.4258/hir.2020.26.2.119>
14. Wildenbos GA, Jaspers MWM, Schijven MP, Dusseljee-Peute LW (2019) Mobile health for older adult patients: using an aging barriers framework to classify usability problems, (in eng). *Int J Med Inform* 124:68–77. <https://doi.org/10.1016/j.ijmedinf.2019.01.006>
15. Jeffrey B et al (2019) Mobile phone applications and their use in the self-management of type 2 diabetes mellitus: a qualitative study among app users and non-app users, (in eng). *Diabetol Metab Syndr* 11:84. <https://doi.org/10.1186/s13098-019-0480-4>
16. Doupis J, Festas G, Tsilivigos C, Efthymiou V, Kokkinos A (2020) Smartphone-based technology in diabetes management, (in eng). *Diabetes Ther* 11(3):607–619. <https://doi.org/10.1007/s13300-020-00768-3>
17. Sittig S, Wang J, Iyengar S, Myneni S, Franklin A (2020) Incorporating behavioral trigger messages into a mobile health app for chronic disease management: randomized clinical feasibility trial in diabetes, (in eng). *JMIR Mhealth Uhealth* 8(3):e15927, 16 Mar 2020. <https://doi.org/10.2196/15927>
18. Alzahrani A, Gay V, Alturki R (2022) Exploring Saudi individuals' perspectives and needs to design a hypertension management mobile technology solution: qualitative study, (in eng). *Int J Environ Res Pub Health* 19(19) 10 Oct 2022. <https://doi.org/10.3390/ijerph191912956>
19. Davis SP, Ross MSH, Adatorwovor R, Wei H (2021) Telehealth and mobile health interventions in adults with inflammatory bowel disease: a mixed-methods systematic review (in eng). *Res Nurs Health* 44(1):155–172. <https://doi.org/10.1002/nur.22091>
20. Moses JC, Adibi S, Shariful Islam SM, Wickramasinghe N, Nguyen L (2021) Application of smartphone technologies in disease monitoring: a systematic review, (in eng). *Healthcare (Basel)* 9(7):14. <https://doi.org/10.3390/healthcare9070889>
21. Márquez Contreras E et al. (Jan 2019) Specific hypertension smartphone application to improve medication adherence in hypertension: a cluster-randomized trial, (in eng). *Curr Med Res Opin* 35(1):167–173. <https://doi.org/10.1080/03007995.2018.1549026>

# The Readiness of a Private Hospital Toward Smart Hospital in Indonesia



Nur Hidayah, Qurratul Aini, and Gofur Ahmad

**Abstract** The research aims to analyze the readiness of a private hospital in Indonesia to become a smart hospital by digitalizing Internet-based information system management. The research adopted a qualitative approach, namely a case study, by conducting in-depth interviews. The informants were eight human resources of a private hospital consisting of the hospital director, finance manager, human resource manager, deputy of human resource manager, nursing manager, education and training manager, and assistant manager of management information system. The findings showed that the hospitals had made efforts to realize smart hospitals by implementing a hospital management information system (HMIS) consisting of electronic medical records, a human resource information system, a billing system for patient registration and payment services, and a pharmacy information system. On the other hand, the accounting and financial information systems have not been fully integrated because they are still being developed. Even though the HMIS development plan is part of one of the strategic programs in the hospital's strategic plan for 2022–2026, it would be better if the hospital had an HMIS-specific strategic plan to result in a more systematic approach to better HMIS planning and development.

**Keywords** Internet of things (IoT) · Management information system · Smart hospital · E-health

---

N. Hidayah (✉) · Q. Aini

Master of Hospital Administration, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia 55183

e-mail: [nurhidayah@umy.ac.id](mailto:nurhidayah@umy.ac.id)

G. Ahmad

Master of Management, Universitas Muhammadiyah Jakarta, Jakarta, Indonesia 15419

## 1 Introduction

Smart cities have been developed and have received widespread global attention [1]. The smart city concept was developed using information and communication technology that invests in human and social capital to improve citizens' quality of life by promoting economic growth, participatory governance, effective and efficient human resource management, sustainability, and efficient mobility, while also ensuring the citizens' privacy and security [2]. Smart healthcare is a critical component of smart city development implemented in several hospitals in developed countries [1]. Information and communication technologies (ICT) in the healthcare industry resulted in the development of the concept of electronic health (e-health), which contributes to cost savings and efficiency gains [2]. In addition, the digital transformation of the healthcare system is considered a critical process that impacts the current and future healthcare systems [3].

The widespread adoption of the Internet of Things (IoT) has resulted in a significant evolution in every single field of human life from a technological standpoint, including the healthcare sector [4, 5]. Therefore, IoT-based smart healthcare has begun to be developed globally, including remote patient monitoring, emergency patient care, treatment and routine medical examination reminders, remote patients' medical prescription, and looking for the patient's nearest health resources [5]. IoT-based smart healthcare enables more effective communication between patients, equipment, doctors and other medical personnel, and hospital systems. Also, it enhances the role of medical care and resources. Thus, by adopting IoT-based smart healthcare, medical care can be more efficient, and patients can receive better medical treatment [6].

Smart hospitals in industry 4.0 or Hospital 4.0. are increasingly being promoted as the key to increasing productivity and flexibility, promoting economic growth, and ensuring industrial systems' resilience [1]. A "smart hospital" is a term used to describe digitalized hospitals that use automated and optimized IoT-based processes. Smart hospitals prioritize improving patient care by introducing new modern healthcare services. These provide a comprehensive description of the patient's disease, allow doctors to respond quickly, and highlight patient monitoring continuously [7]. Integrating technology and healthcare has big opportunities in the current era of personalized healthcare, particularly during the Coronavirus disease (COVID-19) pandemic. Nowadays, systems that integrate connected devices, tracking technology, and continuous monitoring are increasing in demand [8].

A private hospital in Central Java, Indonesia, has made efforts to integrate Internet-based information technology (IT) into the hospital units. The hospital comprises several units: outpatient, inpatient, and intensive care, supporting service, financial, human resources, facilities and infrastructure. However, not all hospital units have been integrated into information technology. The hospital service unit has provided online patient registration since 2020, but still uses a smartphone application like WhatsApp. Also, financial and human resources management information systems are still in the planning stage, while patient service management and electronic

medical records (EMR) are under development [9]. EMR enables hospitals to provide better services to patients with a higher level of efficiency, easier access to information, and more effective patient care [10]. The hospital implemented an information system for outpatients in 2021, including an online registration system and an EMR. However, the EMR was limited to pediatric and neurology polyclinics.

Currently, not all hospital units have been integrated with internet-based information systems, except for online patient registration, EMR for pediatric and neurology clinics, and billing systems. Additionally, as the person responsible for developing a hospital management information system, the hospital management information system does not yet have a strategic plan for developing a hospital management information system that is integrated with the hospital's strategic plan. Several studies have proven that an integrated management information system has become a strategy to achieve efficiency, effectiveness, and productivity in achieving the hospital's vision and mission. The previous study suggested that the management information system unit should develop a separate strategic plan and implement evaluation and control integrated with the hospital strategic plan [11]. Hence, this research aims to assess the readiness of a private hospital in Central Java, Indonesia, to become a smart hospital by digitalizing an Internet-based hospital management information system (HMIS).

## 2 Research Methods

This research used a qualitative approach called a "case study" to get detailed information from the participants regarding the readiness of a private hospital to become a smart hospital by conducting a digitalization Internet-based hospital management information system. A case study is a qualitative research approach widely used in many fields, especially social sciences. A case study aims to gain an in-depth understanding of a phenomenon or complex issue in a real-life context [12]. The participants were eight human resources of a private hospital consisting of a hospital director, finance manager, human resource manager, deputy of human resource manager, nursing manager, education and training manager, and assistant manager of the management information system. The data was collected by conducting an in-depth interview. The audio recordings of the interview were transcribed into written statements. Furthermore, the interview transcripts were analyzed using coding analysis. Each research finding is categorized and labeled using coding analysis to generate a set of common themes [13].



### 3 Results

The results of this research indicated that a private hospital in Central Java, Indonesia, has developed a hospital management information system (HMIS) to become a smart hospital. Several HMISs, namely electronic medical records (EMRs), financial information system, accounting information system (AIS), human resources information system (HRIS), billing system, and pharmacy information system (PIS). The following paragraphs explain the research results in further depth.

**Management Information System Development:** The development of the hospital management information system (HMIS) received a positive response from the hospital leaders. Both hospital directors and managerial directors have begun to realize the importance of developing HMIS. The hospital supports the development of HMIS, particularly from a financial standpoint. One informant said: *"I see hospital leaders becoming more aware of the development of management information systems. Our current hospital director is also people who are aware of the development of information systems. I also see that the current managerial directors are considerably more transparent regarding the development of information systems."* (Informant 1). One informant added: *"So far, the budget has not been a problem ... it has been budgeted. We have been quite supportive"* (Informant 5).

HMIS has not been carefully developed. It is still adjusting to the information system required by the hospital. The hospital does not have a strategic plan yet for a hospital management information system that is integrated with the hospital's strategic plan. One informant stated: *"So far, the planning has not yet been completely done"*. (Informant 5). One informant said: *"There is no plan yet. It is still based on hospital needs. If the hospital requires an urgent information system, it will be developed by the information technology (IT) unit. The grand design does not exist yet"* (Informant 2). One informant explained: *"The MIS unit doesn't have its own strategic plan yet...there must be a manager who is an expert in the development of information systems and can guide them in creating the strategic plan."* (Informant 1).

Even though the hospital does not have a specific plan yet, it has made efforts to develop HMIS, such as holding discussions, collaborating with external vendors, and proposing the addition of an information technology (IT) team. Discussions on the development of HMIS are still being carried out, such as discussions regarding development planning and suggestions from each unit. One informant revealed: *"The discussions are currently ongoing. The general description of the system that will be developed has already begun to be compiled, and the suggestions from units are also being worked on."* (Informant 1). One informant also argued: *"IT and radiology staff discussed how to facilitate radiology specialists in order can view radiological images."* (Informant 2).

Additionally, the hospital has attempted to propose an additional IT team. Hospitals need more qualified IT staff to enhance the development of hospital information systems. One informant mentioned: *"We have actually tried to propose an additional team."* (Informant 1). Another informant also revealed: *"Now what we feel is lacking*

*the resource to create the notion. It can be developed by more experienced IT staff.*" (Informant 3). One informant added: *"There are two programmers and analysts, and one technician ... this is still insufficient. There are plans to hire more."* (Informant No. 5).

The Implementation of Hospital Management Information System: Implementing a hospital management information system (HMIS) facilitates the hospital in managing the entire hospital service process. Nevertheless, not all hospital service processes are integrated with HMIS, because some are currently in the development process. Below is the HMIS that has been implemented by the hospital.

**Electronic Medical Records (EMRs):** Electronic medical records (EMRs) have become one of the information systems preferred by the hospital. EMR has only been implemented in outpatient unit, and it has not been implemented in other units. Also, the hospital involves the HMIS team in developing EMRs. One informant mentioned: *"We are preferring electronic medical records (EMRs)."* (Informant 4). One informant stated: *"Electronic medical records are still outpatient care. The development is almost complete."* (Informant 5). Also, informant 4 added: *"There is a hospital management information system team consisting of many clinicians, particularly permanent specialists, nurses, and staff from a pharmacy, laboratory, and registration. We make every effort to contribute suggestions or fundamental ideas for the development of electronic medical records. There is a development plan that will be implemented during the following year, although it has not yet been specified in full on how to perform it."* (Informant 4).

All hospital employees, including doctors, nurses, and other users, were given training regarding the use of ERM. The IT staff provided personal training to each doctor. Meanwhile, nurses and administrative staff received the training through a socialization forum delivered by the IT team. One informant said: *"Employees, physicians, and users are trained individually."* (Informant 4). One informant stated: *"The training does not use a class system; instead, the IT team communicates with doctors."* (Informant 5). One informant also argued: *"We invite nursing and administrative employees to attend the socialization meeting held by the IT team."* (Informant 4).

**Financial Information System:** The financial information system cannot be used because it is still developing. One informant mentioned: *"The financial system has been implemented. However, its utilization has not been optimal."* (Informant 1). Again, informant 1 stated: *"Yesterday we had a discussion in the Board of Directors Manager group that discussed information systems. Last January, the Deputy Director of Finance told the Finance Manager to look for partners related to the development of the financial information system, but I do not know the progress yet."* (Informant 1).

**Accounting Information System (AIS):** The accounting information (AIS) system has not yet been fully created because development time with vendors is very limited. The accounting information system cannot be entirely automated. Manual input is still required. Even financial reports are still manually entered using Microsoft Excel. One informant revealed: *"The development is not fully developed because the accounting was developed last, so it only gets a little time from the vendors."*

*Still partially manual. While this is still a double input, the cashier's report cannot be withdrawn automatically. The financial statements from the trial balance are extracted and then entered into Excel. We have been unable yet to produce financial reports or other required reports."* (Informant 5).

**Human Resources Information System (HRIS):** The human resources division already has an information system of human resources. Currently, the information system of human resources is in the trials phase. Each hospital employee can input employment data and supporting documentation into the human resources information system. One informant expressed: *"As far as I know, the human resource information system division already has an human resources. ... Still in the process of building a database, for example, related to training data, proof of marriage certificate, and proof of membership of Social Health Insurance Administration Agency for workers, and others."* (Informant1). One informant also argued: *"Still in the trial phase. Once upon a time, I was given socialization to input the document, like entering the certificates I have this year into the system."* (Informant 4).

However, the hospital's human resource performance evaluation has not been integrated into the system. The performance evaluation is still conducted using the manual form. One informant mentioned: *"As far as I know, the assessment is still using the manual form."* (Informant 1). One informant also said: *"The staff performance appraisal system is also still manual. It still uses paper forms that are distributed to all units to be filled out and then evaluated by the human resource division."* (Informant 2).

**Billing System:** Payment and patient registration are two services in hospitals that have been integrated with the billing system. One informant stated: *"The billing system is one of the information systems that has advanced rapidly since 2016. This system simplifies billing services for patients."* (Informant 2). Again, informant 2 said: *"The laboratory unit is not yet integrated with billing."* (Informant 2).

Payment services have been facilitated using electronic payments. However, there are still certain specialists who use non-electronic payment methods. One informant explained: *"Some no longer utilize paper, opting instead for technological devices. The cashier will provide the patient with instructions for pressing the payment button. Currently, payment is centralized through banks that cooperate with hospitals... still working with conventional banks."* (Informant 4). One informant added: *"But some specialists still use the old system.....which is not yet electronic.....patients are given a receipt to pay the cashier."* (Informant 4).

There are several patient registration systems implemented by the hospital, including online registration using the national health insurance (NHI) mobile application, a self-registration platform, and registration using the *Muyassar* application. One informant mentioned: *"The registration is already integrated with billing system."* (Informant 2). Another informant said: *"Social Health Insurance Administration Agency (SIAA) wants a system... we have to adapt to SIAA. The patient has received a registration number, but verification is still required."* (Informant 4). One informant added: *"The NHI online system for SIAA patients. Patients can arrive at the appointed time based on the doctor's schedule."* (Informant 1).

**Pharmacy Information System (PIS):** Prescribing medicine in hospitals already uses electronics. The hospital has also facilitated online medicine delivery. Payment for medicine delivery can be made with an e-wallet or with cash. One informant mentioned: *“Electronic prescriptions are also available.”* (Informant 4). One informant also said: *“We have just launched a medication delivery service in collaboration with Grab... it can be with two payment systems, an e-wallet or the patient can pay for it on the Grab application.”* (Informant 1).

Pharmacy managers have evaluated the response time of medicine preparation. The response time data for medicine preparation showed that there were still patients who received one polyclinic service for more than three hours. Furthermore, the evaluation result will be used as evaluation material to enhance the quality of hospital services. One informant revealed: *“The Pharmacy manager is trying to evaluate the response time for the medicine preparation... more than three hours for one polyclinic service. The response time data is available so that it can be used as a basis for future quality improvements.”* (Informant 1).

## 4 Discussion

The research results showed that the hospital had made efforts to develop a hospital management information system (HMIS). The HMIS is specifically developed to manage the hospital's routine activities system-based, such as administration, finance, clinical services, and supplementary services. Due to the rapid advancement of digital technology, most healthcare companies adopt digital-based healthcare services to provide the best possible healthcare [14]. The application of digital-based healthcare services has the potential to enhance the standard of treatment provided, reduce costs, and increase patient safety and equity [15]. Thus, HMIS developed by the hospital will be discussed in a further paragraph.

**Electronic Medical Records (EMRs):** The results indicated that outpatient units, namely pediatric and neurology clinics, have used electronic medical records (EMRs). Meanwhile, some polyclinics still use paper to write down the patient's medical information. EMRs are currently limited to pediatric and neurology clinics, as in the prior year [11]. Due to the computer system's development, digital-based clinical exams and medical records have become widely adopted in today's healthcare systems [16]. EMRs are integrated systems that can exchange and distribute patient clinical data information in the healthcare system [17]. Adopting EMRs enable doctors to share patients' information promptly and access accurate and complete data. When doctors have access to accurate and complete data, they can make diagnostic and treatment decisions based on the most recent, thorough understanding of each patient. This situation leads to better healthcare. Also, the identification and intervention of potentially fatal errors are provided. Thus, medicine allergies or contraindications can be prevented before harm occurs [15].

**Financial Information System:** The hospital does not yet have an information technology-integrated financial information system. Even the financial reports are

still created manually. Hence, the financial information system is currently being developed. Most manual records and manual management are used in traditional financial management. On the contrary, due to the rapid development of information and computer technology (ICT), the recording of deposits, income, expenses, household cash, and transfer can easily be managed using a computer-based system [18].

**Accounting Information System (AIS):** The findings indicated that the accounting information system (AIS) has not yet been fully integrated with the system, so there is still work to be done through double input. Due to the rapid development of information technology, many organizations adopted computerized accounting systems to improve organizational performance. These developments have saved time, costs, and resources to enhance faster and better business transactions and communication [19]. AIS is a computerized electronic system that covers a business's transaction processing processes, such as collecting, storing, processing, tracking, and conveying financial and accounting data through financial reports. Also, AIS provides accurate accounting data that the user can utilize to conduct effective planning, control function, and decision to maintain the organization's competitiveness. The computerized accounting system processes and converts data into financial reports that are accessible to both internal and external users. In financial reports, the cash flow, income, and balance sheet are presented [19].

**Human Resource Information System (HRIS):** The human resource information system (HRIS) has been planned since 2020 and has now been successfully realized. The human resources unit already has an HRIS that can be used to input and access health human resources (HRH) data. However, the HRIS has not been integrated into the employee performance assessment. Human resources information systems (HRIS) is a set of computer programs, databases, software, and hardware used to record, store, manage the data, and retrieve and change the data for human resource function needs. By implementing HRIS, the staff workload is minimized. The employee does not document and manually changes the record instead of utilizing a system, hence increasing the employee's productivity [20].

**Billing System:** The results indicated that the hospital had facilitated a registration and payment services billing system. The first is registration services. Previous research has shown that patient registration can already be completed online using the WhatsApp application [9]. Currently, patient registration has been improved. The patient registration can be conducted online using the national health insurance (NHI) mobile application. The Indonesian Social Health Insurance Administration Agency (SIAA) has released NHI mobile application that can be used by society to gain access to quality healthcare whenever and wherever it is needed using an Internet connection. The hospital also provided online application registration, namely *Muyassar*. *Muyassar* application allows patients to look at the doctor's schedule, room, and queue and can register new and previous patients.

Additionally, patients have also been facilitated self-registration platform. Implementing the registration system integrated with the billing system can improve the effectiveness and efficiency of the delivery of health services. The patient's waiting time becomes reduced, particularly during patient registration. Also, work becomes

considerably easier because it does not use paper but a system [21]. Besides, the research's results also show that the system has served some payment services while others still use paper. The billing system has also been integrated with bank transfer payments.

**Pharmacy Information System (PIS):** The pharmacy information system (PIS) has facilitated system-integrated medicine prescribing. The result showed that private hospitals also facilitate telehealth services, namely medicine delivery that collaborates with online delivery services. The purchase of pharmaceutical medicine is already served via an online pharmacy or pharmacy e-commerce [22]. There are two hospital pharmacies: inpatient and outpatient [23]. PIS is now widely adopted in the modern healthcare system to accomplish a variety of pharmacy functions, including entering, managing, and distributing patient orders; managing inventory and purchases; setting prices, charges, and bills; and reporting medications. The hospital can adopt PIS to minimize medication errors, enhance the quality of medication information management, and increase pharmaceutical productivity through drug interaction checking and dose calculation [24].

## 5 Conclusion

A private hospital in Central Java, Indonesia, has been trying to realize a smart hospital by implementing a hospital management information system (HMIS) integrated with hospital health services. Implementing HMIS is becoming one of the most important aspects required to realize a modern hospital. HMIS, namely electronic medical records, a human resource information system, a billing system for patient registration and payment services, and a pharmacy information system, are already integrated with hospital health services. However, the accounting and financial information systems have not been fully integrated, so some tasks are still performed manually and partly utilizing the system. Also, the hospital has established a strategic plan for 2022–2026 that covered a strategic plan, namely the integrated HMIS for developing information technology-based health services. The key performance indicator (KPI) and achievement targets for the next five years have been established. Hopefully, the hospital can establish the HMIS-specific strategic plan so that the development of HMIS can be more focused and better planned. Consequently, hospitals will be closer to achieving a smart hospital.

In today's digital age, hospitals must not only fulfill the social mission but also be able to compete in the healthcare industry and the free market. Hence, the hospital manager is responsible for properly managing healthcare services so that hospitals can compete with other healthcare industries. The role and usage of computer technology in modern hospital management cannot be separated, particularly in the field and scope of workaday. The rapid development of computer technology makes it easy to use and cost-saving [25].

## 5.1 Limitations of the Research

The research's results provided the comprehension regarding implementing a hospital management information system (HMIS) to realize a smart hospital. The research data was only obtained by conducting in-depth interviews. Hence, conducting more than one data collection technique is suggested for further research to gather more data. Also, the other researchers can conduct research regarding HMIS, focusing on the implementation and other aspects such as the planning, evaluation, user satisfaction, etc.

## References

1. Bongomin O, Yemane A, Kembabazi B, Malanda C (2020) The hype and disruptive technologies of industry 4.0 in major industrial sectors: a state of the art. Preprints 1:1–68. <https://doi.org/10.20944/preprints202006.0007.v1>
2. Solanas A et al (2014) Smart health: a context-aware health paradigm within smart cities. *IEEE Commun Mag* 52(8):74–81. <https://doi.org/10.1109/MCOM.2014.6871673>
3. Ricciardi W et al (2019) How to govern the digital transformation of health services. *Eur J Public Health* 29:7–12. <https://doi.org/10.1093/eurpub/ckz165>
4. Shahzad SK, Ahmed D, Naqvi MR, Mushtaq MT, Iqbal MW, Munir F (2021) Ontology driven smart health service integration. *Comput Methods Programs Biomed* 207:1–18. <https://doi.org/10.1016/j.cmpb.2021.106146>
5. Pasha M, Shah SMW (2018) Framework for E-health systems in IoT-based environments. *Wirel Commun Mob Comput*, vol 2018. <https://doi.org/10.1155/2018/6183732>
6. Al-Mahmud O, Khan K, Roy R, Mashuque Alamgir F (2020) Internet of Things (IoT) based smart health care medical box for elderly people. In: 2020 International conference for emerging technology INCET 2020, pp 1–6. <https://doi.org/10.1109/INCET49848.2020.9153994>
7. Rodrigues L, Gonçalves I, Fé I, Endo PT, Silva FA (2021) Performance and availability evaluation of an smart hospital architecture. *Computing* 103(10):2401–2435. <https://doi.org/10.1007/s00607-021-00979-x>
8. Ganji K, Parimi S (2022) ANN model for users' perception on IOT based smart healthcare monitoring devices and its impact with the effect of COVID 19. *J Sci Technol Policy Manag* 13(1):6–21. <https://doi.org/10.1108/JSTPM-09-2020-0128>
9. Hidayah N, Aini Q, Amin M (2021) Innovation in strategic planning at private hospital in central Java district, Indonesia. 518:481–485. <https://doi.org/10.2991/assehr.k.210120.164>
10. Amin M, Setyoningroho W, Hidayah N (2021) Implementasi Rekam Medik Elektronik: Sebuah Studi Kualitatif. *JATISI (Jurnal Tek Inform dan Sist Informasi)* 8(1):430–442. <https://doi.org/10.35957/jatisi.v8i1.557>
11. Hidayah N, Aini Q, Amin M (2021) The implementation of the information technology system to improve health service. *J Contemp Issues Bus Gov* 27(5):2572–2576
12. Crowe S, Cresswell K, Robertson A, Huby G, Avery A, Sheikh A (2011) The case study approach. *BMC Med Res Methodol* 11(1):1–9. <https://doi.org/10.1186/1471-2288-11-100>
13. Creswell JW (2012) Educational research: planning, conducting and evaluating quantitative and qualitative research. 4th ed. Pearson Education, Inc.
14. Handayani PW, Hidayanto AN, Pinem AA, Sandhyaduhita PI, Budi I (2018) Hospital information system user acceptance factors: user group perspectives. *Inform Heal Soc Care* 43(1):84–107. <https://doi.org/10.1080/17538157.2016.1269109>
15. PAHO (2016) eHealth in Latin America and the Caribbean: interoperability standards review

16. Dash S, Shakyawar SK, Sharma M, Kaushik S (2019) Big data in healthcare: management, analysis and future prospects. *J Big Data* 6(1). <https://doi.org/10.1186/s40537-019-0217-0>
17. Reis ZSN, Maia TA, Marcolino MS, Becerra-Posada F, Novillo-Ortiz D, Ribeiro ALP (2017) Is there evidence of cost benefits of electronic medical records, standards, or interoperability in hospital information systems? overview of systematic reviews. *JMIR Med Inform* 5(3). <https://doi.org/10.2196/medinform.7400>
18. Ma R (2021) Design and implementation of financial information system for mobile devices. *J Phys Conf Ser* 1915(4):0–6. <https://doi.org/10.1088/1742-6596/1915/4/042010>
19. Teru SP, Idoku I, Ndeyati JT (2017) A Review of the impact of accounting information system for effective internal control on firm performance. *Indian J Financ Bank* 1(2):52–59. <https://doi.org/10.46281/ijfb.v1i2.89>
20. Irum A, Yadav RS (2019) Human resource information systems: a strategic contribution to HRM. *Strateg Dir* 35(10):4–6. <https://doi.org/10.1108/SD-02-2019-0043>
21. Mohamadali NA, Aziz NFA (2017) The Technology factors as barriers for sustainable health information systems (HIS)—a review. *Procedia Comput Sci* 124:370–378. <https://doi.org/10.1016/j.procs.2017.12.167>
22. Srivastava M, Raina M (2020) Consumers' usage and adoption of e-pharmacy in India. *Int J Pharm Healthc Mark* 15(2):235–250. <https://doi.org/10.1108/IJPHM-01-2020-0006>
23. Nazer LH, Tuffaha H (2017) Health care and pharmacy practice in Jordan. *Can J Hosp Pharm* 70(2):150–155. <https://doi.org/10.4212/cjhp.v70i2.1649>
24. Hamza PA et al (2021) Recruitment and selection: the relationship between recruitment and selection with organizational performance. *Int J Eng Bus Manag* 5(3):1–13. <https://doi.org/10.22161/ijebm.5.3.1>
25. Gomer S, Kusumapradja R (2020) Acceptance model of hospital information management system: case of study in Indonesia. *Eur J Bus Manag Res* 5(5):1–8. <https://doi.org/10.24018/ejbmr.2020.5.5.505>



# Coping with the Business Ethics Issues in the Era of the Internet of Things



Indah Fatmawati

**Abstract** The advancement of technology has brought society to the Internet of Things industry. Businesses become more efficient due the technology development via innovation and creativity. New business model creation due to the Internet has changed employment and transaction practices. The emergence of new business models and efficiency has created ethical issues and challenges such as sharing economy employment sustainability, inequality, robot, artificial intelligence in the workplace, and unemployment due to technological advancement. Business ethics is not an exclusive norm developed by business players. Society has its social standards, which will be applied to every aspect of human life, including business. What is right and wrong in the community will be perceived similarly to business practices. Hence, people will not evaluate the business practice based on company perception but on their norms. Consequently, with the radical changes in business models and practices, a new consensus of business ethics between companies and society may be developed to accommodate and balance between companies and society norms.

**Keywords** Technology advancement · Internet of things · New business model · Business ethics

## 1 Introduction

Business growth supported by technological advances has brought the world to an industrial revolution, and we are arriving at Industrial Revolution (IR) 4.0. IR 4.0 is a current phenomenon that will impact businesses by continuing the development of previous technology. The industrial Internet of Things for Industry 4.0 (IoT) Integrating Internet of Things (IoT) technologies that generate industrial

---

I. Fatmawati (✉)

Universitas Muhammadiyah Yogyakarta, Yogyakarta 55183, Indonesia

e-mail: [indahfatmawati@umy.ac.id](mailto:indahfatmawati@umy.ac.id)

value allows manufacturers to use fully digital, connected, intelligent, and decentralized manufacturing processes [1]. It arose due to advancements in information and data technology. Increasing productivity is at the heart of every industrial revolution, and IR 4.0 is expected to significantly alter product design, processes, operations, and services [1]. Industry 4.0 is predicted to impact management and future employment prospects in several ways, including enabling the development of novel business models with far-reaching consequences for industries and markets, affecting all stages of the product lifecycle, introducing novel production and transactional methods, increasing efficiency and productivity, and making businesses more competitive [1].

The Internet of Things era has triggered the emergence of business competition based on technological innovation, one of which is realized in artificial intelligence (AI). Artificial intelligence (AI) refers to “machines that can interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation” [2]. It has grown popular across various academic disciplines, industry sectors, and business functions. Artificial intelligence has far-reaching effects on all aspects of civilization [2]. There have been changes in the pattern of life in various economic fields, inseparable from technological developments that can benefit aspects of human life. Therefore, technological innovation and creativity provide business opportunities that contribute to economic development. Online commerce has become a significant business channel that becomes the spirit for creativity which significantly impacts the growth of the country’s digital economy.

In Indonesia, the exceptionally high Internet usage rate has the potential to capitalize on commercial opportunities presented by technological advancements. The influence of the digital economy has an important impact on this country’s development which encourage the growth of young entrepreneurs in business. The government and other parties’ employees have constructed an increasingly robust supporting infrastructure to aid the progress of information technology. Information technology’s widespread use has facilitated economic growth as technology spreads into every facet of daily life. Information technology encompasses many tools in the information life cycle (from data creation and storage through analysis and presentation). Micro, small, and medium-sized businesses can now compete worldwide with the help of modern information technology.

The acceleration of new technology development and discoveries due to globalization occurring quickly [3, 4]. Entrepreneurs face challenging decisions about whether to update or start replacing their old technology [5, 6]. Information technology with multiple capabilities and features can demonstrate extraordinary flexibility in business [7], improving business efficiency via innovation and job creation. In business terms, increased speed and accuracy can decrease structure or organizational costs. Technological progress that may happen faster than humans can comprehend, so by the time people learn something new, it’s already out of date [6]. Technology enables companies to alter their business models, eliminating the need for certain existing job sections [6]. The revolution in technology is causing entire business systems to change [8]. Realizable technological innovations in business increase marketplace

data and information availability, clients, and competitors [8]. This rapid advancement of technology and its effect on the complexity of the business environment creates challenges and opportunities for a marketing firm [9–11]. It affects tasks and processes for developing and executing opportunities for growth between and within the business's development process [12] and creates ethical issues, especially regarding the employees' and customers' involvement in the business process.

Business development aims to create long-term value for customers, markets, and associated organizations [13]. Companies can build the capacity to become customer-centric and market-driven by utilizing information technology [8]. Technology enables engagement and interaction personalization, allowing businesses to work collaboratively with and learn from customers to provide services that meet their customers' changing needs [14]. In the world of business competition, information technology is getting more challenging, and companies must update the system used in business to win the match. They need technology systems with well-structured information to take action, realize opportunities, and avoid the threat to the company by ensuring society's sustainability.

## **2 Issues in Business Ethics in the Era of Industrial Revolution 4.0**

The majority of research in the field of business ethics focuses on issues that occur in the workplace. The ethical problems include ethics in consumer and advertising practices, ethics in the marketplace regarding the moral dimensions of competitive and anticompetitive behavior, and ethics in the workplace. Workplace ethics covers establishing moral rights and obligations within employer–employee relationships, job discrimination issues, conflicts of interest, and the concerns of justice, utility, and rights [15].

An all-encompassing perspective on corporate ethics is necessary. Incorporating a business ethics program as an integral part of an overarching effort to better the social environment in which businesses operate is possible when the principle of legal compliance is combined with the principle of conformity with moral standards and is tailored to the concept of business social responsibility.

At an exponentially increasing rate, robots and computational systems have increased automation and displaced current capacities. Technological progress raises a slew of ethical and human-interest concerns. Hooker and Kim [16] analyze what role and responsibility corporate companies may and should take for the future of society by highlighting ethical challenges in artificial intelligence (AI). Hooker and Kim [16] examine various problems regarding the future of work, such as the gig economy, technological unemployment, the meaning of work, basic income, and the obligation to hire.

## ***2.1 The Sharing Economy (Gig Economy)***

A sharing economy is frequently associated with the gig economy. Online services make these practices possible [16]. In the context of employment, the past two decades have witnessed a gradual transition from workers employed on a full-time basis to individuals hired on a contract basis for short-term tasks. As a direct consequence of this, there is a substantial “freelance” or “contingent” workforce [16]. Technology is going to be an essential part of this transition. Workers can become more independent from their employers when working remotely rather than in an office.

Technological change necessitates new skills, which are often easier to find in the gig economy than to teach to long-term employees. Economic incentives also play a significant role. Outsourcing work is frequently less expensive than assigning it to employees since sometimes outsourcing the work is inherently more efficient. Aside from that, they are not covered by overtime or labor laws. Contract workers typically do not receive health insurance or other benefits, and their rates are lower than those of employees doing the same job. Workers’ obvious benefits include flexibility and freedom [16]. Flexible workers can take time off for child care, illness, or other family matters without asking for permission. When individuals have more freedom, they can transition more readily between different sorts of work and focus on hobbies, pro bono activities, or artistic production.

## ***2.2 Unemployment Due to Technology***

The AI revolution may have a more radical impact than the gig economy. Much of manufacturing has already been taken over by AI. There are no humans to be found on many factory floors. Business services will be the next to go. More jobs are created as a result of technological advancement. Workers will almost certainly compete with robots for them [16, 17]. Intelligent systems backed by powerful algorithms and extensive data can gain the necessary skills faster than human labor. Robots will take many of the new employment created due to the dynamic nature of capitalism and technological advancement.

## ***2.3 Inequality***

The initial forecasts are untrustworthy, and we have every reason not to put any stock in them. However, the prospect of technological unemployment in the approaching machine age prompts many inquiries into the ethical obligations of the corporate sector. At least two philosophical arguments for why corporations should be held accountable are presented by Hooker and Kim [16]. The first is the “responsibility

principle,” which states that ethical organizations should work to mitigate the negative consequences of technological unemployment. Second, according to Rawls, technological unemployment is acceptable only if the worst-case scenario (the unemployed or those who systematically lack employment prospects) has a basis to maintain the social structure that permits such widespread unemployment.

### **3 the Meaning of Work: Minimum Wage and the Obligation to Hire**

One of the simplest ways to address structural inequalities that do not rely on the market is by implementing a universal basic income. Widespread usage refers to “an income paid by a political society to all of its members individually, without regard to means or job requirements” [16, 18]. Political opposition to taking enough money from productive sectors to finance an adequate subsidy hinders a successful transfer program. The beneficiaries’ lack of choice is a more pressing worry. Government-sponsored transfer programs have stripped many people of agency.

The corporate-purpose literature, including but not limited to shareholder theory and the stakeholder viewpoint, directly addresses how a meaningful life connects to work. It will rely on the corporate model chosen by society in the future decades on whether or not the problem can be primarily addressed.

Re-distribution is based on economic considerations. Many people look to government redistribution initiatives as the solution to technological unemployment. Redistribution in developed economies is mainly used to share the bounty of the industrial sector, which boasts unusually high labor productivity. Increased labor productivity raises living standards since the economy produces more money for the same amount of work.

### **4 Ethical Issues in Artificial Intelligence**

There are three main concerns when thinking about AI ethics. The development of AI means we will eventually come across systems whose actions are impossible to foresee. The three ethical concerns about AI brought forth by Hooker and Kim [16] are (a) the obligation to program AI to act ethically, (b) the obligation to create AI that can be explained, and (c) the obligation to voluntarily “accept” responsibility.

These worries are associated with questions of machine ethics. How can companies that employ autonomous vehicles safeguard their customers from harm, such as rejecting their loan applications because of their race? A new field of study, machine ethics, has evolved to explore solutions to this problem (e.g., Anderson [16, 19, 20]). The most common approach of machine ethics uses human intuition to inform an AI’s training data. This information will be used to develop “universal, harmonic, and

socially acceptable machine ethics norms” [16]. However, making decisions based on gut feelings isn’t an excellent ethical benchmark for self-sufficient computers. Considerable research suggests that contextual cues might influence people’s innate judgment tendencies. The fact that accidents are often brought on by human error makes developing autonomous vehicles all the more appealing. Machine ethics must be created as an alternative if people cannot be trusted to have good moral judgment.

## 5 Challenges in Business Ethics in Industrial Revolution 4.0

Hooker and Kim [16] list potential difficulties that companies may encounter in the real world: Step one is choosing a community model that best fits the company’s goals. Second, there is a need to increase employees’ levels of knowledge. Thirdly, the shift from fixed employees to independent contractor status. Lastly, humans and machines must collaborate since they have much to offer each other. Transitioning to the fourth industrial revolution presents several issues for the business community and organizational managers, as Carroll [21] described. For instance, product development, marketing, and distribution can all go full circle, each time raising new ethical questions often resulting from business activity. Because it speaks to the underlying worries of both citizens and businesses, corporate social responsibility will remain a pressing problem. As information technology continues to push all companies toward a global frame of reference level and function, the tensions between and among the responsibilities placed on businesses to maximize profit, comply with the law, act ethically, and give back to their communities through philanthropy will only increase in complexity.

Carroll [21] argues that the effect of the implementation of the fourth industrial revolution is that businesses are bracing themselves for digitization and Industry 4.0. However, the potential dangers they confront may lessen their level of readiness. Also, the company’s size may be noteworthy. These are some examples of the kinds of challenges that companies may encounter: (1) insecurity of the data that should be reduced, (2) transferring the benefit of IR 4.0 from vision to reality, (3) encouraging the investment of IR 4.0, and (4) call for internal staff qualification and training programs for school and universities.

One of the toughest challenges in business ethics is finding a middle ground between conservative and liberal critics of charitable giving. This effort can be accomplished by addressing insecurity issues, such as data security or Industry 4.0 maturity, and encouraging investments in Industry 4.0 with public funds [21]. It’s difficult to steer corporate giving without alienating particular people or communities. All business activities occur inside a comprehensive legal framework, and all company choices must consider law and economics. Companies also have a responsibility to act ethically [22]. When we talk about moral responsibility, we mean the norms and standards that exist in a given community but are not formally enshrined in legislation.

Business ethics is developed with more robust and normative economics to meet commitments to market stakeholders and build international legitimacy and justifiable levels of trust among non-market stakeholders. It is a more practical and effective way to improve the moral performance of organizations and practitioners [23]. Kiel [1] lists the following challenges associated with the IoT from the perspective of sustainable manufacturing:

1. Technical integration: in implementing the modern IT infrastructure of intra-firm and inter-firm, standardized communication protocols for data interfaces, modernization of the IT infrastructure for internal and external communication,
2. Organizational change: creating an adaptable corporate culture and hierarchy that can grow, top-level management involvement, and persuasion of internal stakeholders like cross-functional teams and employees at all levels.
3. Data and information security concerning links in the vertical and horizontal value chain market equilibrium are shifting due to new business areas and industry-spanning concentration on IoT. This phenomenon is a significant concern since unprotected data access makes businesses vulnerable to cybercrime and industrial espionage.
4. Competition: shifting market equilibrium and dynamic of competition.
5. Cooperation: Connection throughout whole value chains necessitates openness, cross-company cooperativeness, trust, and suitable technology; disruption and shifting industry boundaries promote market entry of new competitors and the contribution of customers and vendors to the value-generation process.
6. Future viability is in jeopardy due to failing to use emerging IoT technologies and adopting insufficient industry standards. Substantial expenditures in technology development, skilled labor, and data security. The creation and implementation of business models centered on the IoT, emphasizing value addition.
7. Financial resources and earnings: The return on investment for deploying IoT is unclear at the moment.
8. Human resources adequate training and development methodologies are required to ensure that workers are qualified to plan and coordinate processes. Long-term employee loyalty is essential in the face of skilled worker scarcity. The ability to translate customer demand into effective solutions comprising modular hardware and software combinations.
9. Public context: regionally limited bandwidth and Internet transfer speed legal data ownership and security regulations.
10. Increased collaboration and involvement lead to a deeper understanding of individual customers' needs.

## 6 The Concept of Business Ethics

Over the last three decades, the public's perception of business ethics has not changed significantly [21]. This perception was changed, however, when Time Magazine described the twenty-first century as a "biotech century" (Time, 11 January 1999), and we began to consider the business and ethical implications of these realities [21]. The ethical implications of advanced technology in the specific industry present challenges. Several potential problems begin to be considered, resulting in a particular sector. The effects of modern innovations like personal computers, the Internet, online shopping, cloning, and genetic engineering demanded more consideration and speculation [21].

Business ethics is a subset of ethics in the business context, focuses on the ethical examination of business activities, and investigates the moral aspect of the business. As a field of study, it is based on theory or concept and concerned with the application. It's also developed because of competing philosophical, social theory, and economic schools of thought with unique guiding assumptions and conceptual frameworks [15].

Business ethics studies the moral justifications behind business decisions and organizational rules [24]. Business ethics is the study of moral principles relating to, or ought to apply to, business organizations and their constituents. This field of study is grounded in the belief that ethical considerations should be considered when conducting business. Studying the morality of business connections is essential for developing and enforcing policies that benefit those who work in or are affected by companies.

In a nutshell, most business ethics initiatives aim to improve the corporate environment for employees to work by fostering a more moral culture at the individual, group, and societal levels. Business ethics programs require a critical examination of the social value, legal standing, and conformity of business conduct to moral principles buried in the regulatory framework of the complex connections between social actors involved in the economic transaction [15].

## 7 Business Ethic Perspectives

All societies and cultures have right and wrong, fair and unfair, and moral and immoral concepts. However, there are three schools of thought regarding business ethics:

1. The universalism school of thought: According to the ethical universalism school, the most basic ideas of good and wrong apply to everyone, regardless of their culture, occupation, or location.
2. The ethical relativism school of thought: contends that the many religious beliefs, cultural practices, and social conventions in different countries and cultures give rise to many distinct ethical standards. Depending on local ethical standards,



these varying standards determine whether business-related acts are correct or incorrect.

- 3. The theory of the social contract: a set of universal moral standards derived from the consensus of many cultures. According to this school of thought, form a social agreement in which all individuals and organizations must observe all situations. Local cultures or groups might decide whether additional behaviors are ethically permitted under this social contract’s parameters.

8 Business Ethics Model

Business ethics models have been developed due to previous research and practice. One of them is a model developed by Svensson and Wood [24]. According to the model, societal norms and expectations shape business ethics, such as institutional responsibilities, increased education, socially responsible management, competition, and an international business with integrity. These factors are expected to influence the company’s perceptions of leadership relationships, employee relationships, shareholder relationships, external stakeholder relationships, etc. Economic consequences, lawful behavior, better corporate citizenship, paying appropriate taxes, being environmentally friendly, retaining employees, retaining services, and product acceptance result from business practices being perceived as ethical. The overall assessment of those outcomes will affect society. Society evaluation, in turn, will influence society’s expectations (Fig. 1).

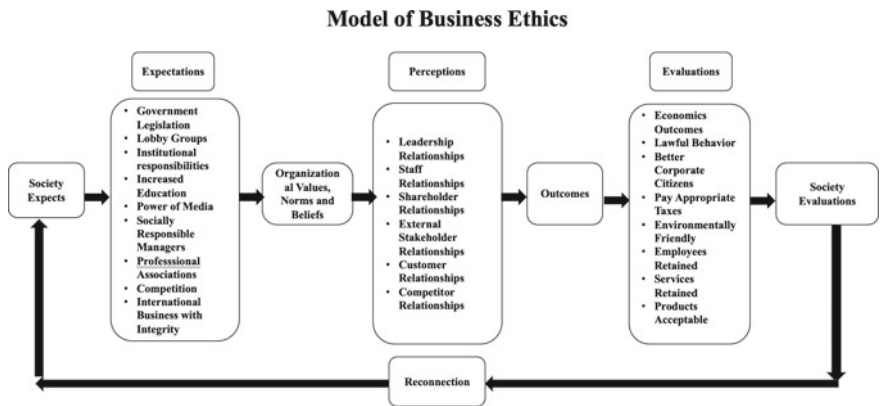


Fig. 1 A model business ethics. Source: Svensson and Wood [24]

## 9 Conclusions: Anticipation Action

To operate ethically, businesses must consider the impact of their decisions on all relevant stakeholders and take actions that maximize the expected social utility of their operations. It's important to remember that the best life for the utilitarian is the one that gives the most to the good of society. Therefore, more opportunities to do so are highly valued [16, 25]. To be decent human beings, we must respect one another by not infringing on or compromising the freedom of others and by helping each person find fulfilment in their own life [16, 24]. Implementing an industry and worldwide agreement that promotes and incentivizes enterprises to meet the accountability of augmentation for all will increase the corporate world's prospects of being more collectively accountable. On the other hand, as we've already established, a basic income wouldn't motivate people to be self-authors who freely exercise economic liberty by engaging in economic cooperation in the second machine age. Business leaders who subscribe to the libertarian ethos should address the risk that an automated economy could lead to a financial system that routinely excludes large segments of the public from participating in the means by which their economic histories are shaped.

Businesses must be aware of the differences in business ethics between countries and industries. This issue appears vital if they intend to export or invest abroad. Inconsistency may result in lower acceptance of the firm's products and services and higher acquisition costs for human or financial resources [26].

Individual and organizational decision-making processes must change for business ethics to be put into practice, as stated by Galukhin [15]. A consistent standard that ensures long-term business success requires: the capacity to make educated decisions, to act following the principles of Advances in Social Science and distributive justice, to protect workers' moral rights, to provide for a diverse set of stakeholders, to consider the consequences of business actions in terms of how they contribute to social welfare and how they conform to the goals of sustainable development, etc. These requirements are the foundation of a morally good corporate culture.

## References

1. Kiel D, Arnold C, Collisi M, Voigt KI (2016) The impact of the industrial internet of things on established business models. In: IAMOT 2016—25th international association management technology of conference processing technology—future thinking, July, pp 673–695
2. Haenlein M, Huang MH, Kaplan A (2022) Guest editorial: business ethics in the era of artificial intelligence. *J Bus Ethics* 178(4):867–869. <https://doi.org/10.1007/s10551-022-05060-x>
3. Joensuu-Salo S, Sorama K, Viljamaa A, Varamäki E (2018) Firm performance among internationalized SMEs: the interplay of market orientation, marketing capability and digitalization. *Adm Sci* 8(3):31. <https://doi.org/10.3390/admsci8030031>
4. Oladimeji MS, Ebodaghe AT, Shobayo PB (2018) Effect of globalization on small and medium enterprises (SMEs) performance in Nigeria. *Int J Entrep Knowl* 5(2):56–65. <https://doi.org/10.1515/ijek-2017-0011>

5. Kaplan DM (2009) Technology and globalization. *Comp Philos Technol* 31:325–328. <https://doi.org/10.1002/9781444310795.ch56>
6. Roos G, Shroff Z (2017) “What will happen to the jobs? Technology-enabled productivity improvement: good for some, bad for others. *Lab Ind J Soc Econ Relat Work* 27(3):165–192. <https://doi.org/10.1080/10301763.2017.1359817>
7. Tavakoli A (2013) Impact of information technology on the entrepreneurship development. *Adv Environ Biol* 7(8):1421–1426
8. Rust RT, Espinoza F (2006) How technology advances influence business research and marketing strategy. *J Bus Res* 59(10–11):1072–1078. <https://doi.org/10.1016/j.jbusres.2006.08.002>
9. Adnan N, Md Nordin S, Hadi Amini M, Langove N (2018) What make consumer sign up to PHEVs? Predicting Malaysian consumer behavior in adoption of PHEVs. *Transp Res A Policy Pract* 113:259–278. <https://doi.org/10.1016/j.tra.2018.04.007>
10. Dhote S, Vichoray C, Pais R, Baskar S, Shakeel PM (2019) Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in ecommerce. *Elect Commer Res Appl* 14:1–16
11. Sun S, Yin Y, Wang X, Xu D, Wu W, Gu Q (2018) Fast object detection based on binary deep convolution neural networks. *CAAI Trans Intell Technol* 3(4):198–207. <https://doi.org/10.1049/trit.2018.1026>
12. Zhao J, Xue F, Khan S, Khatib SFA (2021) Consumer behaviour analysis for business development. *Aggress Viol Behav* 3:101591. <https://doi.org/10.1016/j.avb.2021.101591>
13. Avazdahandeh S, Khalilian S (2019) Estimation of static and dynamic demand function of household water in Qazvin Province and review of the rate of change in consumer behavior over time. *Int J Bus Dev Stud* 11(1):111–126
14. Vargo Stephen L, Lusch Robert F (2004) Evolving to a new dominant logic for marketing. *J Mark* 68(1):1–17
15. Galukhin A, Gusejnov F, Malakhova E, Novikova E (2017) Conceptual frameworks of business ethics. *ICCESSH* 124:709–712. <https://doi.org/10.2991/iccessh-17.2017.172>
16. Hooker JN, Kim TW (2018) Ethical implications of the 4th industrial revolution for business and society. *Research gate*
17. Brynjolfsson E, McAfee A (2014) The second machine age: work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company, London
18. Parijs PV (2004) Basic income: a simple and powerful idea for the twenty-first century. *Polit Soc* 32:7–39
19. Anderson SL, Anderson M (2011) *Machines ethics*. Cambridge University Press, New York
20. Wallach W, Allen C (2009) *Moral machines: teaching robots right from wrong*. Oxford University Press, New York
21. Carroll AB (2000) Millennium: corporate social responsibility and model management morality. *Bus Ethics Quart* 10(I):33–42
22. Carroll AB (1979) A three-dimensional conceptual model of corporate social performance. *Acad Manag Rev* 4:49
23. Petrick JA, Cragg W, Sañudo M (2011) Business ethics in North America: trends and challenges. *J Bus Ethics* 104(1):51–62
24. Svensson G, Wood G (2008) A model of business ethics. *J Bus Ethics* 77(3):303–322
25. Singer P (1995) *How are we to live? Ethics in an age of self-interest*. Prometheus Books, London
26. Scholtens B, Dam L (2007) Cultural values and international differences in business ethics. *J Bus Ethics* 75(3):273–284
27. Thompson MP, MacGregor DG, Dunn CJ, Calkin DE, Phipps J (2018) Rethinking the wildland fire management system. *J For* 116(4):382–390. <https://doi.org/10.1093/jofore/fvy020>

# Sentiment Analysis: Predicting the Position of Islamic Political Parties in Indonesia in the Next Election



Hasse Jubba, Tawakkal Baharuddin, Zuly Qodir, and Suparto Iribaram

**Abstract** The emergence of Islamic political parties was initially thought to represent the voices of the majority of voters in Indonesia, where the majority of voters are Muslim. However, in reality, they could not compete with nationalist political parties. That prompted this study to research Islamic political parties to see opportunities for the next election. This study was conducted using sentiment analysis by relying on the NVivo 12 Plus analysis tool to identify the sentiment results of Twitter users towards Islamic political parties. This study's findings show negative sentiment results for Islamic political parties. It also indicates that Islamic political parties are not ready to appear as parties that dominate elections and challenge nationalist political parties. It also suggests that the nationalist party will remain a strong pivot for the upcoming election period. The result of negative sentiment is also predicted to influence the decisions of Islamic political parties in the future, especially in forming a coalition axis with nationalist political parties. This will consciously encourage Islamic political parties to ignore their own party's ideology. In addition, the result of negative sentiment is also predicted to encourage a change in the future campaign model, which is more focused on inclusive nationalist appeals or exclusive appeals to Islam, especially on social media, to change public sentiment. This will likely be done naturally to secure his political position nationally in the upcoming elections.

**Keywords** Islamic parties · Islamic politics · Predicting elections · Sentiment analysis · Elections · Social media

---

H. Jubba (✉) · Z. Qodir

Department of Islamic Politics, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

e-mail: [hasse@umy.ac.id](mailto:hasse@umy.ac.id)

T. Baharuddin

Government Studies, Universitas Muhammadiyah Makassar, Makassar, Indonesia

S. Iribaram

Islamic Studies, Institut Agama Islam Negeri Fattahul Muluk, Jayapura, Papua, Indonesia

## 1 Introduction

The emergence of Islamic political parties represents the political expression of Islamic society in Indonesia [1]. However, after that, Islamic political parties were deemed to have failed to meet public expectations. This is evident from the recent election results, where Islamic political parties could not compete with nationalist political parties [2]. Ahead of the upcoming 2024 election, Islamic political parties are considered to have an opportunity to rival the strength of nationalist political parties. The reason is that there has always been issues of leadership and representation from Islamic circles in the last several election periods. Even though this momentum raises the issue of religious polarization and has not fully benefited the position of Islamic political parties in the election [3], at least it has spawned massive discussions about Islam and politics [4], especially on social media [5].

Researchers studied how social media is associated with politics [6]. Social media is considered an alternative to reduce financing for politicians and political parties for campaign purposes [7]. Social media can also strengthen interactions between politicians and voters [8]. Social media is also helpful in shaping voter participation in showing political attitudes and expressions [9]. In the context of elections in Indonesia, social media has also proven to accommodate the political interests of politicians or political parties to foster a level of public trust [10]. Social media is also an alternative to Islamic political parties in Indonesia to influence their political positions [11]. Opportunities for politicians and political parties to maximize social media for their political interests depend heavily on the sentiments of social media users. Sentiment analysis focuses on the computational treatment of opinions, sentiments, and text subjectivity [12].

Much research has been done that links political studies with sentiment analysis. However, only a few studies have been found that focus on using sentiment analysis to forecast or predict the positions and opportunities of Islamic political parties in the upcoming elections. However, several previous studies are considered quite relevant. *First*, social media has become a well-known platform for voicing political ideas worldwide [13, 14]. *Second*, sentiment analysis can forecast or predict elections by measuring public discussions and attitudes on social media [15, 16]. *Third*, politicians use social media to influence public discussion by disseminating information about self-profiles and work achievements [17, 18]. It could change the public's evaluation of politicians and political parties.

This study aims to fill the gaps in previous research by relying on the sentiment analysis approach of Twitter users to forecast or predict the positions and opportunities of Islamic political parties in the upcoming elections. The research questions are as follows: (a) What results from Twitter sentiment towards Islamic political parties? (b) How do sentiment analysis results predict the position of Islamic political parties in the upcoming elections in Indonesia? Both of these questions make it possible to determine how public sentiment on Twitter is towards Islamic political parties. In addition, it is also possible to predict the position or chances of Islamic political parties in the next election or future elections in Indonesia. Another reason why this

study is critical is to reduce the gap between Islamic political parties and nationalist parties. It depends on how Islamic political parties can influence public sentiment.

## 2 Method

This study uses a quantitative method with descriptive content analysis. The research subjects were Twitter social media users. The object of research is Islamic political parties. The data source comes from social media Twitter related to topics and issues surrounding Islamic political parties in Indonesia, with Twitter search focusing on searching keywords about Islamic political parties (keywords: PKS party, PKB, PPP, PAN, and PBB) modified on 03 January 2023. The resulting data is a collection of Tweets, where 11.950 Tweets were captured.

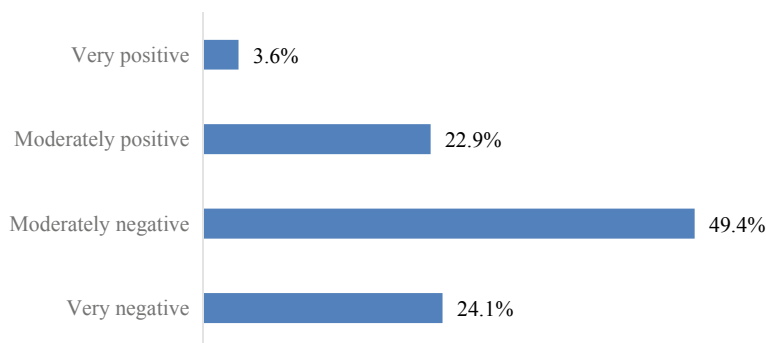
The number of Tweets 11.950 was obtained using the NCapture feature on Google Chrome. NCapture was chosen because it is one of the free web browser extensions that allows researchers to quickly and easily capture content such as data on Twitter. NCapture is also compatible with the analysis tool used in this paper, namely NVivo 12 Plus. The data that NCapture manages to collect is then transferred to an analysis tool like NVivo 12 Plus for encoding the data. Furthermore, the data is classified and coded based on the unit of analysis, namely *identifying sentiment*. Identifying sentiments help identify Twitter users' political expressions towards Islamic political parties in Indonesia. The data coding results were then followed by the visualization stage and analyzed based on data trends and research questions.

## 3 Results and Discussion

### 3.1 Sentiment Analysis on Islamic Political Parties

Analysis of Twitter users' sentiments about Islamic political parties has been successfully identified through Twitter searches and data coding. The results of the sentiment analysis are described as follows:

Figure 1 shows that the sentiments of Twitter users regarding Islamic political parties are identified as still dominantly negative. This identification places Islamic political parties in an unstable position to accommodate their political interests in digital spaces such as social media. It also indicates that Islamic political parties are still not dominant in influencing public judgement on social media such as Twitter. The Islamic political parties identified in this paper all use social media to influence other users' responses. However, they need to be more optimal to accommodate this aspect. It encourages politicians affiliated with Islamic political parties to maximize the use of social media. The more politicians use Twitter to broadcast their thoughts, and the more people participate in discussing it [19].

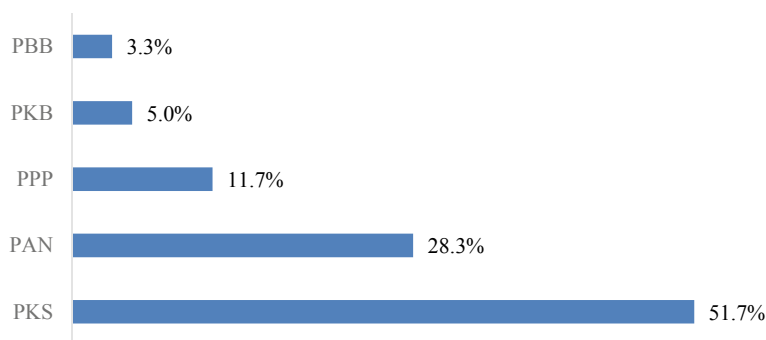


**Fig. 1** Results of sentiment analysis related to Islamic political parties

Islamic political parties need more efforts to change negative sentiments for the better. Approaching the election, social media content generally shows a high level of sentiment from the public, especially in forming political debates [20]. Before elections in Indonesia, Islamic political parties may create social media content closer to the segmentation of other social media users. Social media mechanisms work well if the right messages are delivered to the right people [21], especially paying attention to the segmentation of voters from youth groups. Youth is considered active user of social media [22]. That could only foster youth interest in discussing Islamic political parties and slowly influence the results of public sentiment towards Islamic political parties.

This study also identifies which Islamic political parties are dominant in influencing negative sentiment. The identified political parties are described as follows:

Figure 2 shows that the PKS party is a political party that dominantly influences sentiment results to be negative. The PKS party (51.7%) dominantly influenced sentiment results and was followed by other Islamic political parties, including PAN (28.3%), PPP (11.7%), PKB (5.0%), and PBB (3.3%). There is particular attention to the PKS party on social media. Politically, PKS is quite accomplished because it managed to rank fourth in the 2009 election results, beating PKB, which previously always held the top position representing Islamic political parties [23]. PKS electability decreased after the 2009 elections [24]. Nevertheless, PKS still managed to maintain its position in the top ten votes in the 2014 election [25] and the 2019 election [26]. This is because the PKS can defend itself firmly as a political representative of Islam and as an opposition party [27]. As an opposition party, PKS is the most highlighted by supporters or voters, especially from the nationalist party. This tendency has influenced the discussions and debates of social media users on PKS. This reason also makes PKS the dominant Islamic political party influencing sentiment.



**Fig. 2** Dominant Islamic political parties influence negative sentiment results

### ***3.2 Sentiment Analysis: Predicting the Position of Islamic Political Parties in the Next Election***

The political reality, as illustrated by several previous election results, coupled with the results of negative sentiments by Twitter users, has made the position of Islamic political parties unable to change the substantial domination of nationalist parties for the next election period. As a result, negative sentiment results are predicted to impact the position of Islamic political parties in the upcoming election. The negative sentiment obtained by Islamic political parties is predicted to encourage Islamic political parties to form a coalition axis with nationalist parties. This is a natural political consideration and is relevant enough for Islamic political parties to remain on the axis of Indonesian politics, especially in the next general election. That may be more beneficial for the position of Islamic political parties to continue to guarantee and represent the interests of their voters. In previous election periods, Islamic political parties were also used to form coalitions with nationalist political parties to win candidates [28].

Coalitions between political parties with different ideologies in Indonesia are popular [29]. It draws profits for politicians or political parties who have little power to take roles and get involved in influencing decisions [30]. Islamic political parties can take the coalition path to reduce negative sentiment. Apart from coalitions, the results of negative sentiment are also predicted to affect changes in the model of Islamic political party campaigns in the future, especially on social media. The result of negative sentiment is also predicted to encourage Islamic political parties to appear inclusively. This effort has been observed in several previous political events, where electoral conditions encouraged Islamic political parties to campaign using inclusive nationalist appeals or exclusively Islamic appeals [31]. Another term is moderating its ideological vision or moderating inclusion [32]. It is helpful to open up opportunities to gain support from the public, and slowly reduce negative sentiment.

In order to reduce negative sentiment on social media in the future, Islamic political parties are also predicted to maximize the potential of social media more often. Social



media can bridge social–political attachments and networks between politicians or political parties and other social media users [33]. This means that the communication model relying on social media will be the key to the success of Islamic political parties going forward to initiate public support through online social engagement. Communication models by maximizing the use of social media have been studied to influence the perspectives and attitudes of voters [34], and it depends on how politicians maximize this potential [35]. Using social media by Islamic political parties to change the results of negative sentiment also depends on the information disseminated on social media. It can change public perception and opinion towards Islamic political parties in the future.

Based on the tendency of the negative sentiment results mentioned above, Islamic political parties will be in a difficult position to dominate elections. This illustrates that the nationalist party will still appear as a strong axis for the upcoming election. This study predicts that, in the future, Islamic political parties will continue to form a coalition axis with nationalist parties to secure political status and position nationally, and consistently ignore the ideology of their parties. It also depends on the political situations and conditions in the future. This study also predicts that Islamic political parties will use a campaign model focusing on inclusive nationalist or exclusive Islamic appeals. The negative sentiment analysis results also encourage Islamic political parties in the future to further maximize the potential of social media to change the perceptions and points of view of social media users.

## 4 Conclusion

The political reality, as illustrated by the results of the previous elections, coupled with negative sentiment on Twitter, is predicted to affect the position of Islamic political parties in the next election. This indicates that Islamic political parties cannot appear as political parties that dominate elections. It also argues that nationalist political parties will remain a strong axis for the upcoming election. Negative sentiment results are also predicted to influence the decision of Islamic political parties in the future, to continue to form a coalition axis with nationalist political parties. Consciously, it will encourage Islamic political parties to ignore their party's ideology. In addition, the result of negative sentiment is also predicted to encourage a change in the model of the campaign in the future which is more focused on inclusive nationalist appeals or exclusive appeals to Islam, especially on social media. This is predicted to be done naturally to secure his political position nationally. Islamic political parties can consider the contribution of this research in the future for self-evaluation. This is important to reduce the gap between Islamic political parties and nationalist parties in the upcoming elections. It depends on how Islamic political parties can influence public sentiment. The limitation of this study lies in the data source, which only relies on Twitter to see the sentiment, so future researchers can explore other platforms that contain the same discussion. It may help collect more data to analyze more significant trends.

## References

1. Liddle RW (1996) The Islamic turn in Indonesia: a political explanation. *J Asian Stud* 55:613–634. <https://doi.org/10.2307/2646448>
2. Jubba H, Qodir Z, Abdullah I, Qudsy S, Hidayati M, Pabbajah M, Adawiah R, Iribaram S, Misran M (2022) The contestation of Islamic and nationalists parties in 2019 election. In: International conference on democracy and social transformation, ICON-DEMOST 2021. EAI. <https://doi.org/10.4108/eai.15-9-2021.2315552>
3. Salahudin, Nurmandi A, Jubba H, Qodir Z, Jainuri P (2020) Islamic political polarisation on social media during the 2019 presidential election in Indonesia. *Asian Aff (Lond)* 51:656–671. <https://doi.org/10.1080/03068374.2020.1812929>
4. Lanti IG, Akim, Dermawan W (2020) Examining the growth of Islamic conservatism in Indonesia: the case of West Java. In: Sebastian LC, Hasyim S Arifianto AR (eds) *Rising Islamic conservatism in Indonesia: Islamic groups and identity politics*. Routledge, London, p 26
5. Hui JY (2020) Social media and the 2019 Indonesian elections: Hoax takes the centre stage. *Southeast Asian Aff* 155–174. <https://doi.org/10.1355/aa20-1i>
6. Subekti D, Nurmandi A, Mutiarin D (2022) Mapping publication trend of political parties campaign in social media: a bibliometric analysis. *J Polit Mark* 1–18. <https://doi.org/10.1080/15377857.2022.2104424>
7. Stier S, Bleier A, Lietz H, Strohmaier M (2018) Election campaigning on social media: politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Polit Commun* 35:50–74. <https://doi.org/10.1080/10584609.2017.1334728>
8. Stetka V, Surowiec P, Mazák J (2019) Facebook as an instrument of election campaigning and voters' engagement: comparing Czechia and Poland. *Eur J Commun* 34:121–141. <https://doi.org/10.1177/0267323118810884>
9. Literat I, Kligler-Vilenchik N (2019) Youth collective political expression on social media: the role of affordances and memetic dimensions for voicing political views. *New Media Soc* 21:1988–2009. <https://doi.org/10.1177/1461444819837571>
10. Widayat RM, Nurmandi A, Rosilawati Y, Natshir H, Syamsurrijal M, Baharuddin T (2022) Bibliometric analysis and visualization articles on presidential election in social media indexed in Scopus by Indonesian authors. In: *Proceedings of the 1st world conference on social and humanities research (W-SHARE 2021)*. Atlantis Press. <https://doi.org/10.2991/assehr.k.220402.032>
11. Widayat RM, Nurmandi A, Rosilawati Y, Qodir Z, Usman S, Baharuddin T (2022) 2019 Election campaign model in Indonesia using social media. *Webology* 19:5216–5235. <https://doi.org/10.14704/web/v19i1/web19351>
12. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng J* 5:1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
13. Jensen MJ (2017) Social media and political campaigning: changing terms of engagement? *Int J Press* 22:23–42. <https://doi.org/10.1177/1940161216673196>
14. Keating A, Melis G (2017) Social media and youth political engagement: preaching to the converted or providing a new voice for youth? *Br J Polit Int Relat* 19:877–894. <https://doi.org/10.1177/1369148117718461>
15. Baharuddin T, Qodir Z, Jubba H, Nurmandi A (2022) Prediction of Indonesian presidential candidates in 2024 using sentiment analysis and text search on Twitter. *Int J Commun Soc* 4:204–213. <https://doi.org/10.31763/ijcs.v4i2.512>
16. Budiharto W, Meiliana M (2018) Prediction and analysis of Indonesia presidential election from Twitter using sentiment analysis. *J Big Data* 5:1–10. <https://doi.org/10.1186/s40537-018-0164-1>
17. Darwin RL (2021) Haryanto: women candidates and Islamic personalization in social media campaigns for local parliament elections in Indonesia. *South East Asia Res* 29:72–91. <https://doi.org/10.1080/0967828X.2021.1878928>

18. Vergeer M, Hermans L, Sams S (2013) Online social networks and micro-blogging in political campaigning: the exploration of a new campaign tool and a new campaign style. *Party Polit* 19:477–501. <https://doi.org/10.1177/1354068811407580>
19. Buccoliero L, Bellio E, Crestini G, Arkoudas A (2020) Twitter and politics: evidence from the US presidential elections 2016. *J Mark Commun* 26:88–114. <https://doi.org/10.1080/13527266.2018.1504228>
20. Dang-Xuan L, Stieglitz S, Wladarsch J, Neuberger C (2013) An investigation of influentials and the role of sentiment in political communication on Twitter during election periods. *Inf Commun Soc* 16:795–825. <https://doi.org/10.1080/1369118X.2013.783608>
21. Demir MÖ, Simonetti B, Gök Demir Z (2021) Political segmentation based on pictorial preferences on social media. *Qual Quant* 1–15. <https://doi.org/10.1007/s11135-020-01082-7>
22. Ida R, Saud M, Mashud M (2020) An empirical analysis of social media usage, political learning and participation among youth: a comparative study of Indonesia and Pakistan. *Qual Quant* 54:1285–1297. <https://doi.org/10.1007/s11135-020-00985-9>
23. Hamayotsu K (2011) The political rise of the prosperous justice party in post-authoritarian Indonesia: examining the political economy of Islamist mobilization in a muslim democracy. *Asian Surv* 51:971–992. <https://doi.org/10.1525/as.2011.51.5.971>
24. Suryani (2017) Disalignment of political parties in Indonesia : study on declining electability of prosperous justice party (PKS) in general election 2014. In: Third international conference on social and political sciences (ICSPS 2017). Atlantis Press, pp 201–203. <https://doi.org/10.2991/icsps-17.2018.42>
25. Kramer E (2014) A fall from grace? “Beef-gate” and the case of Indonesia’s prosperous justice party. *Asian Polit Policy* 6:555–576. <https://doi.org/10.1111/aspp.12137>
26. Ummah SM, Rivai AB (2020) Candidacy transformation as electability factor of the legislative candidate from prosperous justice party (PKS) in legislative election 2004–2019. *Polit Indones* 5:437–454. <https://doi.org/10.15294/ipsr.v5i3.21963>
27. Fossati D (2019) The resurgence of ideology in Indonesia: political Islam, Aliran and political behaviour. *J Curr Southeast Asian Aff* 38:119–148. <https://doi.org/10.1177/1868103419868400>
28. Ufen A (2008) From aliran to dealignment: political parties in post-Suharto Indonesia. *South East Asia Res* 16:5–41. <https://doi.org/10.5367/000000008784108149>
29. Prianto AL, Nurmandi A, Qodir Z, Jubba H (2022) Does collective action institutionalize rational Choice? candidate selection in Indonesian political parties. *J Lib Int Aff* 8:63–82. <https://doi.org/10.47305/JLIA2283063p>
30. Pedersen HH (2010) How intra-party power relations affect the coalition behaviour of political parties. *Party Polit* 16:737–754. <https://doi.org/10.1177/1354068809345855>
31. CA Fox, J Menchik (2022) Islamic political parties and election campaigns in Indonesia. *Party Polit* 1–14. <https://doi.org/10.1177/13540688221091656>
32. Rofhani R, Fuad AN (2021) Moderating anti-feminism: Islamism and women candidates in the prosperous justice party (PKS). *J Curr Southeast Asian Aff* 40:156–173. <https://doi.org/10.1177/1868103421989076>
33. Chen HT, Chan M, Lee FLF (2016) Social media use and democratic engagement: a comparative study of Hong Kong, Taiwan, and China. *Chinese J Commun* 9:348–366. <https://doi.org/10.1080/17544750.2016.1210182>
34. Jeroense T, Spierings N (2023) Political participation profiles. *West Eur Polit* 46:1–23. <https://doi.org/10.1080/01402382.2021.2017612>
35. Macafee T, McLaughlin B, Rodriguez NS (2019) Winning on social media: candidate social-mediated communication and voting during the 2016 us presidential election. *Soc Media Soc* 5:1–10. <https://doi.org/10.1177/2056305119826130>

# Digital Leadership in the Development of Digital Competencies in Voter Education Service



Titin Purwaningsih, Bambang Eka Cahya Widodo,  
Moch Edward Trias Pahlevi, and Azka Abdi Amrullobbi

**Abstract** In this digital era, organizational leaders must develop their leadership capacity according to the needs and demands of the community, especially in the public service process. The study aims to analyze the implementation of digital leadership by the General Election Commission (KPU) and the Election Supervisory Body (Bawaslu) in providing voter education services. Considering the decline in the Indonesian democracy index referring to the release, it was recorded at 5.37, down from the previous 5.44. One of the biggest problems is civil liberties issues on social media. So voter education is essential. Implementing digital leadership is one of the keys to building positive narratives or opinions related to voter education. This research uses qualitative methods with data collection techniques, focus group discussion and documentation, and data analysis techniques using Nvivo 12 Plus as artificial intelligence software. This research found that the KPU and Bawaslu still have several obstacles, such as the absence of a framework, SOPs, modules, minimal budget, and minimal human resources qualifications to conduct voter education in the digital world. Even in terms of infrastructure, both are ready and have adequate tools to produce voter education content. It means that KPU and Bawaslu still need to fulfill the values of digital leadership.

**Keywords** Digital leadership · Digital competency · Voter education · General Election Commission · Election Supervisory Body

---

T. Purwaningsih (✉)

Doctoral Program of Government Affairs and Administration, Postgraduate Faculty, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia  
e-mail: [titin.p.widodo@gmail.com](mailto:titin.p.widodo@gmail.com)

B. E. C. Widodo

Undergraduate Program of Government Affairs and Administration, Faculty of Social and Political Sciences, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia

M. E. T. Pahlevi · A. A. Amrullobbi

Komite Independen Sadar Pemilu (KISP), Yogyakarta, Indonesia

## 1 Introduction

COVID-19 has taught election organizers that a series of election activities, including election campaigns, can be carried out online. Mass media and social media, including digital media, are a means of campaigning and building public opinion. The Internet is no longer a new thing, where we can access all information via the Internet. It is a challenge for election administrators to carry out various activities online, including efforts to conduct voter education. In the digital era, voter education is no longer carried out conventionally, but needs to adapt to developments in information and technology. Therefore, we need visionary leadership to carry out institutional tasks and public services, known as digital leadership. Digital leadership is a process of social influence mediated by modern information technology to support changes and improvements in organizational behavior and performance in all stakeholder groups [1]. Digital leadership is inseparable from digital technology, technology leadership, and electronic leadership (e-leadership) [2, 3]. However, digital leadership means the use of digital tools to encourage learning and take advantage of digital literacy in societies where people can access information through digital technologies, such as Internet platforms, social media, and mobile devices [2].

The role of social media as a means of accessing news by the public is increasing, although it tends to be followed by trends in the spread of hoaxes, hate speech, and the like. This phenomenon also occurs in elections in Indonesia. The Ministry of Communication and Information (Kemkominfo) also released about the occurrence of disinformation during the 2019 general election, especially before voting day. According to the author, one disinformation is using electronic identities (E-KTP) to vote at polling stations (TPS) anywhere, even if it does not match the address on the E-KTP [4]. The correct information is that voters who do not have a C6 form (voter invitation) are allowed to use their E-KTP to come and vote at TPS according to their E-KTP address.

The following phenomenon strengthens issues of ethnicity, race, and religion on social media. It causes political polarization. Political polarization is more concerned with religious and ethnic identity politics, not with candidate pairs' different programs. In the 2019 presidential elections, the political campaign war on social media between the camps did look sharper and more challenging. Competition between supporters sharpens polarization [5]. The polarization that occurs is not only related to political issues but extends to geocultural polarization, related to the map of the distribution of voters from the two parties. Polarization and competition between camps are reflected in the competition for news and opinions on social media. Hoax information, black campaigns, and negative campaign color issues in the media. Therefore, political education for voters in responding to information in the media is essential, especially from the General Election Commission (KPU) and Election Supervisory Body (Bawaslu) as election organizers.

This study discusses voter education services provided by the General Election Commission (KPU) and the Election Supervisory Body (Bawaslu). Socialization,

voter education, and community participation are regulated based on General Election Commission Regulation (PKPU) Number 10 of 2018, article 4. Socialization, voter education, and community participation, as referred to in Article 3, aim to: (a) socialization regarding the election stages, schedule, and program; (b) increase public knowledge, understanding, and awareness about rights and obligations in elections; and (c) increasing voter participation in elections. Furthermore, the Law of the Republic of Indonesia Number 6 Year 2020 encourages outreach and campaigns that prioritize the use of digital technology. The use of digital technology is related to the COVID-19 pandemic. The use of digital technology is inseparable from the vision of institutional leaders. Therefore, leaders with digital leadership values are needed who are willing and able to communicate in new ways, channels, and tools with more emphasis on critical thinking, communication, and collaborative ways in an integrated manner. As experts argue, digital progress has formed a new leadership style concept that can unite different generations to work together [6].

This research focuses on three points: the application of digital leadership in voter education; human resource competency in digital leadership; and challenges and supports in implementing digital leadership and developing digital competencies for voter education at the General Election Commission (KPU) and the Election Supervisory Body (Bawaslu).

## **2 Theoretical Framework**

### ***2.1 Digital Leadership***

There are many definitions of digital leadership. Digital leadership is leadership based on the use of digital technology. From a company perspective, digital leadership is the ability of company leaders to find out and make decisions in developing their business using digital business technology [7]. From an organizational perspective, digital leadership is an organizational style to generate organizational knowledge growth by optimizing digital technology [8]. In principle, digital leadership carries out digital transformation, which aims to improve organizational performance.

In order to achieve organizational goals in the digital era as it is today, there are many aspects in its fulfillment, including the element of leadership or digital-minded leaders. Because the success of an organization is not only measured by the performance of its staff or personnel, the most important thing is the competence factor of organizational leaders. A new leadership style is needed that has entrepreneurial skills [9] and even a dynamic digital leadership trait is needed to drive digital transformation [10]. There are several aspects of the quality of digital leadership. It depends on the use of digital technology, support for digital transformation, technology-based professional development, a digital learning culture, and digital leadership skills, including technology, managerial skills, and individual skills [11]. The quality of digital leadership is determined by digital facilities and infrastructure, professional

human resources who master digital technology, the internalization of digital culture, and the quality of digital and managerial mastery of the leaders themselves.

In addition, digital leadership will also support the running of public services. Although sometimes, there are still some obstacles, such as low motivation, service openness, and employee work ethic [12]. The public's concern for public services is public participation and control over the quality of public services. Public demands for fast, transparent, and accountable public services make the government need to continue improving its service quality. Government efforts in terms of public services, including public services for radio frequency spectrum licensing, need to be supported by a digitalization process related to (1) systems, (2) institutions, (3) human resource competencies, and (4) infrastructure.

### 3 Method

This study uses a qualitative method. Qualitative research methods involve several essential efforts, such as asking questions, collecting specific data from informants, analyzing data inductively, and interpreting the meaning of the data [13]. The data collection technique in this research is through focus group discussion (FGD). This data collection technique in qualitative research aims to find the meaning of a theme according to the understanding of a group. The FGD participants are from the member Local Election Commission and Local Election Supervisory Body from North Sumatera, Jakarta, Central Java, West Java and East Java. The five provinces are considered based on the number of voters and Indonesia's most significant number of Internet users.

Based on this, the researcher asked in-depth questions conducted in the FGD to all informants related to digital leadership carried out by the informants. Data analysis uses the Nvivo 12 Plus software, an artificial intelligence software. Artificial intelligence is used to find a problem and solve complex problems in various problems in fields of business, corporations, and government [14]. Artificial intelligence seeks to explain and imitate intelligent behavior through computational processes [16].

### 4 Results and Discussions

Digital leadership is needed in the current digital transformation process to oversee change and use technology quickly on various issues, including the issue of democracy. The General Election Commission (KPU) of the Republic of Indonesia and the General Election Supervisory Board (Bawaslu) of the Republic of Indonesia are institutions that specifically organize General Elections (Elections) as a manifestation of democracy. Both also have the duty and function of providing voter education.

Implementing leadership decision-making in the digital era does not by the leader himself because of his power, control, and position. However, it is a collaboration



between several aspects [15]. This paper will focus on the four aspects of the KPU and Bawaslu's digital leadership: the system, institutional, human resource competency, and infrastructure aspects.

#### ***4.1 General Election Commission (KPU) on Implementing Digital Leadership in Digital-Based Voter Education***

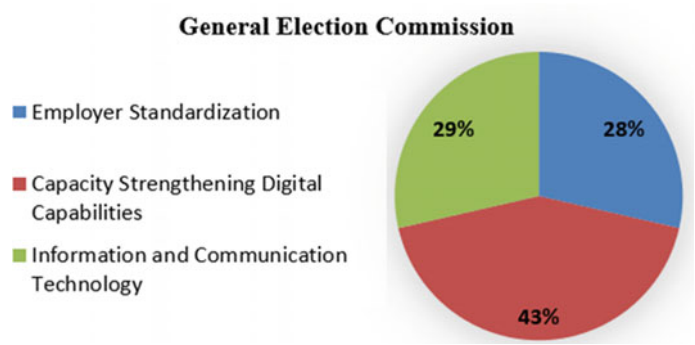
Voter education is the task of election organizers, especially for the General Elections Commission (KPU), both at the central and regional levels. Based on Law Number 7 of 2017 concerning elections, KPU is mandated to educate the voter. Concerning outreach, technical instructions for implementing voter education are shown in General Election Commission Regulation (PKPU) No. 10 of 2018. Voter education is an essential activity in raising public/voter awareness. Society needs to be smart in dealing with the phenomenon of democracy. With community intelligence, voters can determine the choice of leaders responsibly.

In the post-truth era, it encourages changes in people's behavior, one of which is how to use technology to get information. So much information, can have positive and negative impacts, especially on political phenomena. The positive impact is that people can easily get information about what the state or government is doing. However, the negative side is that if it is not wise to use digital technology, it will encourage freedom, resulting in conflict, namely mutual insults, slander, and the use of racial issues for political interests.

Based on a phenomenon in the 2019 election, the election agenda has resulted in polarization, which illustrates that voters are not smart enough to respond to an election contest. It is one of the duties of the General Election Commission to educate voters. Voter education is currently limited to face-to-face activities and leads to digital or social media. With the fast-changing times, there is a need for adaptation in implementing digital-based education. However, in practice, the General Election Commission institution experienced several challenges and obstacles, namely human resource, system, institutional, and infrastructure problems. This study will look at the four problems faced by the General Election Commission.

Figure 1 shows that the most dominant way to implement digital leadership in voter education is strengthening employee capacity in improving skills on social media. Based on the interviews, the absence of special training for KPU employees resulted in KPU employees needing help to create content and narratives that had a positive meaning in society. There is also no standardization of employees in managing social media of voter education. Another area for improvement in human resources is the need for more expertise and competence in the field of communication, which impacts the lack of variety in voter education content on social media. The analysis saw that the absence of special training or strengthening employee capacity on managing social media as a means of voter education shows that the implementing institutions still need to be serious in dealing with the digital era.





**Fig. 1** Human resources in implementing digital-voter education of KPU (*Source NVivo analysis*)

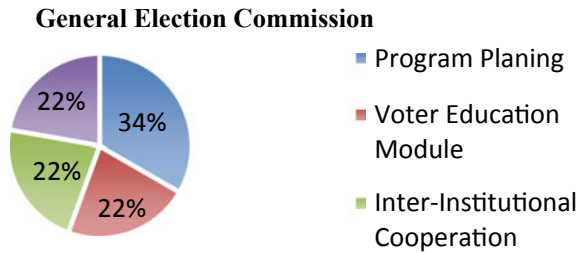
The next challenge is related to the system built by the KPU. Based on the interviews with several provincial KPU, system problems are also an important matter of why implementing voter education could be more optimal. The following are the findings regarding system problems.

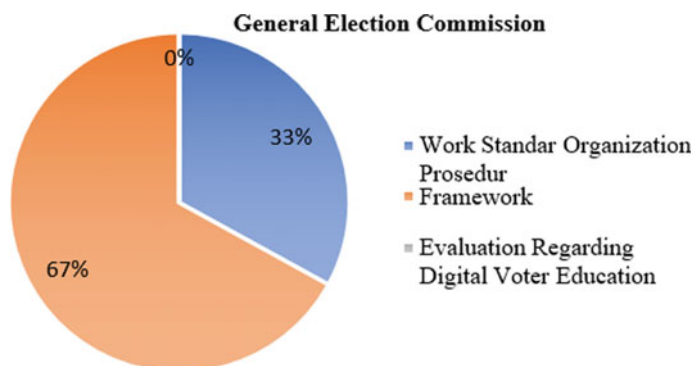
Figure 2 shows no specific program planning for digital-based voter education (34%). The absence of mature program planning has resulted in digital literacy by the KPU for the community only as a complement. Not only program plans, digital-based voter education modules still need to be made available. A lack of cooperation between agencies and a lack of budget causes it. These three aspects where each get a score of 22%, make digital-based voter education less attractive (Fig. 3).

In the institutional aspect, there are two obstacles that the KPU noted, namely the absence of a framework and the absence of SOPs. Both are crucial to ensure that voter education can run properly and reach all voters, in this case the voters, the absence of a framework and SOP show the weaknesses of digital leadership in KPU.

Apart from the obstacles experienced in the system aspect, human resource aspect and institutional aspect, KPU actually has very adequate infrastructure. This can be seen from the value which shows 100%. The KPU is 100% ready to conduct digital-based voter education. Based on the results of the FGD, the KPU already has several infrastructures, such as cameras, video recorders, podcasts, and computers. However, unfortunately, human resources skills, frameworks, and standard operating

**Fig. 2** System in implementing digital-based voter education of KPU (*Source Nvivo analysis*)





**Fig. 3** Institutional in implementing digital-based voter education of KPU (Source NVivo analysis)

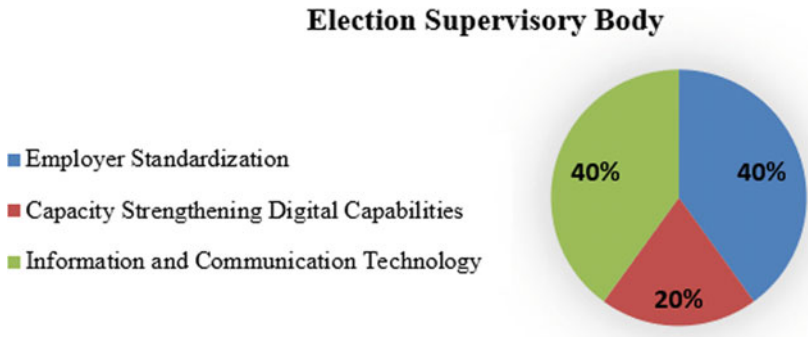
procedures still cannot support the KPU's creativity and productivity in providing political education in the digital world.

#### ***4.2 Election Supervisory Body (Bawaslu) on Implementing Digital Leadership in Digital-Based Voter Education***

Displayed not only the KPU, Bawaslu also has the duty and function of providing voter education to the public. In the midst of technological modernization, Bawaslu also has the challenge of being able to maximize technology as a tool in providing voter education. Not only tools, but Bawaslu must also apply the concept of digital leadership. This is confirmed by the opinions of several experts that technological modernization has formed a new style of leadership concept that can bring different generations to work together [6].

As described in the previous section, digital leadership here is seen from four aspects. *First*, human resources aspect. Based on the analysis, the problems faced by Bawaslu at the provincial level are similar with the KPU at the provincial level problems. In terms of human resources, the biggest obstacle for Bawaslu is the standardization of employees and the need for more skills in information communication engineering (ICT). Both scored 40%. Then, based on the FGD, there has been no standardization for the needs of human resources in managing the digitalization of voter education. So, managing social media is still impromptu, and the content presented to the public is still straightforward. It only fulfills the need for voter education. The mismatch of skills and competencies of the employees also exacerbates this. Many State Civil Apparatus (ASN) employees still need to gain background expertise in ICT or content creation. So, the work of conducting digital-based voter education still needs to be more varied and interesting.

Figure 4 shows that another obstacle in the HR aspect is the need for more capacity building for digital skills. Informants from the regional Bawaslu said there was no



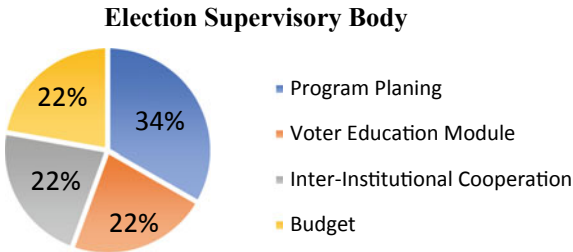
**Fig. 4** Human resouces in implementing digital-based voter education of the Bawaslu (*Source NVivo analysis*)

special training for employees within the Bawaslu to manage social media technology to increase public understanding. Strengthening capacity with training is needed to improve the various abilities and qualities of each individual or organization.

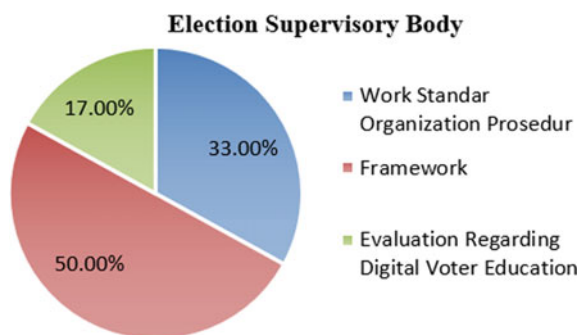
Figure 5 shows that program planning is the most crucial obstacle experienced by Bawaslu in implementing voter education. Program planning is important because it will determine the output and outcome of a program. The problems with planning this program were exacerbated by the COVID-19 pandemic that hit Indonesia. Bawaslu must carry out its duties and functions by maximizing social media or online platform. The Bawaslu RI created a program called Participatory Supervision Cadre School (SKPP) in 2019. The program is a national priority program that aligns with the president’s vision and mission. The originally carried out offline program has now gone online due to the COVID-19 pandemic. The COVID-19 pandemic has impacted changing patterns of activity in society. The pandemic has resulted in many activities involving decisions being carried out online and using data and digital devices [16]. Therefore, digitalization innovation in voter education is crucial.

In the system aspect, implementing voter education in the digital world need a standard module. Primarily, Bawaslu RI must design the module as a guide for provincial Bawaslu, regency/city Bawaslu, and the community. Another obstacle to implementing digital-based education is the need for more budget. The central

**Fig. 5** System problems in implementing digital-based voter education of Bawaslu (*Source NVivo analysis*)



**Fig. 6** Institutional in implementing digital-based voter education of the Bawaslu (Source NVivo analysis)



government often cuts the planned budget. There needs to be a political budget alignment to strengthen human resources' capacity to strengthen digital skills. This lack of budget also explains the need for more enthusiasm for content production and innovation by Bawaslu. As a result, public involvement in participating through digital platforms is minimal. Another obstacle is the need for more cooperation between parties. Collaboration can lighten the workload and improve performance in political education in the digital world.

The third is the institutional aspect. Bawaslu realizes that an immature framework is the most challenging obstacle. The framework is the basic conceptual structure to solve or deal with a complex problem (50%) (Fig. 6).

On the institutional aspect, SOP or standard operating procedure still needs to be improved. Based on the FGD, the Bawaslu RI has an SOP for voter education. This SOP contains guidelines for managing social media, characteristics, interactions of social media, forms of participation in social media, and how to create a virtual space for supervision. Meanwhile, the provincial Bawaslu still needs to have its own SOP but refers to the SOP and guidelines that the RI Bawaslu has made, even though SOPs can cover what staff may and may not do in providing digital-based voter education.

The last obstacle to the institutional aspect is the need for a continuous evaluation of digital-based voter education, with a score of 17%. The results of the in-depth interviews explained that the evaluation is just about accountability for producing voter education content distributed on social media, but not how far the reach and impact of voter education programs are in the digital world.

The fourth is the infrastructure aspect. Infrastructure is essential in running digital-based voter education. With adequate infrastructure, digitalization in the voter education process will work. Information technology infrastructure is the central supporting infrastructure for technological resources in implementing the information dissemination process. According to Turban, Rainer, and Richard [17], information technology infrastructure is the physical facilities, information technology components, services, and information technology management that support the entire company. Information technology infrastructure combines hardware, software, computer networks, facilities, etc. (including all information technology) to develop, test, deliver, monitor, and control information technology services.

Based on the analysis, the infrastructure to support digital-based voter education is sufficient in Bawaslu institutions at both the national and provincial levels. It can be seen from the final value of 100%. The physical infrastructure in each research locus area has met the standards for managing digital content. However, the recognition from many parties in its use is limited. Examples of tools used for audio visuals already exist. Computer support is adequate; it even has a camera to visualize content. The computers provided have met the standards and have sufficient capacity but still need to be improved in creative editing content. In addition, each region already has podcast tools to support the power of voter education.

## 5 Conclusions

Digital leadership is a must-have skill nowadays, especially in state institutions that serve public affairs. Public services must adapt to information and technology developments, so they cannot be carried out conventionally. State institutions, such as the KPU and Bawaslu must be ready to provide public services, especially voter education through digital media.

Much positive information on social media has been eroded by false or hoax information. It is the role of election organizers such as the KPU and Bawaslu to provide organized and creative voter education information through digital space. From this research, there are still many shortcomings that the KPU and Bawaslu have as institutions that implement voter education on social media. The values in the concept of digital leadership have not been fulfilled. It shows the absence of a digital-based voter education system, budget constraints, and limited human resources and infrastructure in implementing voter education.

## References

1. Karakose T, Kocabas I, Yirci R, Papadakis S, Ozdemir TY, Demirkol M (2022) The development and evolution of digital leadership: a bibliometric mapping approach-based study. *Sustainability* 14(23):16171. <https://doi.org/10.3390/su142316171>
2. Ghamrawi N, Tamim RM (2022) A typology for digital leadership in higher education: the case of a large-scale mobile technology initiative (using tablets). *Educ Inf Technol*. <https://doi.org/10.1007/s10639-022-11483-w>
3. Porfirio JA, Carrilho T, Felfício JA, Jardim J (2021) Leadership characteristics and digital transformation. *J Bus Res* 124:610–619. <https://doi.org/10.1016/j.jbusres.2020.10.058>
4. Wiwoho LH (2019) Pemilu 2019 dalam Pusaran Hoaks, Bukti Lemahnya Literasi Digital. *Sorot Media*. [https://www.kominfo.go.id/content/detail/18231/pemilu-2019-dalam-pusaran-hoaks-bukti-lemahnya-literasi-digital/0/sorotan\\_media](https://www.kominfo.go.id/content/detail/18231/pemilu-2019-dalam-pusaran-hoaks-bukti-lemahnya-literasi-digital/0/sorotan_media) (Accessed 07 Apr 2022)
5. Karim AG (2019) Mengelola Polarisasi Politik dalam Sirkulasi Kekuasaan di Indonesia: Catatan bagi Agenda Riset. *Polit J Ilmu Polit* 10(2):215. <https://doi.org/10.14710/politika.10.2.2019.200-210>

6. Aminah S, Saksono H (2021) Digital transformation of the government: a case study in Indonesia. *J Komun Malays J Commun* 37(2):272–288. <https://doi.org/10.17576/JKMJC-2021-3702-17>
7. Zeike S, Bradbury K, Lindert L, Pfaff H (2019) Digital leadership skills and associations with psychological well-being. *Int J*. Query date: 2023-01-10 04:06:27. [Online]. Available: <https://www.mdpi.com/501846>
8. Sasmoko S, Miwardjo L, Alamsjah F (2019) Dynamic capability: the effect of digital leadership on fostering innovation capability based on market orientation. *Manag Sci*. Query date: 2023-01-10 04:06:27. [Online]. Available: <http://growingscience.com/beta/msl/3243-dynamic-capability-the-effect-of-digital-leadership-on-fostering-innovation-capability-based-on-market-orientation.html>
9. Kazim FAB (2019) Digital transformation and leadership style: a multiple case study. *ISM J Int Bus* 3(1):24–33
10. Oberer B, Alptekin E (2018) Leadership 4.0: digital leaders in the age of Industry 4.0. *Int J Organ Leadersh* 7(4):404–412. <https://doi.org/10.33844/ijol.2018.60332>
11. Karakose T, Kocabas I, Yirci R, Papadakis S, Ozdemir TY, Demirkol M (2022) The development and evolution of digital leadership: a bibliometric mapping approach-based study. *Sustain Switz* 14(23). <https://doi.org/10.3390/su142316171>
12. Cahyarini FD (2021) Implementasi Digital Leadership dalam Pengembangan Kompetensi Digital pada Pelayanan Publik. *J Studi Komun Dan Media* 25(1):47–60. <https://doi.org/10.31445/jskm.2021.3780>
13. Dronamraju D (2018) Process improvement strategy for public sector organizations. Linköping University, Linköping
14. Creswell JW (2010) Research design: Pendekatan Kualitatif, Kuantitatif, dan Mixed. Yogyakarta: Pustaka Pelajar
15. Goralski MA, Tan TK (2020) Artificial intelligence and sustainable development. *Int J Manag Educ* 18(1). <https://doi.org/10.1016/j.ijme.2019.100330>
16. Schalkoff RJ (1990) Artificial intelligence: an engineering approach. McGraw-Hill, New York
17. Sarjito A (2019) Model Kepemimpinan digital di Era Revolusi Industri 4.0. Jakarta

# Methodology for the Implementation of FPGA in Technological Applications



Coronel-Villavicencio Edison, Serpa-Andrade Luis ,  
and Garcia-Velez Roberto 

**Abstract** To generate projects with high processing requirements, the use of embedded systems is necessary, and we can appreciate according to the information reviewed in the state of the art that there are many developer modules, which is focused in this work is the study of FPGA and the best methodology. For its application in the practical environment, a methodology is then proposed that involves the study from its hardware description language, its synthesis, adjustments, routing, verification, and analysis of its simulation results, up to the intervention of a manual that involves different modules and syntax for its application in technological applications.

**Keywords** FPGA · Technological applications · Methodology · VHDL

## 1 Introduction

Understanding and properly structuring a learning curve for FPGAs is a complicated task, since knowledge in each of the areas that make up reconfigurable digital systems can be very extensive. It is then possible to structure a logical line of knowledge by areas, which allows learning about FPGAs and achieving the necessary expertise in the labor field.

In the first part of the methodology, you can find the study of gate levels. Gate Level: As it is the input knowledge for FPGA and VHDL, it begins with a general review of VHDL, its structure, and lexical rules, in addition to an introduction to circuit simulation. This preamble is done together with the implementation of simple logic gates, such as AND, OR, NOT, NOR, NAND, and XOR [1–3].

In the next level of Register Transfer (RT-level), MSI circuits are analyzed, such as adders, comparators, and multiplexers; the HDL description of the circuits is analyzed at the module level, as building blocks. In addition, the logic behind concurrent

---

C.-V. Edison · S.-A. Luis (✉) · G.-V. Roberto  
Universidad Politécnica Salesiana GIHEA, Cuenca 010105, Ecuador  
e-mail: [lserpa@ups.edu.ec](mailto:lserpa@ups.edu.ec)

assignments in VHDL is analyzed, and processes are introduced as a way to execute code sequentially [1–3].

Continuing with the methodology, the CAD/CAE tools are those that facilitate the design, synthesis, and implementation of digital systems for FPGA listing them.

- **Vivado:** Vivado Design Suite is a program produced by Xilinx for the synthesis and analysis of HDL designs, which also allows simulation of designs, Xilinx System on Chip design, and high-level synthesis (HLS) design, allowing you to translate high-level code, such as C++, into programmable logic. It has a free (limited) version known as WebPACK. The latest stable version as of the date of this document is 2020.2 [4].
- **ModelSim:** It is a multilanguage environment developed by Mentor Graphics for the simulation of hardware description languages (HDL) such as VHDL, and Verilog. This program is considered as the Gold Standard for HDL simulation since it is independent of the platform and devices used, so it is advisable to get used to simulating in this program. The simulation can be done through the graphical interface, or through scripts. The free version can be found distributed by Intel, in its Lite version 20.1.1 [5].

So a sequential circuit is a circuit with memory, which forms the internal state of the circuit. Unlike a combinational circuit, where the output is a function of its inputs, the output of a sequential circuit depends on its inputs and the internal state of the circuit. Currently, the synchronous design paradigm is used, where all the storage elements are controlled by a single clock, and the data are sampled on the edges of the clock. This methodology is the most important principle for the development of large and complex digital circuits, going into detail are those mentioned below [2].

- **Regular Sequential Circuits:** Regular sequential circuits are those whose outputs present “regular” or repetitive patterns and are the basis for all other types of sequential circuits, e.g., flip-flops, registers, and binary counters [2].
- **FSM:** Finite state machines (FSMs) are mathematical abstractions and are the basic building blocks of digital logic. Unlike regular sequential circuits, next-state logic is designed from scratch, which is why it is sometimes referred to as “random” logic. In practice, the function of the FSM is to act as the controller for large digital systems, where it examines external commands and activates circuits that perform specific actions [2, 3, 6].
- **FSMD:** A finite state machine with data path (FSMD) combines a FSM and regular sequential circuits. These structures can be used to implement systems described by register transfer operations (RT operations), which is the methodology for performing software algorithms on hardware [2].



## 2 State of the Art

Until now it has been assumed that all the knowledge presented in this knowledge curve, in addition to deal with VHDL knowledge, the physical implementation, and development of digital systems, has been accompanied by their simulation. The simulation of VLSI digital systems is extremely important, since, in the industry, implementing wrong designs on FPGAs, ASICs, and so on can be considerably expensive. To accompany simulation and development in VHDL (and basically any other HDL), you need to learn Tcl, since many simulators and synthesis tools use Tcl on their command lines. Tcl: It is the acronym for Tool Command Language and is used as a standardized scripting language for multiple HDL development platforms. Tcl is a relatively simple language, with very basic functionality, but even so, it is possible to create complex programs with it [7, 8].

### 2.1 Crossing Clock Domains

A clock domain is a section of logic where all synchronous elements are governed by the same clock network. Most FPGA digital design books are very strict about using only one clock domain for the entire design; although this would be very useful and would eliminate a number of problems, very often FPGAs are used as an interface means between systems that have their own predefined clock domains, so it is very important to know the synchronization and information management strategies between multiple systems' clock domains [9, 10]. Clock domain crossover (CDC) occurs when two or more clock domains interact with a digital logic block and their frequencies and/or phases are out of sync. There are multiple solutions for the different types of CDC that may occur, among which the following are considered [9, 10].

**Phase control:** It is used to synchronize clocks that have strict relationships with each other, and its phase or delay is a multiple of the main clock. Phase-locked loop (PLL) or delay-locked loop (DLL) is used to branch the secondary circuits from the main one.

**Double flip-flops:** It is a technique used to pass 1-bit signals between two asynchronous clock domains. It uses structures of at least 2 flip-flops to handle the metastability of the input signal. When this method is used, it is not possible to completely predict in which clock cycle the transition will occur. It can be used in control logic where a short period of time (such as one clock cycle) before the response is not critical for operation or synchronization with other communication lines.

**FIFO structures:** It is a more sophisticated way to pass data between two asynchronous clock domains. Some common applications of these buffers are as intermediaries between communication buses and sources of sporadic data generation and that generate information in batches at known frequencies. Given the finite memory

capacity of a buffer, it is necessary to establish certain controls to handle communication, such as handshaking protocols between the domains that read and write to the buffer, or prior knowledge of the transmission type and size. maximum queue that can be generated.

**Gray code:** Gray coding is usually implemented within FIFO structures to pass multi-bit counters between different clock domains only when they are consecutive numbers.

## 2.2 *High-Level Synthesis*

High-level synthesis, sometimes known as C synthesis, algorithmic synthesis, is an automated design process that interprets an algorithmic description, into programming code, and synthesizes digital hardware that implements the algorithm. These tools are widely used in the democratization of digital systems similar to FPGAs, meaning that more people with different academic backgrounds will be able to access and operate these devices. This is seen especially when it is necessary to implement control algorithms, DSP, or artificial intelligence, where the routines are originally designed to behave sequentially (which would make them suitable for programming languages), on hardware, with the aim of to speed up mathematical functions, lighten the computational load of a mixed embedded system ( $\mu P + \text{FPGA}$ ), or parallelize computational processes [11–16].

## 2.3 *Embedded SoC*

As the capacity of FPGA devices continues to grow, these devices tend to integrate specialized hardware blocks to fulfill certain functions, such as integrating microprocessors, and digital signal processors [2].

**Soft microprocessors:** It is a microprocessor core that can be fully implemented using logic synthesis, that is, using the logic and memory elements of FPGAs. In the case of Xilinx, it is known as a microBlaze. They have multiple lines of implementation, such as behavior as 32-bit microcontrollers, real-time processors, and application processors, which generally integrate an operating system [17].

**SoC FPGA:** A System on Chip is a single silicon device that can be used to implement the functionality of an entire system, rather than requiring different physical chips to implement it. In the past, the term SoC was associated with ASIC, but currently, within the FPGA world, it refers to the integration within the same silicon device of a microprocessor system and FPGA logic. In the case of XILINX, it integrates one (or more) ARM processor cores together with FPGA programmable logic, interconnected by communication buses. The design flow for these systems is hybrid; that is, it is necessary to design both the hardware and the software that will use the desired system [18, 19].

## 2.4 *Advanced Simulation of VHDL*

Once the VHDL for synthesis, simulation, and verification of digital systems has been properly understood and handled, it is possible to take a step forward, using new verification methodologies, based on VHDL libraries. These methodologies offer tools for simplifying or automating modeling, report generation, synchronization between simulation frameworks, memory modeling, and more. These libraries are Open Source VHDL Verification Methodology (OSVVM) and Universal VHDL Verification Methodology (UVVM). OSVVM is probably the easiest methodology to learn and the most flexible, since it is a group of packages that are added to the test frameworks (testbenches) as needed. On the other hand, UVVM is more like a framework than a package of tools; to be able to use it freely, it is necessary to structure the entire testbench in the way required for UVVM, but if done this way, you can get great capabilities for tests and modules that are more reusable and compatible. It is even possible to use OSVVM inside UVVM [20–22].

## 2.5 *Technological Applications of FPGA in the Professional Field*

It is important to know the professional fields in which FPGAs are used, but to understand why, you need to understand when to use an FPGA or microcontroller in academic projects. A microcontroller ( $\mu\text{C}$ ) is generally easier to develop than an FPGA hardware. In a  $\mu\text{C}$ , code debugging can be done easily and in real time through JTAG, while in FPGA, JTAG can be used, but only to read fragments of signals and analyze them. So why use an FPGA?

The main reason to use FPGAs is related to performance, connectivity, or reliability issues, since they are vastly superior to  $\mu\text{C}$  and  $\mu\text{P}$ . FPGAs are the devices that have the best time response, and logic can be created to handle all kinds of digital interfaces up to a few hundred megahertz. On the other hand, microcontrollers generally have built-in hardware interfaces such as UART, SPI, or I2C, but if you need speed higher than what a  $\mu\text{C}$  can support, more communication ports per protocol, or specific protocols for communications between peripherals, the solution is FPGAs. In an FPGA, all I/O pins can be reconfigured as high-speed interfaces, plus FPGA chips have a greater number of pins than a microcontroller. Another main advantage of these devices is the ability to perform operations in parallel, without generating procedure queues that could hinder the operation of a  $\mu\text{C}$ . There are other areas in which FPGAs are highly used, such as real-time processing and handling of large volumes of data. Some of the most important applications of FPGAs these days are found in [23–26].

- **Telecommunications:** They are used in both wired and wireless communications. Some of the main fields of development are in WiMAX, 5G/6G, and HSDPA standards. FPGAs are also widely used for DSPs for noise and artifact removal.

- **Defense applications:** FPGAs enable high-tech solutions for RADAR used in defense and electronic intelligence systems. They are widely used in signal processing, from the interpretation of multiple analog channels (high-speed, synchronous reading of sensors) to the handling of military-standard or confidential communication ports.
- **Space applications:** In space applications, they are widely used in data transmission, high-resolution optical and radar data processing, trajectory control with sensors, and video processing to compress and decompress information depending on the available bandwidth. This is one of the fields that largely depends on the use of FPGAs, since they usually need to be remotely reconfigured to correct errors or adapt to new circumstances that are found along the way, adding or removing processing modules as needed.
- **Automotive applications:** Implementation of neural networks at the edge, as well as Light Detection and Ranging (LiDAR), allows new cars their autonomous driving capabilities and risk prevention for drivers.
- **Server applications:** Due to the very high processing capacities of large amounts of data in real time, FPGAs are used for servers or data centers. Their main function is as function accelerators that facilitate the work of the CPUs associated with them.
- **Medical applications:** In the field of medical applications, FPGAs are widely used for diagnostic and monitoring purposes. They are used in equipment to process data, such as MRIs, ultrasounds, and electrocardiographs, and given their fidelity and reliability, they allow secure data to be obtained in critical environments.
- **Artificial intelligence:** FPGAs have become one of the most important focuses of development and implementation of artificial intelligence, reducing the maintenance costs of any system by implementing AI algorithms on the same device, and given their nature of parallelism, they help enormously in the development of machine learning algorithms, even before implementing them on GPUs.

### 3 Proposal

In order to understand how VHDL works and the application to which we assign the hardware description language, we can see in Fig. 1 a diagram of the steps and procedures, which includes two stages, one the development and the other realization.

The present work is a technological research, with a qualitative and experimental approach, where the validity of a manual of practices for the use of the Nexys 4 DDR embedded system for teaching FPGA will be analyzed, in Fig. 2. We can see a scheme that encompasses the study of VHDL and at the same time the applications as well as the progressive model of implementation.

The studies that are carried out have characteristics of pre-experiments, being post-test studies, applied to a group of engineering stakeholders made up of students and teachers, who are exposed to the practice manuals in multiple sessions, and a survey is applied to them that seeks to validate the practices, considering the visual

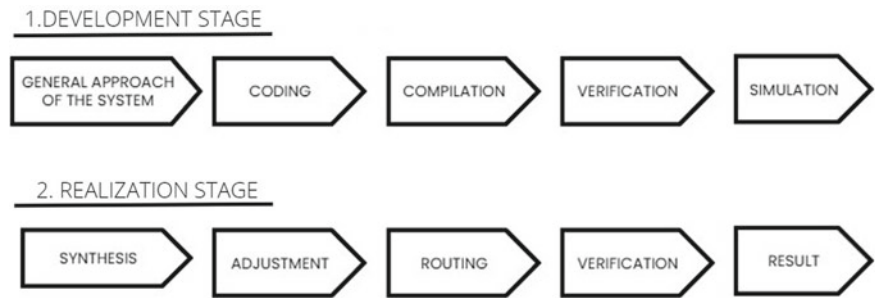


Fig. 1 Development of an application in VHDL

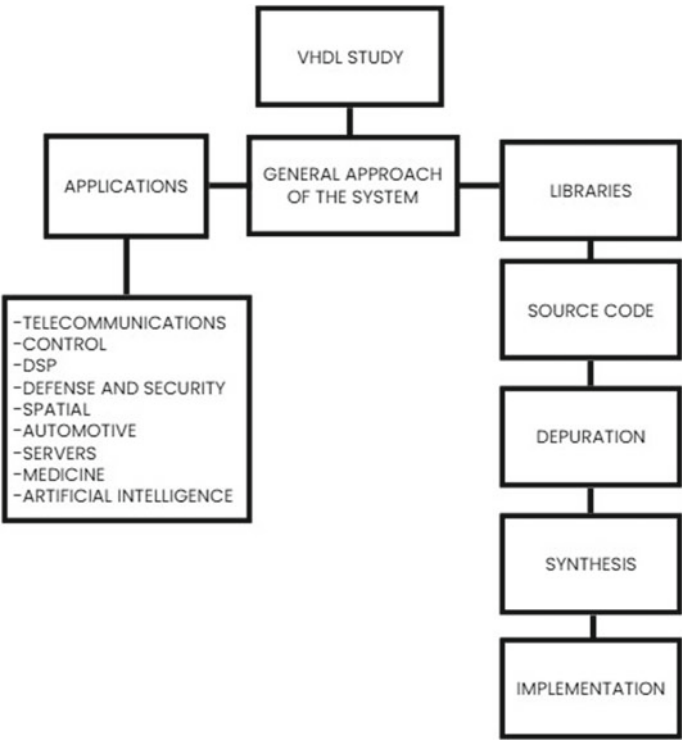


Fig. 2 Study of VHDL applications and progressive implementation scheme

and procedural aspects. In addition, a section is considered where the participating stakeholders can write their criticisms and observations to the manual of practices as feedback to improve its validity. Next, the methodology for the development and validation of the practice manual is described:

1. Analyze the scope of the study of FPGA in the engineering curriculum.

2. Design a practice for each theme, using a developer card.
3. Structure a module of practical guides so that it can be delivered to the test groups.
4. Design the validation tools for the practices.
5. Form focus groups for the application of the practice manuals, which will respond to the validation instruments of the practices.
6. Carry out a satisfaction survey, after the execution of the entire proposed practice manual.
7. Apply the pertinent corrections and observations to the design of the laboratory practices manual.

## References

1. Machado F, Borromeo S, Rodríguez-Sánchez MC (2011) Diseño de sistemas digitales con VHDL. Universidad Rey Juan Carlos, Madrid
2. Chu PP (2017) FPGA prototyping by VHDL examples: Xilinx MicroBlaze MCS SoC. Wiley, Amsterdam
3. Wolf W (2004) Fpga-based system design. Prentice Hall Press, New York
4. Xilinx Inc. (2021) Vivado design suite. <https://www.xilinx.com/products/designtools/vivado.html>. Accessed 05 May 2021
5. Intel (2021) Intel® FPGA simulation; ModelSim\*-Intel® FPGA. <https://www.intel.com/content/www/us/en/software/programmable/quartusprime/model-sim.html>. Accessed 05 May 2021
6. Mealy B, Tappero F (2015) Free range VHDL, p 200. [http://freerangefactory.org/pdf/df344hdh4h8kjfh3500ft2/free\\_range\\_vhdl.pdf](http://freerangefactory.org/pdf/df344hdh4h8kjfh3500ft2/free_range_vhdl.pdf)
7. Ousterhout JK (1994) Tcl and the Tk toolkit
8. TCT (2013) TCL developer site. <http://www.tcl.tk>. Accessed 06 May 2021
9. Kils S (2006) Advanced FPGA design: architecture, implementation, and optimization
10. Churiwala S, Garg S (2011) Clock domain crossing (CDC). Principles of VLSI RTL design. Springer, New York, pp 73–89
11. Maas E, Herrmann D, Ernst R, Rueffer P, Hasenzahl S, Seitz M (1997) Processor-coprocessor architecture for high end video applications. ICASSP IEEE Int Conf Acoust Speech Sig Process Proc 1:595–598. <https://doi.org/10.1016/b978-155860702-6/50064-8>
12. Sun Y, Wang G, Yin B, Cavallaro JR, Ly T (2012) High-level design tools for complex DSP applications. DSP for embedded and real-time systems. Elsevier Inc., Amsterdam, pp 133–155
13. Lucía Ó, Monmasson E, Navarro D, Barragán LA, Urriza I, Artigas JI (2018) Modern control architectures and implementation. Control Power Elect Convert Syst 2:477–502
14. Xilinx Inc. (2021) Vitis high-level synthesis user guide revision history. [www.xilinx.com](http://www.xilinx.com). Accessed 09 May 2021
15. Coussy P, Gajski DD, Meredith M, Takach A (2021) An introduction to high-level synthesis. <http://www.systemc.org>. Accessed 09 May 2021
16. ScienceDirect (2021) High level synthesis: an overview|ScienceDirect Topics. <https://www.sciencedirect.com/topics/engineering/high-level-synthesis>. Accessed 09 May 2021
17. Tong JG, Anderson IDL, Khalid MAS (2006) Soft-core processors for embedded systems. In: Proceedings of the international conference on microelectronics, ICM, pp 170–173. <https://doi.org/10.1109/ICM.2006.373294>
18. Crockett LH, Elliot RA, Enderwitz MA, Stewart RW (2014) The Zynq Book. <https://doi.org/10.1017/CBO9781107415324.004>
19. Intel Corporation (2014) What is an SoC FPGA? Architecture Brief. [www.altera.com/socarchitecture](http://www.altera.com/socarchitecture). Accessed 09 May 2021

20. Osvvm.org (2021) Open source VHDL verification methodology. <https://osvvm.org/>. Accessed 09 May 2021
21. IEEE (2021) osvvm/osvvm GitLab. <https://opensource.ieee.org/osvvm/osvvm>. Accessed 09 May 2021
22. Uvvm.org (2021) Universal VHDL verification methodology. <https://uvvm.org/>. Accessed 09 May 2021
23. Intel Corporation (2021) Consumer electronic FPGA applications: Intel® FPGA. <https://www.intel.com/content/www/us/en/smarthome/products/programmable/applications.html>. Accessed 09 May 2021
24. IAEA (2013) Application of field programmable gate arrays (FPGAs) in instrumentation and control systems of NPPs. <http://www.iaea.org/Publications/index.html>. Accessed 09 May 2021
25. Arrow.com (2018) What is FPGA? FPGA basics, applications and uses. Arrow, 2018. <https://www.arrow.com/en/research-and-events/articles/fpga-basicsarchitecture-applications-and-uses>. Accessed 09 May 2021
26. HardwareBee (2021) FPGA applications: HardwareBee. <https://hardwarebee.com/fpga-common-applications/>. Accessed 09 May 2021

# Technical Requirements Survey on Multimodal Biometric Selection for Deployment in Governments



Mapula Elisa Maeko and Dustin Van Der Haar

**Abstract** Biometric selection is an increasingly popular security measure used in public services due to its potential to improve security and privacy, reduce fraud and improve service delivery in governments. Selecting the appropriate biometrics for an organisation is difficult in today's sphere, as no single solution fits all cases into this issue. However, a need arises to understand the requirements for biometric selection and its potential impacts on the public before deploying biometrics in governments. The strategy was to conduct the requirement with the participants to understand the shortcomings. The survey was distributed to 120 participants from the general public and those with knowledge of licencing centres processes. The results from the survey were derived from the summary on Google Forms as the survey was completed. Furthermore, for better analysis, SPSS 25 statistical research software for analysing data was used as a tool to analyse and interpret the results. The survey, distributed to South Africans between 17 and 55 years of age, had a response rate of 89%. The requirements survey results showed that 68.2% of respondents believed that the current person authentication system at driver licencing centres needed improvement, and they were willing to use a biometric system. This research aims to determine which biometrics are most reliable for identifying and verifying an individual's identity and which biometric selection processes are the most efficient and cost-effective.

**Keywords** Biometric · Multimodal biometrics selection · Security · Deployment requirements · Government

---

M. E. Maeko (✉) · D. Van Der Haar

Academy of Computer Science and Software Engineering, University of Johannesburg, CNR  
University Road and Kingsway Avenue, Auckland Park, Gauteng 2006, South Africa  
e-mail: [elisa.maeko@gmail.com](mailto:elisa.maeko@gmail.com)

D. Van Der Haar

e-mail: [dvanderhaar@uj.ac.za](mailto:dvanderhaar@uj.ac.za)



## 1 Introduction

Biometrics are a rapidly growing technology that is increasingly being used in public with the potential to improve the security and accuracy of public services to protect sensitive information. In this essay, I will present the results of a requirements survey on biometric selection systems to be deployed in public services. It is important to consider the multimodal biometric requirements of the technology to ensure its successful deployment. To assess the suitability of various biometric technologies for public services, a requirements survey was conducted to assess the ease of using biometric technology, its security, privacy, accuracy and cost of deploying biometric technologies. The use of multimodal biometric technologies can reduce corruption, eradicate fraudulent driving documents, identity fraud and improve the integrity and reliability of the driver's licence document [1, 2]. However, choosing the suitable biometric modality or correct combination of biometric modalities is difficult considering the varying complexities in each environment, along with their user base. Selecting multimodal biometric authentication technology for use in a large-scale environment such as public services depend on its needs, the budget cost, technical infrastructure and user acceptance of the technology [3].

The study argues that DLTCs could benefit from multimodal biometric access control that focuses on user awareness, positive perception and the usability assessment of multimodal biometric technologies. A model that promotes selecting appropriate multimodal biometric attributes for biometric authentication solutions in driver licencing centres is proposed. The paper discusses the implications of the survey results and proposes possible solutions for the deployment of biometric technological systems in public services. In the following sections, the study first explores and reviews the literature on selecting biometrics and the survey design, then summarise the survey results, key findings and finally discusses the implications of the results and proposed solutions.

## 2 Problem Background Biometric Selection

The study reviewed the literature to identify the factors that should be considered for the implementation. A literature review was selected to critically explore previous research concerning the guidelines for biometric selection in driver licencing centres and glean information about approaches that have worked and those that have not worked [4]. The literature review explored different types of access control and various authenticators. The purpose was to understand better the type of access control mechanism suitable for implementation in the driver licencing environment. Issues of cost, integration with the existing IT infrastructure, privacy concerns and sustainability that may positively or negatively affect the deployment of the technology solution were discussed. Authentication has become essential in all types of organisations; therefore, improving access control is a must for an organisation

to ensure that only appropriate people receive services. The review gives a better understanding of research in this field and better guidance to support the outcome of the research [5]. Multimodal authentication based on face and fingerprints could be a compatible match for driver licencing centres. The method is used digitally to infuse an individual's fingerprint and face image for a multimodal biometric authentication of the person seeking services at the driver licencing centres. In this context, the fingerprint and face images are captured and saved as a template. The purpose is to improve accuracy and decrease fraud. The reliability of a fingerprint verification system depends on certain factors, such as the manner a fingerprint pattern is presented, the fingerprint sensing and the algorithmic matching level [6].

## **2.1 Face**

To achieve multimodal biometric authentications, face biometrics may be fused with other biometrics such as face and fingerprint, face and ear and face and iris biometrics for identification and verification of a person. In driver licencing centres, iris and face biometrics can be combined for a single multimodal biometric authentication system. The method offers benefits in terms of cost, improved authentication and performance [7]. A single scanner may be used to scan the iris and face images using various scanning capabilities, making it cheaper to use and maintain one scanner rather than deploying two scanners which may be costly.

Furthermore, face and ear biometric images are combined for a multimodal authentication technology, as observed in the study of Patil and Dhole [8]; they integrated face and ear images fused at the decision level and proposed a multimodal biometric recognition system. The advantages of an ear biometric over other biometrics are that the ear is robust for application, and three-dimensional image resolution is realised. The ear biometrics does not change over time. The intensity of the ear image is uniformly distributed, showing less inconsistency, as seen in the case of facial expressions. Face biometrics may be combined with palm print biometrics, and the images together are instantly used for two touchless scanners [4].

## **2.2 Iris**

The ear and iris multimodal biometric authentication methods are proposed in [9]. The method works by acquiring an ear image and an iris image during pre-processing and uses PCA at the feature extraction fusion level to achieve an acceptable rate. Pre-processing of the eye image occurs to segment the iris by detecting the radius of the iris and the pupil. Two phases occur: segmentation and normalisation; normalisation ensures that the texture image of the iris is maintained. For increasing the reliability of multimodal authentication and verification, the study proved to be a better method in multimodal systems where the fusion of biometrics is employed [10]. Therefore,

the ear and iris multimodal authentication may be suitable for law enforcement due to the reliability of this multimodal authentication system for individuals.

The [11] proposed a study on the multimodal biometric system based on the face and iris. The fusion of the face biometric and iris biometric occurs at the feature fusion level, the decision fusion level and the matching score decision level. The fusion results show a better recognition accuracy level than the other multi-biometric recognition systems. In the context of the driver licencing environment, the iris and face recognition may be suitable to realise a desired level of accuracy in the authentication of persons and ensure the system's stability.

### ***2.3 Hand and Palm Biometrics***

The hand and palm multi-biometric recognition system may be either contactless or through a touch screen. In [12], the author proposed a multimodal palmprint technology extracting the right palmprint image and left-hand image; the images showed similarities, and those similarities are used for improving palmprint multimodal performance and accuracy. The fusion level of hand and palmprint is attained at the feature extraction fusion level.

The study conducted by [13] proposed the combination of palmprints and fingerprints for multimodal biometric recognition to improve accuracy and efficiency in identifying a person. Fingerprint and palm print biometrics are the leading modalities in the market today for personal identification. The unique feature of both fingerprint and palmprint is extracted using the Zernike moment invariant algorithm using a low-resolution scanning device for deciding at the decision fusion level. Furthermore, [14] proposed the integration of face and palm prints; multi-biometric feature sets are extracted using Gabor–Wigner Transform, and the extracted feature is normalised or reduced using particle swarm optimisation.

Hand and palmprint authentication systems can be used in driver licencing centres because of the benefits of technology. A single scanner may scan both the palm and hand, reducing the cost of acquiring two scanners. The scanning device needs to have the capability to perform such tasks. The advantages of handprint and palmprint multimodal authentication technology include identification accuracy, increased level of security, privacy and ease of using the technology. However, challenges have been identified using both handprints and palmprints: a database containing two or more biometric traits that belong to one person could be lacking, and execution could be more time-consuming [15]. Other similar work includes the study conducted by [16] that investigated the reasons for fraud in banks and security controls to detect and prevent fraud. Similarly, [17] investigated technology and data analysis advances that banks can use to prevent and combat fraud. Driver licencing centres can significantly reduce fraud losses through specialised multimodal biometric solutions.

Therefore, multimodal biometric technology is cost-effective because a single scanner may scan two biometric traits, thereby reducing security risks. The integration of two or more biometric modalities improves the accuracy of identification significantly [7].

Multimodal biometric technology ensures that the universal population is adequately covered by capturing two or more biometric characteristics. Due to the universal nature of multimodal biometric technology, the system can capture a person’s other form of biometrics, even when a person has a disability [18]. Multimodal biometric technology provides capabilities that prevent an intruder from spoofing the multiple biometric characteristics of an individual instantly. Spoofing of biometric characteristics is removed, thereby increasing the level of security [19].

3 Phases of Biometric Deployment

Implementing multimodal technology to driver’s licencing centres is considered a substantial investment. In this context, deployment is treated as a project for a large population group; therefore, it is referred to as a large-scale deployment project for DLTCs. A biometric deployment project follows various project management phases to maximise deployment success. The following categories are discussed: project planning, technology evaluation, implementation, training and support and maintenance. Figure 1 illustrates the biometric deployment phases showing the various phases that need to be considered to avoid biometric deployment failure.

3.1 Biometric Deployment Planning

A biometric deployment project requires sufficient management to ensure that the overall success of the technology is deployed. Planning ensures a clear deployment purpose, the potential scope creep as well as the direction of the deployment process. During the planning phase, it is common to have a pilot study. The pilot stage ensures that failure is avoided, and issues are faced and resolved before full deployment [20]. The purpose is to ensure that before investing in large-scale deployment is undertaken, the project can be tested on a smaller scale, enabling the technology impact to be assessed, time complexity tested, errors detected and ease of use determined. It



Fig. 1 Phases of biometric deployment

must be noted that piloting a project is not the initial phase of project deployment. In every deployment, there are various stakeholders involved in fulfilling different roles. Multimodal biometric systems are most suitable for organisations where a high level of accuracy and security is required such as the driver licencing department and law enforcement, the financial sector, the healthcare sector and home affairs, [21]. The technical limitation is a drawback identified in the area where a large-scale population authentication technology is used.

The system's performance is also tested against uncertainty and acceptance of the biometric system [22]. Users are essential players in multimodal biometric authentication for deployment, and user acceptance is required. Ensuring privacy and confidentiality of collected personal information and that the information collected is only used for the purpose it is collected. Reliable, accurate and secure biometric authentication systems are widely accepted for large-scale deployment [21].

The project sponsors, whether internal or external, need to be considered. The total cost of biometric technology procurement for rollout to large organisations such as driver licencing centres must be considered, and a tender process must be in place. In South Africa, public entities need to follow the State Information Technology Agency (SITA) procurement process. For any project above R500,000, the project needs to go through a tender process; the process may be lengthy and could result in delays and rescheduling of deployment scheduled dates [23]. The steering committee is a critical part of a project for deployment as the team provides guidance, support and advice for better decision-making about the technology for deployment. Sponsors are critical in monitoring project risks and timelines and providing input for supporting deployment [21].

### **3.2 Technology Evaluation**

The biometric technology under consideration needs to analyse the threat and the risk of deployment, such as user acceptance of the deployed technology. The physical infrastructure, network infrastructure, database and infrastructure for decision-making must be evaluated to ensure enough hardware, software, capacity, and bandwidth support should special technology requirements be desired to meet demands [24]. The return on investment (ROI) significantly contributes to deploying biometric technology in an organisation. The accrual interest in short-term and long-term benefits to the daily production of the workforce is another. Yet other critical factors are the project's capital cost, the consultation fees with various suppliers, the training cost and the testing cost of the project. The biometric technology security issues are considered during deployment phases, such as access rights and access controls to prevent unauthorised access. The threat and risk of deployment must be analysed.

### ***3.3 Implementation***

After all the plans are completed, another phase is implementing the biometric technology within the driver's licencing centre. Here the key is to prioritise the effectiveness of technology deployment, and the plan to ensure a good ROI and that stakeholders are satisfied must be followed [22]. Quality assurance is another critical factor during deployment to ensure the biometric technology functions as intended. Pilot testing and performance testing are essential testing requirements during biometric deployment [20].

### ***3.4 Training***

Training personnel is as necessary as the planning of the deployment project to ensure that the staff are essential to successfully deploy the biometric technology. Training enables personnel to interact, take ownership of the deployment and train other users [24].

### ***3.5 Support and Maintenance***

The successful implementation of the deployed biometric technology requires continuous support and maintenance. The support and maintenance phase ensure that change follows the planned direction and that it is integrated into existing infrastructure and used to fulfil the daily activities in the best way possible. Therefore, planned change management is necessary to deploy a biometric system [19]. The next section discusses the biometric deployment processes.

### ***3.6 Summary Checklist for Biometric Deployment***

The successful implementation of a biometric authentication method in an organisation depends on the adequately followed biometric deployment processes. The entire IT infrastructure and the business organisation are then assessed. Before implementation, the IT unit analyses and determines the future success or failure of the proposed biometric authentication solution to improve the security of information in the organisation [3]. It is to ensure that the deployed solution optimally meets the organisation's needs.

Therefore, an audit of the entire IT infrastructure is assessed to determine how the deployment integrates with the organisation's existing systems and policies. The integration is beneficial to the organisation in that it reduces costs in acquiring new

IT infrastructure and the training and education of staff to ensure user acceptance of the new biometric solution. The hardware equipment chosen for deployment must be effective against security attacks, such as spoofing [24]. The organisation needs to know which type of biometric solution suits its environment. The organisation may consider the task of the biometric solution to cover things such as identification and verification, the location for deployment, the estimated number of users and the type of risk to information if not secured [25]. The choice between a unimodal and a multimodal biometric solution needs some consideration. Multimodal biometric measures offer more security due to multiple layers of protection than a single biometric solution though it may come at the expense of more resources.

## 4 Survey Design

A qualitative research design was selected because of the method's benefits: participants could express their experiences, feelings and thoughts about the problem being explored. It further offers a platform that describes and explains the information [8]. The advantages of selecting a qualitative design ensure a broader perspective of the entire research problem, enabling the exploration of the overall problem using the tools available to collect data. Furthermore, to ensure the validity of the findings, the qualitative design will be applied using a survey to better understand the problem under investigation [26]. The research philosophy used was the interpretive paradigm because of its scientific nature to discover knowledge. The research paradigm, first introduced in 1962 by [27], 'explains the research paradigm as the common framework outlining the cultural research observation, values, beliefs researchers have concerning the nature and conduct of research' [27].

Interpretive research is applicable in this study due to its qualitative process to uncover the most relevant and interesting research questions and interpret them with an understanding of the different participants' experiences. The interpretive paradigm in this context formulates a theory, tests a hypothesis and analyses the findings. Qualitative and interpretive research are used to understand the research problem under investigation and the aspects that cannot be quantified better [28]. The interpretive paradigm in the study describes and explores the scientific knowledge that has been discovered, focusing on the specific issue of authentication within driver licencing departments [5].

Based on the research question, the research design answers the initial plan to be followed, which includes the data collection procedure, the research strategy, the framework, the study population and data analysis [26]. The selected research design gives direction on the most economical, affordable and simple data collection and analysis process for conducting the study research. This study targets the public and those working within driver licencing centres using a survey to develop the requirements. The survey questions were presented in two sections. The first section (A) consisted of questions about the background and general information of the participants. The second section (B) used a Likert 5-point scale to collect data regarding

the participants’ perceptions and the factors affecting the deployment of multimodal biometric authentication.

5 Survey Results

The survey results of participants from the general public between 17 and over 55 showed a response rate of 89%. For this study, 107 of the surveys were completed correctly. The study was set to stop taking any responses once the overall total of 120 survey results had been received. Surveys were purposively and decisively sampled because participants and settings are selected to gather the required information [28]. The survey results were analysed using Google Form response summary and IBM SPSS against the results gathered from the literature review. Purposive sampling is a strategy to select participants intentionally. The observation approach was also conducted during visits to DLTCs to understand the environment better. The survey questions were analysed according to the sections in the survey.

The results were illustrated using frequency tables, bar charts, pie charts and histogram graphs. The frequency tables contain a summary of measures in four columns. The columns show frequency, which include the total number of observations in that category. The column with the percentage indicates the percentage according to the frequencies. The valid percentage column shows the percentage in that category per the total number of non-missing responses. The cumulative percentage shows the sampled total percentage accounted for in that row. After the tables, the graphs are presented [14] (Table 1).

The gender pie graph in Fig. 2 includes females, males and those who prefer not to disclose the type of their gender. It illustrates that 52.33% of the respondents are female, 41.12% are male, and 6.54% choose not to reveal their gender. The results indicate that more females took part in the survey than males (Table 2).

Figure 3 illustrates the awareness of the respondents of challenges experienced in DLTCs. The pie graph indicates that 69.2% of the respondents (the majority) indicate that they understand some problems faced within DLTCs. The results indicate that 69.2% are a considerable number of respondents who have experienced problems, showing a need for intervention to ensure good service delivery. Of the respondents using the DLTC services, 33% indicate that they are not aware of any problem areas.

Table 1 Gender of respondents

		Frequency	%	Valid %	Cumulative %
Valid	Female	56	52.3	52.3	52.3
	Male	44	41.1	41.1	93.5
	Prefer not to disclose	7	6.5	6.5	100.0
	Total	107	100.0	100.0	



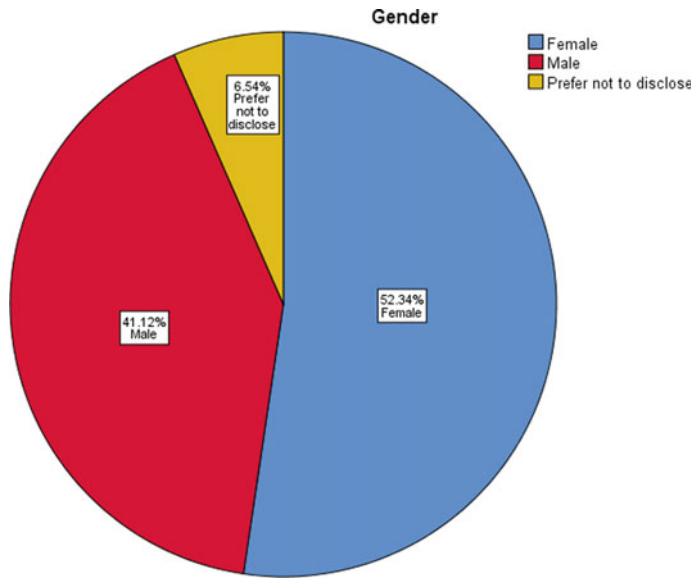


Fig. 2 Gender of respondents

Table 2 Respondent’s awareness of any problem areas within DLTCs

		Frequency	%	Valid %	Cumulative %
Valid	No	33	30.8	30.8	30.8
	Yes	74	69.2	69.2	100.0
	Total	107	100.0	100.0	

Figure 4 illustrates that 45.79% of respondents indicate inadequate identification and verification of applicants, and 24.30% indicate that cases of fraud and corruption (bribery) are prevalent in DLTCs. The results suggest that fraud and corruption are high and require investigation to understand the root cause of the problem area and its impact on the public and road users. Furthermore, 3.74% of respondents indicate a problem of impersonation and identity theft, while 2.80% of the respondents indicate faulty or slow licencing systems. In addition, 23.36% of respondents indicate the problem of issuing fake learner’s and driver’s licences.

In this item, the researcher wanted to establish the respondents’ perception level regarding the need to improve the current driving licence system, procedures and verification of applicants. Table 3 and Fig. 5 illustrate that 68.2% of respondents perceive that the current DLTC system and procedures need improvement, and 27.1% of respondents indicate that there might be a need for improvement. In comparison, 4.7% of respondents indicate that the current system does not need improvement. The study survey results show that most respondents perceive the need to improve the current system at DLTCs; an improvement could bring about a solution to the

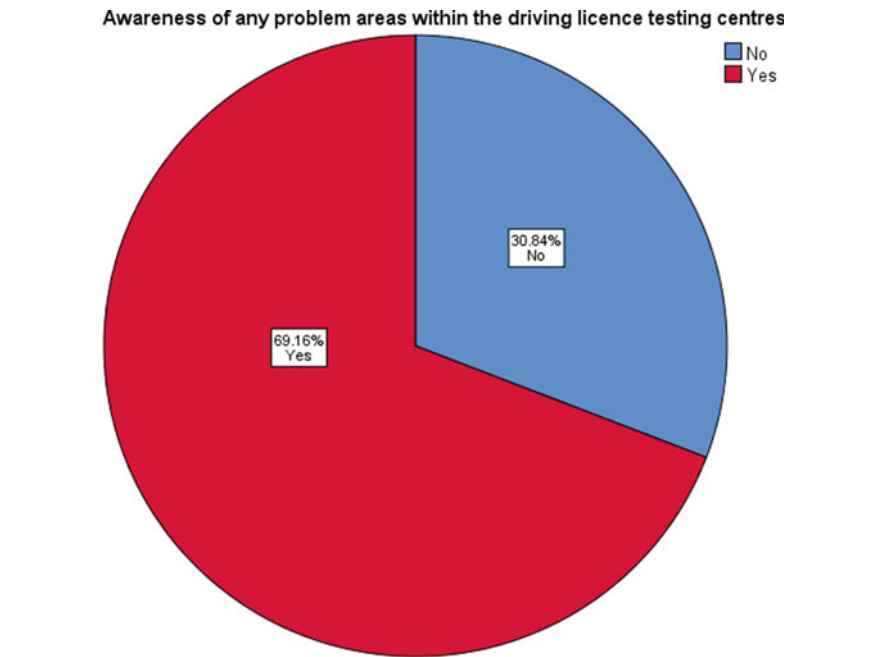


Fig. 3 Respondents’ awareness of any problem areas in DLTCs

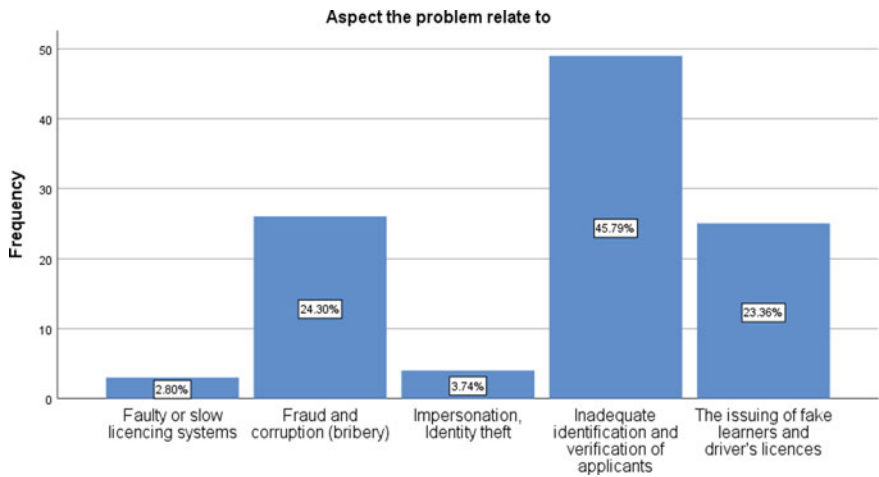


Fig. 4 Problem areas in driving licences

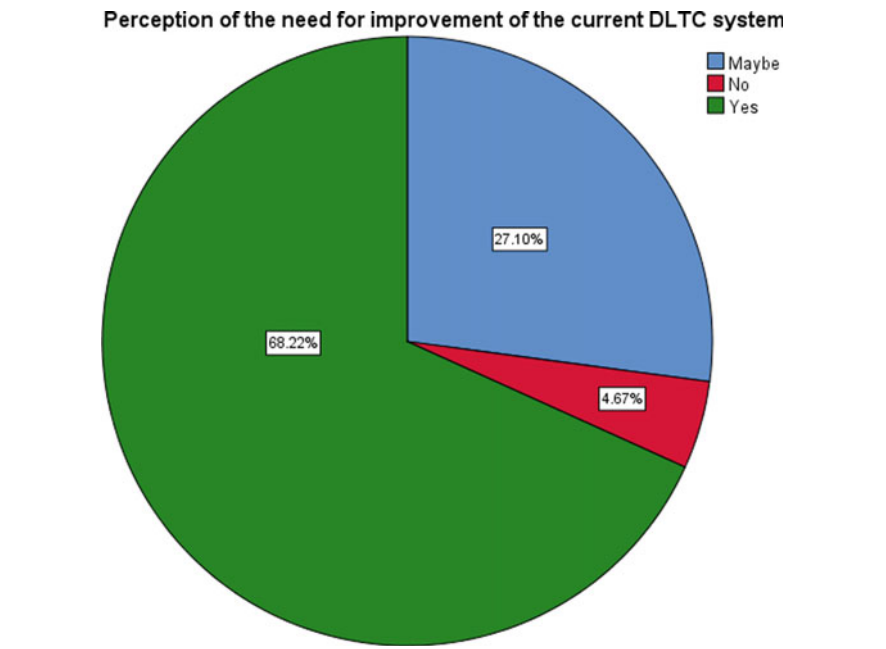
current driver licencing challenges and the urgency of the situation at DLTCs. Figure 6 illustrates that the respondents preferred two combinations of biometric technology.

The researcher asked the respondents to indicate which combination of technology they are willing to use. The graph in Fig. 6 illustrates that 48.60% of respondents are willing to use face and fingerprint technologies, 28.04% of respondents indicated they are willing to use fingerprint and iris technologies, 2.80% are willing to use signatures and voice biometrics, and 9.35% are willing to use face and iris technologies. In comparison, 5.61% of the respondents indicate they are willing to use biometric iris, voice, gait and retina biometrics.

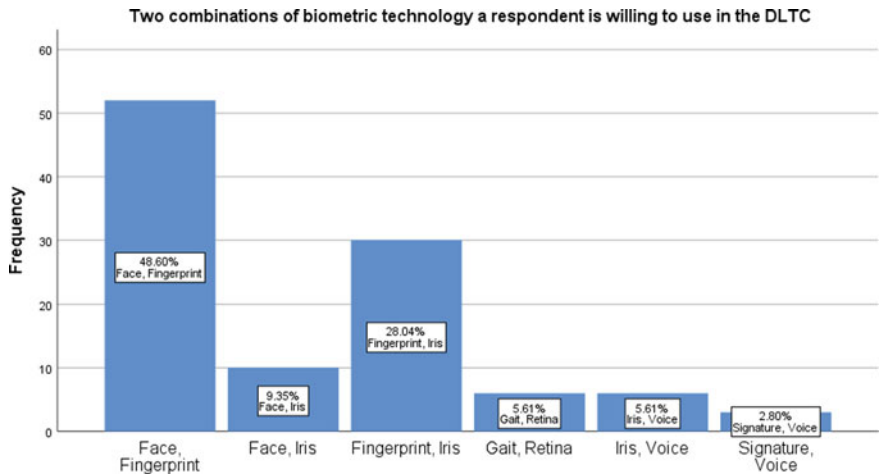
The results indicate that more respondents are willing to use fingerprint, face and iris biometric technology for authentication. Fewer people are willing to use gait, retina, signature and voice biometric technologies for authentication (Table 4).

**Table 3** Respondent’s perception of the need for improvement of the current DLTC system

		Frequency	%	Valid %	Cumulative %
Valid	Maybe	29	27.1	27.1	27.1
	No	5	4.7	4.7	31.8
	Yes	73	68.2	68.2	100.0
	Total	107	100.0	100.0	



**Fig. 5** Perception of the need for improvement of the current DLTC system



**Fig. 6** Two combinations of biometric technology respondents are willing to use in DLTCs

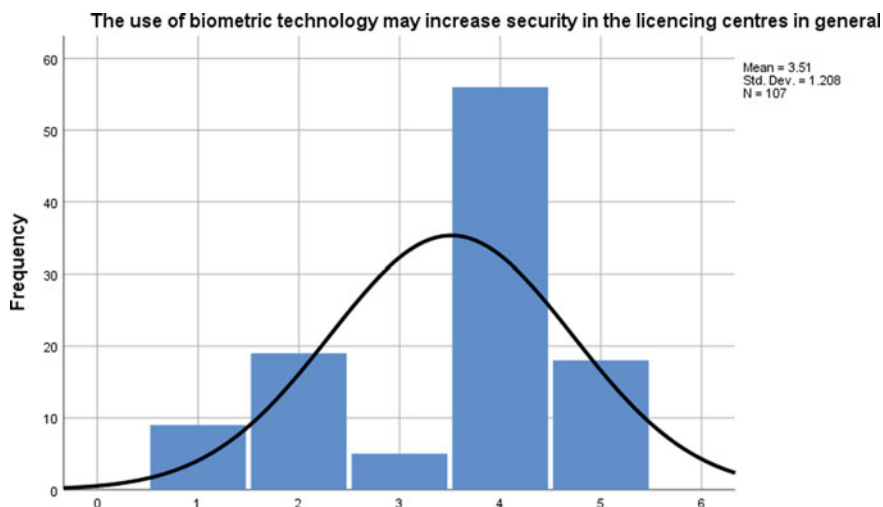
**Table 4** Respondents’ use of biometric technology may increase security in DLTCs in general

		Frequency	%	Valid %	Cumulative %
Valid	1	9	8.4	8.4	8.4
	2	19	17.8	17.8	26.2
	3	5	4.7	4.7	30.8
	4	56	52.3	52.3	83.2
	5	18	16.8	16.8	100.0
	Total	107	100.0	100.0	

Figure 7 illustrates that using biometric technology may increase information security in DLTCs. The researcher explored the respondent’s perception of using biometric technology to increase the security of authentication systems. Figure 7 illustrates that 52.3% of respondents agree, 16.8% of respondents strongly agree, 17.8% disagree, 4.7% are neutral, and 8.4% strongly disagree. Furthermore, the histogram graph shows a mean value of 3.51, a standard deviation of 1.208 and an n-value of 107. The results show that using a biometric authentication technology may increase access control and security in general in DLTCs.

6 Key Findings

The survey results of this study were analysed according to the respondents’ perception, awareness and willingness to use a suitable multimodal authentication method to authenticate applicants and other stakeholders. The results show that 68.2% of the



**Fig. 7** Respondents' use of biometric technology may increase security in DLTCs in general

respondents perceive that DLTCs need improvement to ensure proper authentication to significantly reduce the number of illegal and fake driver's licences on the roads. The results show that the department needs to improve the security and authentication of all stakeholders to reduce identity fraud and improve the administration of the DLTCs significantly. The study explored current issues, such as impersonation, inadequate identification and verification of applicants, fraud and corruption (bribery).

In summary, the most significant findings of this study include the following points:

1. The results show that biometric technology is relatively new to the public, and several concerns require consideration.
2. The general public needs are informed of the use of biometric technologies to understand the benefits these technologies offer.
3. Most respondents have never used biometrics and do not know why a biometric attributes-based system is essential.

The results show that respondents are willing to use multimodal biometric technology, but that, due to a lack of education and training, the acceptability rate of the solution might be compromised. Furthermore, it is revealed that respondents acknowledge the importance of the awareness of biometric technology before its implementation. The respondents acknowledge that biometric verification may improve the speed of licence verification at DLTCs.

Given the study's significant findings, problems exist at DLTCs. The respondents noted that unfair practices, such as fraud and corruption (bribery), that cost DLTCs and place road users at risk are avoidable. Inadequate management and the use of information and communications technology to identify and verify applicants

adequately lead to inefficiencies, a backlog, bribery and fraudulent driving documents [5]. The study established that respondents are willing to use biometric technologies if they improve security and control access to systems and applications. However, the study showed a need for more knowledge regarding multimodal biometric technologies. The wrong selection of a multimodal authentication method in an environment is a critical challenge to successful deployment. About 57% of respondents stated the readiness to use biometric systems to ensure proper authentication in public services. Why using a biometric attributes-based system is essential? Lack of awareness and knowledge of biometric technologies leads to negative perceptions. Adherence to deployment standards and guidelines will be needed to ensure minimal delays and failure.

## 7 Implications of the Results and Proposed Solutions

The results of this survey have shown that biometric selection can be used to identify individuals accurately. However, the survey also indicates implications associated with biometric selection. For example, it could create databases of personal information about individuals. The information could be used for purposes not intended for such as identity theft or tracking a person's movements. To protect individuals from these risks, it is important to take measures to ensure that biometric selection is used securely and responsibly. This could include developing legal and technical measures to regulate biometric selection, such as introducing privacy-enhancing technologies.

Additionally, organisations that use biometric selection should be held accountable for any misuse of the technology. Finally, education and awareness programmes should be developed to inform individuals about the potential risks associated with biometric selection. With these measures in place, biometric selection can be a powerful tool for accurately identifying individuals while protecting their privacy [29].

Technical requirements on the biometric selection are an important consideration to ensure the accuracy and security of biometric systems, as they can guarantee the accuracy of the data collected, prevent unauthorised access and protect the privacy of those using the system. Privacy requirements on a biometric selection are essential. They can be an excellent tool for protecting individual rights and privacy of information, as they protect personal information and create a secure environment. To protect individuals' rights to privacy and maintain the security of biometric data, strict requirements must be implemented to ensure that only the most accurate and reliable biometric selection methods are used. The cost requirements of biometric selection significantly impact its deployment for public services use, as they can affect the affordability of the technology, the complexity of the systems and the rate of adoption and maintenance cost, which may prevent the adoption of the technology in some instances, in this case, the time complexity of the proposed technique.

Biometric selection has advantages, such as being less vulnerable to fraud or identity theft than traditional methods. However, there are also some potential risks

associated with biometric selection. For example, if a person's biometric data is compromised, they may not be able to access their digital identity. Additionally, it is important to train individuals on how to use their biometric information securely. The results of the survey establish that biometric selection has a significant impact on the accuracy of the results.

## 8 Conclusion

Multimodal biometrics offer a unique mechanism to secure crucial sensitive information. The ability of the multimodal biometric system to spot duplicate identities is a crucial advantage of the system. Furthermore, due to the unique method of authentication based on multiple biometrics, there is an increased level of security in using multimodal biometric systems [8]. Due to the availability of multiple biometric characteristics, authentication tends to be reliable. Due to the use of multiple biometric modalities, multimodal biometrics ensure high accuracy in human identification.

Due to various information security threats and the risk of unauthorised exposure, the demand for information security in organisations has increased. A robust, reliable and secure authentication system is a critical requirement for organisations today. The need to render services to the public requires a high-security level to store and process information and ensures that only authorised or authenticated persons can access IT resources. Some users are relatively new to biometric use, and a negative or a positive attitude may influence a user in terms of using or not using the technology. Furthermore, user acceptance of multimodal biometric technologies over the short-term and long-term usage of the system plays a crucial role in the technology for deployment. An increased security level in terms of storage, the security of the biometric template, privacy, ease of using the system and hygiene influence a user's acceptance of the authentication technology [21]. Incorporating a form of training can also improve acceptance. In conclusion, the results of this survey demonstrate that biometric selection is an effective requirement for deployment in public services, provided the proper requirements and safeguards are in place. The research findings of this survey will be used to help public services departments to determine the best multimodal biometric technology for deployment.

## References

1. Chanukya PS, Thivakaran TK (2020) Multimodal biometric cryptosystem for human authentication using fingerprint and ear. *Multimedia Tools Appl* 79(1):659–673
2. El-Abed M (2017) Usability assessment of keystroke dynamics systems. *Int J Comput Appl Technol* 55(3):222–231
3. Kannammal A, Prasanalakshmi B (2009) Analysing security measures with unimodal and multimodal biometrics. In: Paper presented at the international conference on sensors, security, software and intelligent systems: Beijing

4. Onwuegbuzie AJ, Frels R (2016) Seven steps to a comprehensive literature review: a multimodal and cultural approach. Sage, Los Angeles
5. Yin RK (2015) Qualitative research from start to finish. Guilford Publications
6. Kisku DR, Gupta P, Sing JK (2009) Feature level fusion of biometrics cues: human identification with Doddington's caricature. In: Paper presented at the international conference on security technology. Springer, Berlin, pp 157–164
7. David DB, Suganthi K (2017) Survey on integrating face and iris biometrics for security motive using change detection mechanism. *Int Res J Eng Technol (IRJET)* 4(5):283–286
8. Patil VH, Dhole MSA (2016) An efficient secure multimodal biometric fusion using palm print and face image. *Int J Appl Eng Res* 11(10):7147–7150
9. Nadheen MF, Poornima S (2013) Feature level fusion in multimodal biometric authentication system. *Int J Comput Appl* 69(18)
10. Nageshkumar M, Mahesh PK, Swamy MS (2009) An efficient secure multimodal biometric fusion using palmprint and face image. Available from: <https://arxiv.org/abs/0909.2373>
11. Larbi N, Taleb N (2018) A robust multi-biometric system with compact code for iris and face. *Int J Electr Eng Inf* 10(1)
12. Sjarif N (2019) Multimodal palmprint technology: a review. *J Theor Appl Inf Technol* 97(11)
13. Jazzar MM, Muhammad G (2013) Feature selection-based verification/identification system using fingerprints and palm print. *Arab J Sci Eng* 38(4):849–857
14. Saini N, Sinha A (2015) Face and palmprint multimodal biometric systems using Gabor-Wigner transform as feature extraction. *Pattern Anal Appl* 18(4):921–932
15. Ulain N, Hussain F (2020) Fighting governmental corruption in Pakistan/Pakistanska borba protiv korupcije: an evaluation of anti-corruption Strategijesevaluacija antikorupcijskih strategija. *Hrvatska i komparativna javna uprava* 20:439–468. <https://doi.org/10.31297/hkju.20.3.2>
16. Khanna A, Arora B (2009) A study to investigate the reasons for bank frauds and the implementation of preventive security controls in Indian banking industry. *Int J Bus Sci Appl Manage (IJBSAM)* 4(3):1–21
17. Bhasin ML (2016) The role of technology in combatting bank frauds: perspectives and prospects. *Ecoforum J* 5(2)
18. Banyal RK, Jain P, Jain VK (2013) Multi-factor authentication framework for cloud computing. In: Paper presented at the 2013 fifth international conference on computational intelligence, modelling and simulation. IEEE, pp 105–110
19. Dahea W, Fadewar HS (2018) Multimodal biometric system: a review. *Int J Res Adv Eng Technol* 4(1):25–31
20. Jackson G, Read M (2012) Connect 4 success: a proactive student identification and support program. ECU, Joondalup, Australia
21. Goldstein J, Angeletti R, Holzbach M, Konrad D, Snijder M (2008) Large-scale biometrics deployment in Europe: identifying challenges and threats. JRC Scientific and Technical Reports, Seville
22. Lwoga ET, Lwoga NB (2017) User acceptance of mobile payment: the effects of user-centric security, system characteristics and gender. *Electron J Inf Syst Dev Countries* 81(1):1–24
23. Matemotsa MB (2018) Adjudication of bids within the public sector supply chain management process. Doctoral dissertation. University of Pretoria, Pretoria
24. Alcaraz C, Zeadally S (2015) Critical infrastructure protection: Requirements and challenges for the 21st century. *Int J Crit Infrastruct Prot* 8:53–66. <https://doi.org/10.1016/j.ijcip.2014.12.002>
25. Bolle RM, Connell JH, Pankanti S, Ratha NK, Senior AW (2013) Guide to biometrics. Springer, New York
26. McCusker K, Gunaydin S (2015) Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusion* 30(7):537–542
27. Ferraioli D, Kuhn DR, Chandramouli R (2003) Role-based access control. Artech House, London
28. Rajathi A, Chandran P (2019) SPSS for You. MJP Publisher
29. Furnell S, Evangelatos K (2007) Public awareness and perceptions of biometrics



# A Path Recommender System for Enjoyment Improvement of the Cultural Heritage



Francesco Colace, Dajana Conte, Maria Pia D'Arienzo,  
Domenico Santaniello, Alfredo Troiano, and Carmine Valentino

**Abstract** The need for cultural heritage enhancement in the Italian context requires an effort from new technologies. Indeed, various tools can be functional to guarantee the improvement of the artistic artifacts' maintenance or users' enjoyment. This paper aims to present an architecture to improve the cultural experience through recommender systems and digital storytelling techniques. Moreover, the construction of a mathematical model allows for the building of personalized paths. An experimental phase, based on the development of a prototype, permits the evaluation of the proposed architecture, and obtained results are promising.

**Keywords** Recommender system · Context awareness · Path recommendation · Digital storytelling

---

F. Colace (✉) · D. Santaniello · C. Valentino  
DIIN University of Salerno, Fisciano, Italy  
e-mail: [fcolace@unisa.it](mailto:fcolace@unisa.it)

D. Santaniello  
e-mail: [dsantaniello@unisa.it](mailto:dsantaniello@unisa.it)

C. Valentino  
e-mail: [cvalentino@unisa.it](mailto:cvalentino@unisa.it)

D. Conte  
DIPMAT University of Salerno, Fisciano, Italy  
e-mail: [dajconte@unisa.it](mailto:dajconte@unisa.it)

M. P. D'Arienzo  
DISUFF University of Salerno, Fisciano, Italy  
e-mail: [mdarienzo@unisa.it](mailto:mdarienzo@unisa.it)

A. Troiano  
NetCom Group, Napoli, Italy  
e-mail: [a.troiano@netcomgroup.eu](mailto:a.troiano@netcomgroup.eu)

## 1 Introduction

The Italian artistic and cultural heritage require a significant effort to new technologies for achieving enhancement. Then, the information and communication technology (ICT) field developed various research works to respond to the enhancement question. In particular, this topic produces the analysis of tools finalized to the two main ways for enhancement objective: conservation of archeological heritage [1, 2] and improvement of cultural experience [3, 4].

The maintenance of archeological assets requires the use of tools able to predict restoration works. Moreover, it takes advantage of digital tools to operate on digital reproductions of archeological heritage [5]. In this field, data collection through sensors allows for monitoring archeological artifacts and elaborating predictions about future damages [6, 7].

Instead, cultural experience improvement requires satisfying the preferences of the person who interacts with the artistic and cultural heritage. In this case, the achievement of enjoyment improvement takes advantage of the knowledge about the users and the environment in which he enjoys the cultural experience. Indeed, the museum experience differs from archeological sites, and indoor environments represent a different challenge from outdoor ones. To respond to the variety of situations related to the cultural experience, various tools and paradigms allow adapting systems to user preferences and the environment in which he moves.

Meeting the personalization demand exploits tools able to understand what users prefer and suggest the proper cultural object of all those available, namely recommender systems (RSs) [8, 9]. They consist of information filtering tools that analyze users' preferences and provide suggestions among all possible choices.

RSs work on three elements: user, item, and transaction [10]. In particular, the singular transaction represents an interaction between the user and the system. The RSs' main objective consists of suggesting items with which the user did not interact. Based on the modality exploited by the RS for forecasts, RSs can be classified as content-based, collaborative filtering, and hybrid [11]. In particular, content-based approaches calculate rating forecasts creating vectors of features related to users and items [12]. Collaborative filtering RSs take advantage of interactions among users and items [13]. Finally, hybrid RSs exploit two different recommendation techniques to overcome the limits of the singular technique [14]. Moreover, RSs can improve their performance by integrating contextual data [15, 16] in the forecasts' elaboration. The term context represents "any information useful to characterize the situation of an entity that can affect the way users interact with systems" [17]. Contextual data can be integrated into the recommendation phase in three different ways [18]:

- Contextual pre-filtering techniques use contextual data to select inputs for recommender systems;
- Contextual post-filtering approaches filter RSs outputs based on contextual data;
- Contextual modeling approaches integrate contextual data in the forecast elaboration.

RSs that take advantage of contextual data to improve the users' suggestions are defined context-aware recommender systems (CARSs) [19].

This work aims to exploit a CARS to provide personalized paths to users. In particular, this paper will describe an architecture in which the RS is able to understand user preferences and provides the user-cultural object affinity coefficient to elaborate a personalized path in an archeological site. The path creation is enriched through digital storytelling techniques [20, 21]. It allows the improvement of the interaction between the user and the cultural object through the utilization of multimedia related to the visiting item. Then, this architecture permits for improving user enjoyment and cultural experience. The structure of this paper is the following: Sect. 2 presents related works, Sect. 3 describes the proposed architecture and focuses on the analysis of the path engine, Sect. 4 contains the experimental phase, and Sect. 5 presents conclusions and future works.

## 2 Related Works

The recommender systems utilization in the cultural heritage field represent a powerful tool for personalizing the users' experience and providing appropriate suggestions related to cultural objects. Then, the literature presents various examples of approaches that exploit recommendation engines to improve users' enjoyment.

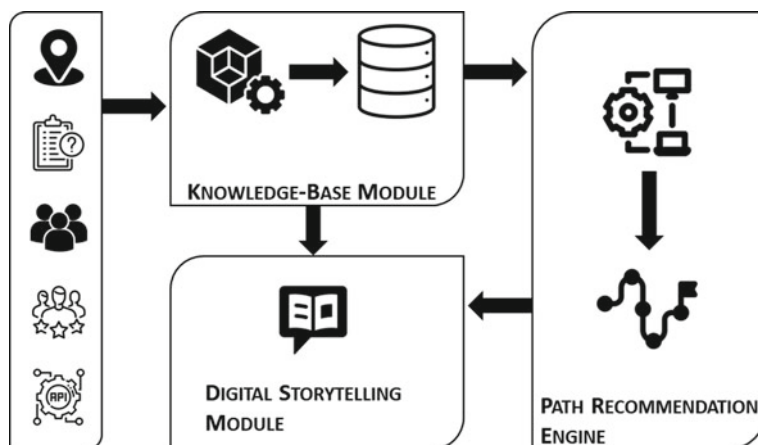
Chianese and Piccialli [22] describe the context evolution system that takes advantage of a context-aware recommendation engine to improve the user experience in the case of the exhibition "The Beauty and the Truth" in Naples. This system exploits context awareness to integrate the environmental characteristics in the services' suggestions.

Bartolini et al. [23] present a content-based recommender engine integrated with context awareness. Contextual data permits the selection of cultural objects based on the location context. Then, after the suggestion elaboration, other contextual information filters the recommendation outputs to the appropriate visiting path creation.

Ruotsalo et al. [24] propose the SMARTMUSEUM approach based on a content-based RS. In particular, this approach exploits Web data and contextual data to describe the environment in which the user acts. Moreover, the creation of users' profiles takes advantage of ontologies and collected contextual information. Then, contextual data allows the selection of the appropriate cultural objects on which the suggestion are elaborated.

## 3 The Proposed Architecture

This section focuses on the proposed architecture finalized for the personalized paths creation and the use of digital storytelling to improve the users' enjoyment. The architecture, summarized in Fig. 1, presents three main components:



**Fig. 1** Proposed architecture

- The knowledge-based module processes and stores data acquired from various fonts;
- The path recommendation engine elaborates stored data and creates personalized paths;
- The digital storytelling module takes advantage of the created path and stored data to present cultural objects to users through images and descriptions.

The architecture operativeness is strictly related to acquired data. It allows the elaboration of the recommendation engine and the digital storytelling module. Then, the data acquisition takes advantage of various fonts:

- User location allows for analyzing user position and selecting the appropriate starting point for the path to elaborate.
- A questionnaire permits the construction of the user profile. If the user does not answer the questionnaire, the user profile will be a vector of equal components.
- The sensors' installation consents to acquire information related to the crowdedness of cultural objects.
- Cultural heritage experts provide information for the construction of cultural objects' profiles.
- External APIs allow the acquisition of multimedia for the digital storytelling module.

The knowledge-based module operates the pre-elaboration of the raw data to make it suitable for data processing. Then, the storage happens through a structured query language (SQL) database [25]. The path recommendation engine works in two phases. The first phase consists of the recommendation process. The system acquires users' and cultural objects' profiles to calculate rating forecasts according to the context-aware recommender system described in [26]. This context-aware recommendation

approach takes advantage of users and items profiles and some data collected for the interaction among users and cultural heritages to elaborate rating forecasts. Then, in the second phase, the system selects elaborated recommendations for the specific user and solves the optimization problem (1)–(9).

$$\max \sum_{i,j=1,i \neq j}^n r_{uj} g_{ij} x_{ij} \quad (1)$$

$$\sum_{i,j=1,i \neq j}^n d_{ij} x_{ij} \leq D_{max} \quad (2)$$

$$\sum_{j=1,j \neq i}^n g_{ij} x_{ij} - \sum_{j=1,j \neq i}^n g_{ji} x_{ji} = \begin{cases} 1 & i = P, P \neq A \\ -1 & i = A, P \neq A \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \dots, n \quad (3)$$

$$\sum_{j=1,j \notin NI}^n g_{ij} x_{ij} \leq 1, \quad i = 1, \dots, n \quad (4)$$

$$\sum_{i=1,i \notin NI}^n g_{ij} x_{ij} \leq 1, \quad j = 1, \dots, n \quad (5)$$

$$\sum_{j=1,j \notin NI}^n g_{Pj} x_{Pj} = 1, \quad j = 1, \dots, n \quad (6)$$

$$\sum_{i=1,i \notin NI}^n g_{iA} x_{jA} = 1, \quad j = 1, \dots, n \quad (7)$$

$$g_{ij} x_{ij} + g_{ji} x_{ji} \leq 1 \quad i, j = 1, \dots, n \quad i, j, \notin NI \quad (8)$$

$$x_{ij} \in \{0, 1\} \quad i, j = 1, \dots, n \quad (9)$$

where

- $V$  represents the cultural objects set and  $n = |V|$  is the number of cultural objects;
- $d_{ij}$ : distance of cultural object  $i$  to cultural object  $j$ ;
- $D_{max}$ : maximum distance to travel;
- $r_{uj}$ : affinity between the user  $u$  and the cultural object  $j$  provided by the context-aware recommender system.
- $x_{ij} = \begin{cases} 1 & \text{if user goes from the cultural object } i \text{ to the cultural object } j \\ 0 & \text{otherwise} \end{cases}$

- $g_{ij} = \begin{cases} 1 & \text{if the cultural object } i \text{ is linked to the cultural object } j \\ 0 & \text{otherwise} \end{cases}$
- $NI$ : set of isolated nodes, namely connected only to another node;
- $P$ : starting point;
- $A$ : ending point.

The optimization problem (1)–(9) has the objective function (1) that aims to maximize the number of visited cultural objects based on the user preferences. The constrain (2) manages the maximal distance that the user wants to traverse, instead constrain (3) imposes  $P$  as starting point,  $A$  as the ending point and other cultural objects as passage points. Constrains (4) and (5) regulate, respectively, the uniqueness of the entrance in the node and the exit from the node. The constrain (6) imposes the exit from the starting point  $P$ , and the constrain (7) imposes the entrance in the ending point  $A$ . The constrain (8) allows to avoid cycles in the isolated nodes. Finally, the constrain (9) requires that variables  $x_{ij}$   $i, j = 1, \dots, n$  have to be binary.

The resolution of the optimization problem (1)–(9) returns the personalized path for the user. Then, the digital storytelling module acquires cultural objects from the selected path and data related to cultural objects to present them to the user.

## 4 Experimental Phase

This section describes the experimental phase projected to test the proposed architecture. The experiment took place in the Archeological Sites of Paestum and Velia and required the development of a prototype. In particular, the resolution of the optimization problem takes advantage of the PuLP package [27].

The experimental phase involved 18 people aged ranged from 19 to 27 years old. They are students of the University of Salerno, and after the developed prototype use, they evaluated the proposed architecture through a questionnaire. The questionnaire is composed of four sections:

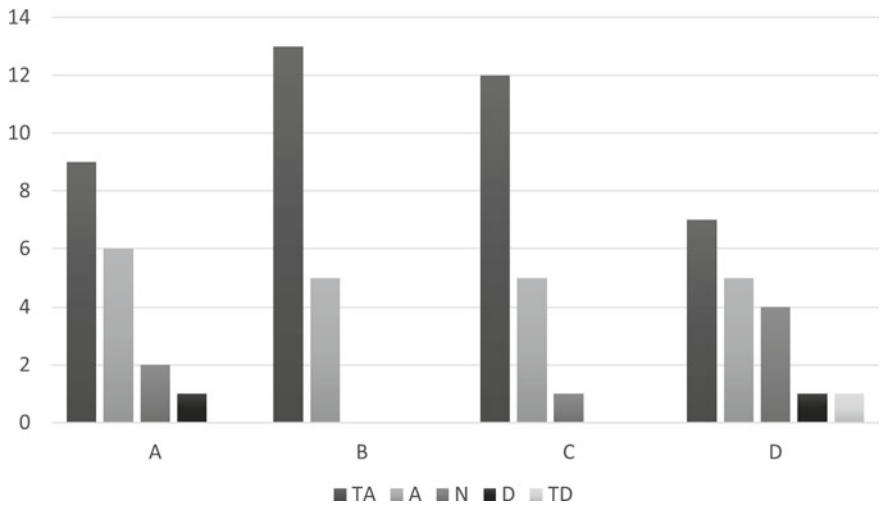
- Section A (presentation) aims to estimate how the proposed architecture presents cultural objects through the digital storytelling module.
- Section B (path recommendation) wants to evaluate the suggested path. In particular, the user has to rate the affinity of the cultural objects with his preferences.
- Section C (reliability) represents the rate of the cultural experience lived by the user.
- Section D (usability) aims to evaluate the usability of the prototype.

Users can choose from five possible responses: totally agree, agree, neutral, disagree, or totally disagree Table 1.

Figure 2 summarizes obtained results. Section B, related to path recommendation, returns the best results because there are only positive evaluations. Then, Section C, related to the enjoyment evaluation, also returns positive reviews, and only one person estimates the experience with the value Neutral. Then, Section A returns good results,

**Table 1** Users’ responses to the questionnaire

	Totally agree	Agree	Neutral	Disagree	Totally disagree
Section A	9	6	2	1	0
Section B	13	5	0	0	0
Section C	12	5	1	0	0
Section D	7	5	4	1	1



**Fig. 2** Questionnaire results

and there are two evaluations equal to neutral and one to disagree. These results confirm that the digital storytelling techniques use improves user interaction with cultural objects. Finally, the worst results are obtained in Section D related to the prototype usability.

### 5 Conclusions and Future Works

This paper presents an architecture to improve cultural heritage enhancement through enjoyment improvement. This architecture takes advantage of a context-aware recommender system, digital storytelling techniques, and the resolution of the optimization problem (1)–(9). The experimental phase confirms the good working through the positive results obtained by Section B, Section C, and Section A. To improve the proposed architecture, the experimental phase has to be increased. Moreover, the prototype has to be improved because of the results obtained by Section D. Moreover, the contextual analysis will be increased through the increase of the sensor typologies.

## References

- Colace F, Elia C, Guida CG, Lorusso A, Marongiu F, Santaniello D (2021) An IOT-based framework to protect cultural heritage buildings. pp 377–382
- Rao BN, Rao BB, Challa NP (2019) Predictive maintenance for monitoring heritage buildings and digitization of structural information. *Int J Innov Technol Explor Eng* 8(8):1463–1468
- Chianese A, Piccialli F (2015) Improving user experience of cultural environment through IOT: the beauty or the truth case study. *Smart Innov Syst Technol* 40:11–20
- Chianese A, Piccialli F, Valente I (2015) Smart environments and cultural heritage: a novel approach to create intelligent cultural spaces. *J Location Based Serv* 9(3):209–234
- Marra A, Gerbino S, Greco A, Fabbrocino G (2021) Combining integrated informative system and historical digital twin for maintenance and preservation of artistic assets. *Sensors* 21(17)
- Lorusso A, Guida D (2022) IOT system for structural monitoring. *Lecture Notes in Networks and Systems*, 472 LNNS, pp 599–606
- Perles A, Perez-Marin E, Mercado R, Segrelles JD, Blanquer I, Zarzo M (2018) An energy-efficient internet of things (IOT) architecture for preventive conservation of cultural heritage. *Future Gene Comput Syst* 81:566–581
- Fayyaz Z, Ebrahimian M, Nawara D, Ibrahim A, Kashef R (2020) Recommendation systems: algorithms, challenges, metrics, and business opportunities. *Appl Sci (Switz)* 10(21):1–20
- Felfernig A, Polat-Erdeniz S, Uran C, Reiterer S, Atas M, Tran TN, Azzoni P, Kiraly C, Dolui K (2019) An overview of recommender systems in the internet of things. *J Intell Inf Syst* 52(2):285–309
- Francesco R, Bracha S, Rokach L (2015) Introduction and challenges. In: *Recommender systems*
- Colace F, Conte D, De Santo M, Lombardi M, Santaniello D, Valentino C (2022) A content-based recommendation approach based on singular value decomposition. *Connection Sci* 34(1):2158–2176
- Lops P, De Gemmis M, Semeraro G (2011) Content-based recommender systems: state of the art and trends. In: *Recommender systems handbook*, pp 73–105
- Koren Y, Bell R (2015) Advances in collaborative filtering
- Al Fararni K, Nafis F, Aghoutane B, Yahyaouy A, Riffi J, Sabri A (2021) Hybrid recommender system for tourism based on big data and AI: a conceptual framework. *Big Data Min Anal* 4(1):47–55
- Clarizia F, Colace F, De Santo M, Lombardi M, Pascale F, Santaniello D (2019) A context-aware chatbot for tourist destinations. pp 348–354
- Clarizia F, Colace F, De Santo M, Lombardi M, Pascale F, Santaniello D, Toker A (2020) A multilevel graph approach for rainfall forecasting: a preliminary study case on London area. *Concurrency Comput Pract Experience* 32(8)
- Abowd GD, Dey AK, Brown PJ, Davies N, Smith M, Steggles P (1999) Towards a better understanding of context and context-awareness. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1707:304–307
- Adomavicius G, Mobasher B, Ricci F, Tuzhilin A (2011) Context-aware recommender systems. *AI Mag* 32(3):67–80
- Shaina R, Chen D (2019) Progress in context-aware recommender systems—an overview. *Comput Sci Rev* 31:84–97
- Lu F, Tian F, Jiang Y, Cao X, Luo W, Li G, Zhang X, Dai G, Wang H (2011) Shadowstory: creative and collaborative digital storytelling inspired by cultural heritage. pp 1919–1928
- Psomadaki OI, Dimoulas CA, Kalliris GM, Paschalidis G (2019) Digital storytelling and audience engagement in cultural heritage management: a collaborative model based on the digital city of Thessaloniki. *J Cult Heritage* 36:12–22
- Chianese A, Piccialli F (2016) A smart system to manage the context evolution in the cultural heritage domain. *Comput Electr Eng* 55:27–38



23. Bartolini I, Moscato V, Pensa RG, Penta A, Picariello A, Sansone C, Sapino ML (2016) Recommending multimedia visiting paths in cultural heritage applications. *Multimedia Tools Appl* 75(7):3813–3842
24. Ruotsalo T, Haav K, Stoyanov A, Roche S, Fani E, Deliai R, Makela E, Kauppinen T, Hyvonen E (2013) Smartmuseum: a mobile recommender system for the web of data. *J Web Seman* 20:50–67
25. Sequeda JF, Tirmizi SH, Corcho O, Miranker DP (2011) Survey of directly mapping SQL databases to the semantic web. *Knowl Eng Rev* 26(4):445–486
26. Colace F, Conte D, Gupta B, Santaniello D, Troiano A, Valentino C (2023) A novel context-aware recommendation approach based on tensor decomposition. *Lecture Notes in Networks and Systems* 448:453–462
27. Mitchell S, OSullivan M, Dunning I (2011) Pulp: a linear programming toolkit for python. The University of Auckland, Auckland, New Zealand, p 65

# Artificial Intelligence Applications in Healthcare



Usman Ahmad Usmani , Ari Happonen , Junzo Watada ,  
and Jayden Khakurel

**Abstract** The demand of healthcare services is rising, for e.g., Europe and the US are experiencing shortage of healthcare professionals. Artificial intelligence (AI) holds a promise to assist healthcare professionals in wide range of tasks. There is already a large amount of clinical and non-clinical evidence that AI algorithms can analyze both structured and unstructured clinical data (including images) data from electronic medical records (EMRs) with the characterization and prognosis of the disease. However, there is lack of study that provides an overview on what are the clinical applications that currently exist. This study provides an overview on the most powerful AI applications in healthcare, including those that are directly related to healthcare as well as those that are part of the healthcare value chain, such as drug development and ambient assisted living. Moreover, this article also provides an overview on the ethical concerns that may arise with the use of AI in healthcare domain.

**Keywords** Artificial intelligence · Healthcare · Ethic · Employee shortage · Pandemic · Industry 4.0 · Machine learning · Digitalization · Healthcare sustainability

---

U. A. Usmani

Universiti Teknologi Petronas, 32610 Seri Iskandar, Perak, Malaysia

e-mail: [usman\\_19001067@utp.edu.my](mailto:usman_19001067@utp.edu.my)

A. Happonen (✉)

LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland

e-mail: [Ari.Happonen@lut.fi](mailto:Ari.Happonen@lut.fi)

J. Watada

Waseda University, Tiergartenstr. 17, Heidelberg 69121, Japan

J. Khakurel

University of Turku, 20500 Turku, Finland

e-mail: [Jayden.khakurel@utu.fi](mailto:Jayden.khakurel@utu.fi)

## 1 AI in Healthcare Industry

In the healthcare industry, AI refers to machine learning algorithms and software that mimic human cognition in the processing, presentation, and interpretation of complex medical and healthcare data. The primary objective of health-related AI research is to determine how clinical parameters influence patient outcomes [1]. Diagnostics, pharmacological research, medicine, and patient monitoring and care are just a few of the fields where AI is used. The ability of AI technology and utilization of ML for data analysis and to offer a well-defined output to the end-user sets it apart from traditional healthcare solutions. ML models need to be trained on a large amount of data. Specially in healthcare and in pandemic like situation, speed is a key to solve issues [2] and efficient means are needed to detect new patterns and react on them.

In pattern recognition, AI algorithms are literal: the algorithm can only do from the data it gets and understand what it is instructed to do. These algorithms are basically black boxes, offering little to no explanation to their decisions [2]. In short, AI is still in somewhat on its infancy in healthcare utilization. Still Bayesian networks, artificial neural networks (NN), and fuzzy set theory have all been applied in healthcare [3]. Also, there is an increase in the number of genetic sequencing databases [4]. Electronic health record (EHR) systems are becoming more common [5]. Hospitals are also turning to AI to save money, improve patient satisfaction, and satisfy staffing and demand demands [6]. The US government [7] is currently spending billions of dollars on AI in healthcare. The contributions of the paper are as follows:

- We give a brief overview of the clinical applications of the role of AI in healthcare domains, such as radiology, oncology, pathology, etc.
- The system applications of AI, such as disease diagnosis, telemedicine, drug intervention, etc. are described.

## 2 Clinical Applications

AI and ML are gaining popularity illness detection and treatment in digital healthcare context. By using massive datasets, illness diagnosis e.g., for cancers is progressing utilizing AI and ML advancements. Here, deep learning (DL) is as a subset of ML, a tool that is mimicking the human brain's capacity to analyze data in order to recognize photos, objects, improve diagnostics, and aid people in making choices. It may be capable of operating without a need for human intervention [8]. Clinical oncology research is increasingly focusing on researching the complicated biological architecture of cancer cell proliferation in order to comprehend the genetic causes of cancer. It also aimed to use big data and computational biology to examine millions of cases in order to solve the existing global picture of growing cancer mortality [9]. Furthermore, by combining NGS sequencing with high-resolution imaging technologies, AI in clinical decision-making is expected to improve the chances of early illness prediction and diagnosis. Following sections will discuss the clinical applications of AI.

## 2.1 *Cardiovascular Medicine*

Cardiac imaging analysis has a lot of space to grow. DL help in coronary and electrocardiograms (ECG). Previously, heart diseases such as CHD and acute coronary syndrome have mostly been treated by the cardiac intervention (ACS). DL and AI are able to detect coronary atherosclerotic plaques more accurately than doctors in the near future. Two examples of how AI can be used to detect images are automated chamber size appraisal and left ventricular function assessment [10]. It can also be used to look for structural anomalies such as valve disease and help and stage disorder. Jimenez et al. [11] observed that DL could predict survival with better accuracy after training the data of several patients.

Cardiac single-photon emission computed tomography and optical coherence tomography are used to identify the lumen and border, and cardiac single-photon emission computed tomography (to the three layers of the coronary artery) (to diagnose myocardial ischemia) [12] (to diagnose myocardial ischemia [13], improve the myocardial perfusion imaging diagnostic accuracy, and MRI (to diagnose myocardial ischemia [14] and improve the diagnostic accuracy of myocardial perfusion imaging) [15] (for the efficient and fast of the cardiac segmentation in short-axis MRI) [16].

AI combined with minimally invasive surgery technology like the Da Vinci Surgical Robot can help to enhance surgical safety, shortening hospital stays and lowering patient trauma [17]. It's also possible, that AI could be able to learn to do surgery quicker than humans.

## 2.2 *Dermatology*

AI is being investigated to enhance and/or complement existing melanoma and skin cancer screening approaches (NMSC). With sensitivity and specificity of 0.81 and 0.80, Hogarty et al. [18] were the first to train a NN for melanoma detection. Young et al. [19] investigated whether deep learning technology might be utilized to develop a viable skin cancer classificatory, based on few clinical photos. The findings were evaluated by 13 board-certified dermatologists and nine dermatology trainees [20]. Although board-certified dermatologists' malignant and benign categorization accuracy was statistically higher than that of dermatology trainees, the DCNN's accuracy was significantly higher, at around 21% [21]. Han et al. [22] studied the application of a DL system on clinical photographs of several skin conditions, inc. melanoma. The performance of the assessed algorithm was found to be equivalent to that of 16 dermatologists [23]. The program's external validity, on the other hand, is limited, as Han et al. [22] discovered while testing it on a different patient population that it had a much lower sensitivity. Only 29 out of 100 lesions were correctly diagnosed [24].

Dermatologists found an overall sensitivity of 89.4% and a specificity of 64.4% when identifying melanoma using the authors' criteria [25]. Although AI for skin

cancer detection using clinical images and photos is still in its early stages, it shows promise [19]. Dermatitis of the skin, often known as skin dermatitis, is a kind of dermatitis that affects the skin. The importance of natural language processing and ML in EHRs is shown by these results [26]. To improve the accuracy of the present research, contextual data might be needed to be made accessible for the AI systems.

Important system performance factors including feature retention power, aggregated feature effect, dependability index, and standard metrics like accuracy, sensitivity, and specificity are measured as a consequence. The best system was a blend of SVM and FDR, which employed a database of 670 psoriasis photos and achieved a cross-validation accuracy of 99.84%. With cross-validation technique, the SVM-FDR system has a 99.99% dependability. They put the procedure to the test by comparing the findings of automatically segmented lesions to those of manually segmented lesions, and discovered that the results were almost identical. They were able to achieve sensitivity and specificity ranges of 82.7–96.7% and 69.3–96.7%, respectively, during validation. The AUROC value varied between 0.82 and 0.98 [27].

## 2.3 Gastroenterology

Newest ML models have shown promise in diagnosing diseases and predicting prognoses and overcome the limitations of classical linear statistics. New ANNs are well-suited for dealing with complications. Furthermore, the ANN has the potential to discover complex correlations between demographic, environmental, and clinical factors. In 2005, Pace et al. [28] created an ANN model that successfully identified illness in 159 patients using 45 clinical parameters. In a similar pilot investigation, Cao et al. [29] used linear discriminant analysis to detect gastritis in 350 outpatients using just clinical and biochemical data.

The APACHE II and the Ranson criteria grading system have both been shown to be reliable [30]. Some authors employed an ANN to predict the prognosis of 190 people who suffered from significant lower gastrointestinal in 2003. The authors compared the performance to a previously validated scoring system (BLEED).

Mungo et al. [31] used an ANN model to predict five-year and one-year survival in 418 esophageal cancer patients in 2005. The ANN model outperformed traditional linear discriminant analysis approach in terms of accuracy. In the input, the quantity of training data bits has recently risen from hundreds to thousands. Mungo et al. [32] utilized a supervised ANN model to predict death in 2380 people who had an upper gastrointestinal hemorrhage. That approach outperformed the full score in terms of specificity (97.5 versus 52.0%), accuracy (96.8 versus 52.9%), area under the receiver operating characteristic (AUROC), the sensitivity (83.8% versus 71.4%), and of prediction performance (0.95 versus 0.67). Yang et al. [33] developed an ANN model for predicting prognosis in colitis patients after treatment, with a sensitivity and specificity of 96 and 97% for the necessity of surgery, respectively. In reference [34], an ANN model was developed that successfully predicted healing in people

with inflammatory bowel disease (IBD). Qiu et al. [35] employed an artificial neural network (ANN) model to predict the development, severity, and relapse of IBD.

When it came to misdiagnosing lymph node metastases, an SVM-based prediction model cut the number of unnecessary treatments in half (compared to a prediction model based on European (91%), Japanese recommendations (91%), American (85%)); Yang et al. [36] gathered data and 23 immunologic markers from 483 patients who had curative surgery for esophageal squamous cell carcinoma to construct an SVM-based model. As a result, they are interested in AI applications, particularly those that use images. Non-neoplastic gastrointestinal illnesses like infection, inflammation, and bleeding are also treated with AI.

## 2.4 Infectious Diseases

Authorities have established systems to identify those at risk of infectious disease transmission due to concerns about disease transmission. As a result, temperature checks are carried out regularly across Singapore airport terminals, with a thermal camera used to detect overheating people. This simple method is only one of several that can prevent illnesses from spreading. New mathematical modeling-based techniques are being applied to improve this kind of monitoring. Vital sign categorization was used to establish a comparable system for detecting unwell persons [37].

The precision of biological samples in real-life, such as patient's blood samples, highlighted the limitations of building tools based on data collected in a controlled environment (laboratory). It is unclear if this was caused by bacterial interactions that normally occur or by poor biological sample quality.

Practical considerations like as sample quality and laboratory procedure time must, however, be included into the mathematical models offered. It was discussed how to diagnose TB, the world's second-leading cause of infection-related mortality [38]. The following are some of the drawbacks of employing ML: it detects data patterns without requiring a perfect match to known patterns, which results in poor accuracy and (ii) noise and loss may occur if  $k$  is too small or big.

SVM was employed to categorize TB cohort, as it was considered as robust classifier. ML algorithms were developed to identify malaria-infected red blood cells using data from digital in-line holographic microscopy, a low-cost approach [39]. Eleven parameters on individual RBC segmented holograms differed significantly in healthy and sick people. The SVM-trained model achieved the highest accuracy in both training ( $n = 280$ , 96.78%) and testing ( $n = 120$ , 97.50%). This DIHM-based AI system is easy to use and does not need blood samples to be processed. Epidemiological studies may be conducted at the population level or at the bedside of patients (clinical epidemiology).

Epidemiological studies give a diverse variety of input data, enabling a broad range of AI assets to be used. Epidemiological studies and forest sickness have suggested that based on a very low incidence, we may be able to predict an epidemic, as demonstrated by a recent research on a tick-borne viral infectious disease. The

researchers exhibited a high rate of prediction and suggested that future databases incorporate localization data to improve transmission control using a modified NN.

Aside from ARIMA, there are various methods for detecting a potential danger associated with a disease pandemic. The major characteristics linked with infectious risk in the spread of Rift Valley fever (RVF) throughout Africa and the Arabian Peninsula were discovered using maximum entropy ML methods.

Because of the variety and quantity of available malaria control data, such an approach was possible (e.g., Malaria Immunology Database, Mapping Malaria Risk in Africa, PlasmoDB, and Malaria Atlas Project). The accuracy of the prediction model is determined by the quality and specificity of the input data rather than its quantity. Using SVM and the radial basis function (RBF) kernel, scientists were able to predict and avoid high morbidity rates. The most dependable indicator has little to do with the weather and everything to do with the dengue-carrying mosquitos' infection rate. In most cases of infectious disease, the efficiency of transmission halting is closely linked to outreach. Based on age, gender, and other socioeconomic factors, the best technique for connecting with and reaching out to varied groups should be determined. AI should be utilized to predict infectious disease pandemics and treatment results. The high-level AI analytical approach discussed above is possible when each of the various databases has a high degree of truth.

## 2.5 *Oncology*

Due to worries about disease transmission, authorities have devised methods to identify persons at risk of infectious disease transmission. As a consequence, frequent temperature inspections are undertaken across Singapore's airport terminals, with a thermal camera used to detect those with dangerously high temperatures. This straightforward strategy is only one of several that may be used to prevent infections from spreading. To enhance this form of monitoring, new mathematical modeling-based techniques are being applied. As a comparable strategy for recognizing unwell people, vital sign classification [40] was developed.

As a consequence, unlike other techniques of clustering, a single point may be allocated to many groups. This demonstrates the ability to develop effective strategies for identifying vulnerable populations. Even in the case of severe infectious diseases, triage is required. In the detection of isolated bacteria, the combination of HRM and SVM was 100% accurate [41].

Real-world biological samples, such as patient blood samples, were used to demonstrate the limits of developing tools based on data acquired in a controlled setting (laboratory). It's uncertain if spontaneous bacterial interactions or poor biological sample quality caused this. However, practical concerns like sample quality and laboratory process time must be included into the mathematical models presented. It was debated how to diagnose tuberculosis (TB), the world's second-leading cause of infection-related death [42]. Early indicators of infection were sought since obtaining a formal diagnosis may take a long time. Artificial immune

recognition systems (AIRS) and other modern technologies are capable of detecting a wide range of diseases. The immune system's ability to create AIRS was taken advantage of.

There are a few disadvantages of using ML: it finds data patterns without needing a perfect match to known patterns, resulting in low accuracy; also, if  $k$  is too small or too large, noise and loss issues may develop. Using supervised ML methods, the AIRS achieved great accuracy. By replacing the classifier using SVM instead of, Osamor et al. [43] improved the accuracy of a classification tool. Because SVM is a far more robust classifier, it was used to classify a TB cohort. Using data from digital in-line holographic microscopy, a low-cost method, ML algorithms were constructed to detect malaria-infected red blood cells [44].

Eleven metrics on individual RBC segmented holograms changed considerably between healthy and unwell individuals. In both testing ( $n = 120$ , 97.50%) and training ( $n = 280$ , 96.78%), the SVM-trained model had the greatest accuracy. This DIHM-based AI technology is simple to use and does not need the analysis of complicated blood samples. Over the course of five years, a recent research collected data on a variety of diabetes, cardiovascular disease, hypertension, and other disorders from 50 American states [45]. A total of 30 states' data was utilized for training, while the remaining 20 states' data was used for testing. The results were quite precise due to the vast quantity of data and ML used.

NCD, on the other hand, is a disease that is difficult to spread quickly from one patient to another due to close proximity or a shared environment. Predicting the epidemic's future magnitude and location are crucial due to the virus's severe symptoms. Forest sickness, a tick-borne viral infectious disease, has shown that we may be able to predict an epidemic based on a very low incidence, according to epidemiological studies. The researchers demonstrated a high rate of prediction and suggested that future databases use localization data to better transmission control using a modified NN. In response to current life-threatening infections such as Ebola, researchers have developed new predicting technologies. Researchers used logistic regression (LR), artificial neural networks (ANN), ML, SVM classifiers [46], and decision trees to create an ensemble of predictors (DT).

## 2.6 Pathology

A deep NN is used to process images. Kernels, a class of filters, are used to produce a pooling layer that successfully reduces the dimensions of visual input while preserving its characteristics. Kernels flatten an image and eliminate or decrease its dimensions, allowing computer vision and machine vision models to assess, appraise, and classify digital images, or sections of images, into preset categories. As slide scanning technology advances in speed and reliability, WSI data becomes accessible for training and testing NN models. Computational pathology will become the



future standard of treatment if combined with clinical data and data [47]. Pathologists can better address the course of severe diseases and offer better patient care using computational pathology, which speeds up the pathology process and delivers a more detailed picture [47].

## 2.7 *Primary Care*

The world's healthcare systems are under strain from a number of sociopolitical issues. It is now more important than ever to provide healthcare with the fewest resources feasible while maintaining patient safety. Two major factors exerting pressure on healthcare systems are an ageing population with numerous chronic ailments, as well as growing global healthcare costs. Primary healthcare (PHC) helps to achieve these goals to some extent at the population and community levels. Primary healthcare is fast developing, not just in terms of regulation, but also in terms of technology. The majority of primary healthcare practitioners are now digital and treat their patients using health information systems. AI ideas such as ML and DL may now be used to exploit these health information systems, thanks to advancements in computer and informatics technology. Fourth, AI isn't a new idea; it's been around for about 50 years, with NNs becoming prominent in the 1980s and 1990s. However, because of limits in hardware processing capabilities at the time, this trend did not persist long.

Clinicians' combined experience can avoid these biases, assisting in the formulation of suitable treatment recommendations. Another ethical worry is that practitioners' use of AI might alter the nature of the patient-clinician contact. The studies mentioned above, along with a few others, demonstrate how PHC can use AI in the future. Only a few AI algorithms have been employed in everyday clinical practice, despite all of AI's methodological and computational accomplishments. The difficulties and problems associated with AI's deployment in PHC, we feel, are one of the main causes. Furthermore, a broad collection of stakeholders, including doctors, informaticians, AI researchers, and AI practitioners, all have opposing viewpoints on AI's use in PHC [48]. The International Medical Informatics Association (IMIA Primary)'s HealthCare Informatics Working Group conducted this Delphi poll in an attempt to reach a consensus on AI attitudes, issues, and concerns in primary care.

## 2.8 *Psychiatry*

Psychiatric diseases impact children as early as five-years old, making them one of the top causes of disability worldwide [49]. They impose a significant financial burden on those afflicted, both in terms of lost years of life owing to disability or death and in terms of societal healthcare expenses [50]. Despite its youth, AI has already changed the way physicians diagnose, forecast, and treat mental diseases [51]. Despite recent

breakthroughs in our knowledge of what causes mental illnesses, clinicians have struggled to diagnose them due to a lack of objective and exact clinical testing. One approach is to use technology to help in mental disorder diagnosis. Latent semantic analysis (LSA), a high-dimensional automated technique for evaluating speech transcripts, has been effectively used by doctors to help in the diagnosis of mental diseases.

For concept-based text analysis, LSA is a natural computational language processing approach. It creates a semantic knowledge representation from a set of natural-language keywords in order to find connections between words and their meanings [52]. The primary premise behind the development of a meaning model is that words used in similar situations are more semantically connected than those used in different contexts. Sarzynska-Wawer et al. [53] studied speech transcripts from schizophrenic patients and healthy controls. It was able to distinguish if a message came from the sick or the control group. LSA has also been used with structural speech analysis to uncover minute disparities between schizophrenia patients' first-degree relatives and unrelated healthy persons.

Using ML and electronic health data, Walsh et al. [54] successfully predicted future suicide attempts in a sample of adult patients with a history of self-injury. For example, the moderated online social therapy (MOST) project is an Internet-based solution that employs an online social therapy system to assist persons with mental health problems [55]. MOST [56] is the first program of its kind in the world, integrating online peer assistance with social networking in a clinician-moderated setting. The MOST model has been demonstrated to be useful in trials involving young individuals with psychosis or depression [57], but further research is required to confirm its long-term efficacy.

## 2.9 Radiology

AI is being used in fields such as remote patient monitoring, medical diagnostics, drug discovery, imaging, hospital administration, virtual assistants, and risk management. AI is anticipated to help a variety of enterprises that rely on large amounts of data, such as DNA and RNA sequencing data analysis. Radiology, pathology, dermatology, and ophthalmology, all of which rely on image data, have already realized the benefits of AI deployment. Radiologists are doctors who evaluate medical pictures and provide information to help diagnose, categorize, and monitor illnesses.

Because imaging data is collected as part of routine clinical practice, large datasets are theoretically straightforward to obtain, offering a vast resource for scientific and medical discovery. The fast growth and diffusion of medical imaging as a research discipline has been fueled by the integration of pictures and clinical outcome data [A64]. Image mining was originally used to seek for form, intensity, and texture, all of which have an influence on features. Many of these qualities are therapeutically relevant, according to a recent study that used DL algorithms to generate feature representations from sample photos.

Finally, we discuss the challenges and limitations of putting these concepts into reality. The goal of using AI to medical imaging is to improve the efficacy and efficiency of clinical care. In contrast to the number of professional readers available, imaging data is rising at an unsustainable rate.

Although discriminatory, such characteristics are chosen by experts, and hence may not be the optimum feature quantification approach for the discriminating task at hand. In addition, established features are often incapable of responding to changes in imaging modalities such as PET, MRI, and CT, as well as their signal-to-noise quality. Radiologists employ their manual perceptual skills to detect abnormalities, then use their cognitive abilities to question the results of the manual detection approach. Radiologists visually assess image stacks and make necessary adjustments to the viewing planes, window width, and level settings. Radiologists are taught to detect abnormalities based on changes in imaging intensity or the appearance of aberrant patterns using their education, experience, and understanding of the healthy. These and other considerations are part of a somewhat subjective decision-making matrix that may be used to a range of conditions, including lung nodules, breast polyps, and colon polyps.

Several mammography investigations found that radiologists seldom altered their diagnoses after using clinical integration and feature-based systems, and that these technologies had no statistical influence on radiologists' performance. Some of this is due to the systems' human-like capacities. DL has been used to detect lung nodules in CT and prostate cancer in imaging, such as MRI<sup>44</sup>, in recent research.

### 3 Systems Applications

Throughout the information age, the introduction of new technology has had an impact on a variety of enterprises. In the healthcare industry, there are no exceptions. Doctors, hospitals, insurance companies, and other healthcare-related enterprises have all been affected by automation, ML, and AI—often more favorably and profoundly than other industries (AI). In fact, AI is employed by 86% of healthcare provider companies, life science enterprises, and healthcare technology vendors, according to a 2016 CB Insights survey. By 2020, these companies will have invested an average of \$54 million on AI. So, what are some of the most well-liked alternatives?

#### 3.1 Disease Diagnosis

According to Jiang et al. [58], a number of AI techniques, such as support vector machines, NNs, and decision trees, have been used to treat a variety of disorders. Each method has its own set of advantages and disadvantages. "In order to provide information for illness diagnosis and classification, there are two ways for classifying various diseases." The two kinds of NNs are artificial neural networks (ANN) and

Bayesian networks (BN). “In order to provide information for illness diagnosis and classification, there are two ways for classifying various diseases.” The ability of ANN to properly diagnose diabetes and cardiovascular illness has been shown.

Using massive amounts of electronic health records (EHRs) and medical learning classifiers, AI has been able to considerably aid physicians in patient diagnosis (MLCs) [59]. When medical illnesses get more complicated, case duplication becomes more common, particularly for patients who have a lengthy history of providing electronic medical data [59]. Despite the fact that people with uncommon diseases are less likely to be alone, doctors have a hard time discovering individuals who have similar clinical conditions [59]. Not only is AI being used to uncover similar cases and treatments, but it is also being used to factor in critical symptoms and aid physicians in asking the proper questions, enabling patients to get the most accurate diagnosis and therapy possible [59].

### ***3.2 Telemedicine***

The expansion of remote patient care has increased the availability of AI applications [60]. AI may be able to aid with remote medical care by monitoring data from sensors [61]. A wearable device might allow for continuous patient monitoring and identification of changes that would otherwise go unnoticed. The data might be compared to previously collected data, with AI algorithms alerting doctors if any discrepancies are found [61]. Another study discovered that mental health treatment lacked the reciprocity and responsibility for care that should be present in interactions between mental health consumers and practitioners (whether or not they are psychologists) [62]. AI can efficiently aid in the care of the elderly, since the average age has grown owing to increasing life expectancy [63]. For example, sensors in the surroundings and on people’s bodies might monitor a person’s daily routine and warn a caregiver if one of them, or a recorded vital, is abnormal [63]. Despite the benefits of the technology, there are worries about monitoring limits in order to safeguard a person’s privacy, especially since technologies exist to map out home layouts and identify human actions [63].

### ***3.3 Electronic Health Records***

Electronic health records (EHRs) are essential for data interchange in the healthcare business. When AI is used in more than 80% of medical contexts, it will be able to review data and provide fresh insights to doctors [64]. In one scenario, natural language processing (NLP) is being used to make shorter reports by matching comparable medical words [64]. Heart attack and myocardial infarction, for example, are equivalent terms, although physicians may favor one over the other. NLP systems

correct for these inconsistencies, allowing for more thorough analysis. Finding duplicate language in a physician's notes owing to repetition and preserving the key information to make the notes easier to read is another use of NLP [64]. Other apps employ concept processing to look for similar circumstances in the data supplied by the current patient's doctor and to help the doctor remember to include all relevant information [65]. In addition to making content changes to an EHR, AI algorithms may study an individual patient's record and anticipate sickness risk based on past data and family history [66]. A rule-based system is a kind of algorithm that works similarly to how individuals use flow charts to make decisions [67]. This system gathers a significant amount of data and develops a set of rules that connect specific observations to final diagnoses. As a consequence, depending on fresh patient data, the algorithm may be able to predict whether or not a patient would acquire a certain sickness or condition [67]. Because the algorithms can combine a patient's data, they may be able to discover any unsolved issues and bring them to the attention of a physician, thus saving time.

According to a study conducted by the research, using EHR data to predict therapeutic response has a 70–72% accuracy rate [66]. Since the amount of online health records grows every five years, these choices are intriguing [68]. Because physicians may not have the time or resources to manually enter all of this information, AI may be able to assist them in providing better care to their patients [66].

### **3.4 Drug Interactions**

Algorithms for detecting drug-drug interactions in the medical literature have been developed, thanks to advances in natural language processing [69–72]. Drug-drug interactions are a worry for patients who are taking many medications at the same time, and they're growing more common as the number of medications increases [73]. To cope with the challenge of monitoring all known or suspected drug-drug interactions, ML algorithms have been created to extract medical literature on interacting medications and their possible consequences. In 2013, a group of researchers from Carlos III University set out to gather data on drug-drug interactions in order to establish a test for such algorithms [74]. The contestants were judged on their ability to correctly identify which medications interacted with one another in the text, as well as the features of such interactions [75]. Researchers continue to utilize this corpus to assess algorithm efficiency. Other algorithms scour user-generated data, such as electronic health records and/or adverse event reports, for drug-drug interactions [70, 71]. Clinicians may report suspected adverse pharmacological responses to the World Health Organization's VigiBase and the FDA's Adverse Event Reporting System (FAERS). DL algorithms were developed in order to look for patterns in data that might reveal drug-drug interactions.

### ***3.5 Creation of New Drugs***

DSP-1181 is a drug developed by a British start-up to treat obsessive-compulsive disorder (OCD) (Japanese pharmaceutical firm). Drug development might be completed in a year if pharmaceutical companies devote five years on such attempts. DSP-1181 has been tested on humans. Medicine started using AI in September 2019 to uncover six possible inhibitors of the DDR1 gene, a kinase target connected to fibrosis and other illnesses. A prospective lead contender was evaluated in mice after the new medications were produced in only 21 days using a process known as generative reinforcement learning (GENTRL). Deep Genomics, a Canadian firm, revealed in the same month that their AI-powered drug discovery platform has discovered a target and therapy option for Wilson's disease. Exon-skipping impact of Met645Arg, a genetic mutation that renders the copper-binding protein ATP7B inefficient, might be addressed with DG12P1.

### ***3.6 Industry***

When huge health businesses band together, health data becomes more accessible. More health data allows AI systems to develop. In the healthcare industry, clinical decision support systems account for a significant proportion of AI research. ML algorithms improve as more data is gathered, allowing for more reliable replies and solutions [76]. Big data is being investigated by several firms in the healthcare sector. Many firms are competing in the market for data analysis, storage, administration, and analytical technologies, which are all important aspects of the healthcare sector. The following firms have made significant contributions to the development of AI algorithms for healthcare: IBM is working on Watson Oncology at Cleveland Clinic and Memorial Sloan Kettering Cancer Center. IBM is also collaborating with CVS Health on AI-based chronic illness management systems, as well as Johnson & Johnson on scientific document analysis to identify potential drug development partnerships.

IBM and Rensselaer Polytechnic Institute formed the Health Empowerment by Analytics, Learning, and Semantics (HEALS) joint venture in May 2017 to investigate how AI may improve healthcare. SoftBank Robotics and UBTECH, for example, are working on robots that are similar. A WhatsApp group set up by an Indian firm answers inquiry regarding dangerous conditions in India. As the AI sector evolves, Amazon, Google, and Apple, for example, all have their own AI research teams and millions of dollars set aside to buy smaller AI-based businesses. A number of automakers are using AI into their healthcare systems. Toyota, Tesla, GE, BMW, and Volvo have all launched new research programs to assess a driver's essential data to guarantee that they are alert, paying attention to the road, and not compromised by drugs or suffering from mental distress.

### 3.7 Regulation

While healthcare AI research is seeking to establish its usefulness in improving patient outcomes before it is widely implemented, its usage might expose patients and healthcare staff to new sorts of hazards, such as do not resuscitate implications, algorithmic prejudice, and other machine morality problems. Because of the concerns with AI's therapeutic usage, restrictions may be essential.

The group's stated goals for publicly sponsored AI research and development were verified in the National Artificial Intelligence Research and Development Strategic Plan, which was released in October 2016 (both in government and academics). The only agency that has voiced unhappiness with the situation is the Food and Drug Administration (FDA). The International Telecommunication Union and the World Health Organization formed the Joint Focus Group on AI for Health to test and assess AI technology in the health sector (FG-AI4H) [76]. As of November 2018, there were eight use cases, including using imaging to predict breast cancer risk, delivering anti-venom drugs based on snake photos, and diagnosing skin problems.

## 4 Discussion

In medicine, algorithms have already shown a number of potential benefits for both professionals and patients. Regulating these algorithms, on the other hand, is a challenging undertaking. The Food and Drug Administration (FDA) has approved a variety of assistive algorithms in the United States, but there are no universal approval processes in place.

Furthermore, clinical algorithm creators aren't often the same individuals who treat patients; as a consequence, computationalists may need to learn more about medicine, while clinicians may need to learn more about the tasks that a certain algorithm is or isn't well suited to. While AI may aid in diagnostics and basic clinical tasks, it's tough to envision automated brain surgery, where doctors must often adjust their plans depending on what they observe within the patient. In these and other ways, AI's promise in medicine currently outnumbers its ability to provide patient care. Clarified FDA standards, on the other hand, may make creating algorithm criteria simpler, perhaps leading to more algorithm usage in clinical practice.

In the modern digitalizing world [77–82], AI technologies have become more common in the business and everyday life and are now making inroads into healthcare. AI has the potential to help healthcare personnel in a variety of areas, including patient care and administrative operations, by enabling them to improve on existing solutions and solve problems more quickly. While most AI and healthcare technologies are vital to the healthcare business, the strategies that they allow may vary significantly across hospitals and other healthcare organizations. Furthermore, although certain research on AI in healthcare suggest that AI may perform as well as or better than humans in some tasks, such as illness detection, it will take many years before

AI totally replaces people in a variety of medical professions. However, for many, the solution remains a mystery. What is AI in healthcare, and how might it benefit you? What is the present status of AI in healthcare, and how will it evolve in the future? Will it ever be able to take over crucial operations and medical care? Machines should be taught a global code of ethics as well as a variety of local ethical norms. AI's ethical roots are not the same as humans'. Human ethical principles aren't etched in stone; they've been a cause of debate and cultural variance for a long time. As a consequence, in order to standardize medical practice ethics, determining which ethics to educate robots must come first.

Approaches based on top-down and bottom-up virtue ethics must all be investigated when assessing what works best for medical AI in general and in specific areas. A total of 96 machine judgments are given a theoretical educational framework in top-down ethics. Machines obtain ethics via growth or experience in a bottom-up way. Hybrid approaches are used to compensate for the shortcomings of both strategies. A computer-based model of moral psychology is known as moral calculation. If/then rules and yes/no cascades are used to feed binary ethical dialogues and problems into an algorithm. Algorithmic equations are used to quantify ethical issues and feed them into computers. Implicit moral computers are given ethical rules. Explicit moral machines are moral robots that can compute their own ethical rules [83]. The autonomy and moral consciousness of ethical robots are assessed. Machine ethics, on the other hand, are much too complex to be included into algorithms; nonetheless, including these ethics into algorithms may enable algorithms to exceed human judgment and decision-making [84].

## 5 Conclusion

Clinical decision support systems (CDSS) allow doctors to quickly get relevant data or research to make medical, therapeutic, and patient-related decisions, as well as mental health decisions. AI is being utilized in medical imaging to identify lesions or other issues that a human radiologist could miss in MRIs, x-rays, CT scans, and other pictures. Many healthcare institutions across the world are putting AI-enabled technologies to the test, such as AI-powered COVID-19 patient screening systems and algorithms for patient assistance and monitoring. The findings of these investigations are being collated now, and broad recommendations for the use of AI in medicine are being produced. In any case, AI's potential to help doctors, researchers, and the people they serve is fast expanding. In the creation and management of digital health systems, AI will almost certainly play an important role.

A number of challenges must be addressed, including informed use authorization, algorithmic fairness, security and transparency, data privacy, and biases. All stakeholders, including healthcare professionals, AI developers, regulatory bodies, and patients, must work together to overcome the challenges listed above in order to ensure that AI is ethical and legal. To achieve the ultimate societal goal of AI benefitting everyone, we must develop a system founded on public trust. Important



factors include data security, informed consent, algorithmic fairness and resilience, privacy levels, regulatory supervision, and appropriate openness with greater safety. If an AI-driven healthcare system based on the slogan health AIs for all of us is to be established efficiently, effectiveness standards and an acceptable liability regime for AIs are critical factors that must be investigated and addressed; in this context, we must examine current legal frameworks and make changes to reflect new technology breakthroughs. However, social and political discussions about AI-driven healthcare ethics, as well as the implications for society and the human workforce, are crucial. The potential for AI to enhance our healthcare system is great, but we won't be able to do it until the present ethical and legal hurdles are addressed.

## References

1. van de Sande D, Van Genderen ME, Smit JM, Huiskens J et al (2022) Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health Care Inform* 29(1)
2. Happonen A, Tikka M, Usmani U (2021) A systematic review for organizing hackathons and code camps in Covid-19 like times: literature in demand to understand online hackathons and event result continuation. In: 2021 International conference on data and software engineering (ICoDSE), pp 7–12. IEEE. <https://doi.org/10.1109/ICoDSE53690.2021.9648459>
3. Bazoukis G, Hall J, Loscalzo J, Antman EM, Fuster V, Armondas AA (2022) The inclusion of augmented intelligence in medicine: a framework for successful implementation. *Cell Rep Med* 3(1):100485
4. Holzinger A, Längs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. *Wiley Interdisc Rev: Data Min Knowl Discov* 9(4):e1312
5. Kaul V, Enslin S, Gross SA (2020) History of artificial intelligence in medicine. *Gastrointest Endosc* 92(4):807–812
6. Kulkarni S, Seneviratne N, Baig MS, Khan AHA (2020) Artificial intelligence in medicine: where are we now? *Acad Radiol* 27(1):62–70
7. Becker A (2019) Artificial intelligence in medicine: what is it doing for us today? *Health Policy Technol* 8(2):198–205
8. Jotterand F, Bosco C (2022) Artificial intelligence in medicine: a sword of Damocles? *J Med Syst* 46(1):1–5
9. Alrefaei AF, Hawsawi YM, Almaleki D, Alafif T, Alzahrani FA, Bakhrebah MA (2022) Genetic data sharing and artificial intelligence in the era of personalized medicine based on a cross-sectional analysis of the Saudi human genome program. *Sci Rep* 12(1):1–10
10. Usmani UA, Usmani MU (2014) Future market trends and opportunities for wearable sensor technology. *Int J Eng Technol* 6(4):326
11. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P et al (2020) Artificial intelligence in cardiology: present and future. In: Mayo clinic proceedings, vol 95, no 5, pp 1015–1039. Elsevier
12. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R et al (2018) Artificial intelligence in cardiology. *J Am Coll Cardiol* 71(23):2668–2679
13. de Marvao A, Dawes TJ, Howard JP, O'Regan DP (2020) Artificial intelligence and the cardiologist: what you need to know for 2020. *Heart* 106(5):399–400
14. Seetharam K, Shrestha S, Mills JD, Sengupta PP (2019) Artificial intelligence in nuclear cardiology: adding value to prognostication. *Curr Cardiovasc Imaging Rep* 12(5):1–6
15. Siegersma KR, Leiner T, Chew DP, Appelman Y, Hofstra L, Verjans JW (2019) Artificial intelligence in cardiovascular imaging: state of the art and implications for the imaging cardiologist. *Neth Hear J* 27(9):403–413

16. Usmani UA, Happonen A, Watada J (2022) A review of unsupervised machine learning frameworks for anomaly detection in industrial applications. *LNNS*, vol 507, Chapter: 11, pp 158–189. [https://doi.org/10.1007/978-3-031-10464-0\\_11](https://doi.org/10.1007/978-3-031-10464-0_11)
17. Usmani UA, Happonen A, Watada J (2023) Enhancing artificial intelligence control mechanisms: current practices. In: *Real life applications and future views, lecture notes in networks and systems*, vol 559, pp 287–306. [https://doi.org/10.1007/978-3-031-18461-1\\_19](https://doi.org/10.1007/978-3-031-18461-1_19)
18. Hogarty DT, Su JC, Phan K, Attia M, Hossny M et al (2020) Artificial intelligence in dermatology-where we are and the way to the future: a review. *Am J Clin Dermatol* 21(1):41–47
19. Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML (2020) Artificial intelligence in dermatology: a primer. *J Investig Dermatol* 140(8):1504–1512
20. Li CX, Shen CB, Xue K, Shen X, Jing Y, Wang ZY et al (2019) Artificial intelligence in dermatology: past, present, and future. *CMJ* 132(17):2017–2020
21. Cirillo D, Núñez-Carpintero I, Valencia A (2021) Artificial intelligence in cancer research: learning at different levels of data granularity. *Mol Oncol* 15(4):817–829
22. Usmani UA, Happonen A, Watada J (2022) Enhanced deep learning framework for fine-grained segmentation of fashion and apparel. In: *Intelligent computing, SAI 2022, Lecture Notes in Networks and Systems*, vol 507, Chapter: 3, pp 29–44. [https://doi.org/10.1007/978-3-031-10464-0\\_3](https://doi.org/10.1007/978-3-031-10464-0_3)
23. Aractingi S, Pellacani G (2019) Computational neural network in melanocytic lesions diagnosis: artificial intelligence to improve diagnosis in dermatology? *Eur J Dermatol* 29(1):4–7
24. Patel S, Wang JV, Motaparthi K, Lee JB (2021) Artificial intelligence in dermatology for the clinician. *Clin Dermatol* 39(4):667–672
25. Usmani UA, Haron NS, Jaafar J (2021) A natural language processing approach to mine online reviews using topic modelling. In: *International conference on computing science, communication and security*, pp 82–98. Springer, Cham
26. Stiff KM, Franklin MJ, Zhou Y, Madabhushi A, Knackstedt TJ (2022) Artificial intelligence and melanoma: a comprehensive review of clinical, dermoscopic, and histologic applications. *Pigm Cell Melanoma Res*
27. Usmani UA, Watada J, Jaafar J, Aziz IA, Roy A (2021) A reinforcement learning algorithm for automated detection of skin lesions. *Appl Sci* 11(20):9367
28. Nicholson C, Davies JM, George R, Smith B, Pace V et al (2018) What are the main palliative care symptoms and concerns of older people with multimorbidity? A comparative cross-sectional study using routinely collected phase of illness, Australia-modified Karnofsky performance status and integrated Palliative care outcome scale data. *Ann Palliat Med* 7(Suppl 3):S164–S175
29. Cao JS, Lu ZY, Chen MY, Zhang B, Juengpanich S et al (2021) Artificial intelligence in gastroenterology and hepatology: status and challenges. *World J Gastroenterol* 27(16):1664
30. Kröner PT, Engels MM, Glicksberg BS, Johnson KW, Mzaik O et al (2021) Artificial intelligence in gastroenterology: a state-of-the-art review. *World J Gastroenterol* 27(40):6794
31. Mungo B, Molena D, Stem M, Yang SC, Battafarano RJ et al (2015) Does neoadjuvant therapy for esophageal cancer increase postoperative morbidity or mortality? *Dis Esophagus* 28(7):644–651
32. Greenhill AT, Edmunds BR (2020) A primer of artificial intelligence in medicine. *Tech Innov Gastrointest Endosc* 22(2):85–89
33. Yang YJ, Bang CS (2019) Application of artificial intelligence in gastroenterology. *World J Gastroenterol* 25(14):1666
34. Zhu J, Wang L, Chu Y, Hou X, Xing L et al (2015) A new descriptor for computer-aided diagnosis of EUS imaging to distinguish autoimmune pancreatitis from chronic pancreatitis. *Gastrointest Endosc* 82(5):831–836
35. Qiu W, Duan N, Chen X, Ren S, Zhang Y et al (2019) Pancreatic ductal adenocarcinoma: machine learning-based quantitative computed tomography texture analysis for prediction of histopathological grade. *Cancer Manage Res* 11:9253
36. Usmani UA, Jaafar J (2022) Machine learning in healthcare: current trends and the future. In: *International conference on artificial intelligence for smart community: AISC 2020, 17–18 Dec, Universiti Teknologi Petronas, Malaysia*, pp 659–675. Springer Nature Singapore, Singapore

37. Agrebi S, Larbi A (2020) Use of artificial intelligence in infectious diseases. In: *Artificial intelligence in precision health*, pp 415–438. Academic Press
38. Sharma A, Rani S, Gupta D (2020) Artificial intelligence-based classification of chest X-ray images into COVID-19 and other infectious diseases. *Int J Biomed Imaging*
39. Usmani UA, Watada J, Jaafar J, Aziz IA, Roy A (2021) A reinforced active learning algorithm for semantic segmentation in complex imaging. *IEEE Access* 9:168415–168432
40. Usmani UA, Watada J, Jaafar J, Aziz IA (2022) A systematic review of privacy-preserving blockchain in e-medicine. In: *Biomedical and other applications of soft computing*, pp 25–40
41. Shimizu H, Nakayama KI (2020) Artificial intelligence in oncology. *Cancer Sci* 111(5):1452–1460
42. Chua IS, Gaziell-Yablowitz M, Korach ZT, Kehl KL, Levitan NA et al (2021) Artificial intelligence in oncology: path to implementation. *Cancer Med* 10(12):4138–4149
43. Osamor VC, Okezie AF (2021) Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis. *Sci Rep* 11(1):1–11
44. Rudie JD, Rauschecker AM, Bryan RN, Davatzikos C, Mohan S (2019) Emerging applications of artificial intelligence in neuro-oncology. *Radiology* 290(3):607–618
45. Scheetz J, Rothschild P, McGuinness M, Hadoux X, Soyer HP et al (2021) A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Sci Rep* 11(1):1–10
46. Forghani R, Savadjiev P, Chatterjee A, Muthukrishnan N, Reinhold C, Forghani B (2019) Radiomics and artificial intelligence for biomarker and prediction model development in oncology. *Comput Struct Biotechnol J* 17:995
47. Stenzinger A, Alber M, Allgäuer M, Jurmeister P, Bockmayr M et al (2021) Artificial intelligence and pathology: from principles to practice and future applications in histomorphology and molecular profiling. In: *Seminars in cancer biology*. Academic Press
48. Sun TQ, Medaglia R (2019) Mapping the challenges of artificial intelligence in the public sector: evidence from public healthcare. *Gov Inf Q* 36(2):368–383
49. Lin E, Lin CH, Lane HY (2020) Precision psychiatry applications with pharmacogenomics: artificial intelligence and machine learning approaches. *Int J Mol Sci* 21(3):969
50. Doraiswamy PM, Blease C, Bodner K (2020) Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med* 102:101753
51. Ray A, Bhardwaj A, Malik YK, Singh S, Gupta R (2022) Artificial intelligence and psychiatry: an overview. *Asian J Psychiatry*, p 103021
52. Brunn M, Diefenbacher A, Courtet P, Genieys W (2020) The future is knocking: how artificial intelligence will fundamentally change psychiatry. *Acad Psychiatry* 44(4):461–466
53. Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I et al (2021) Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res* 304:114135
54. Jotterand F, Bosco C (2020) Keeping the “Human in the Loop” in the age of artificial intelligence. *Sci Eng Ethics* 26(5):2455–2460
55. Burr C, Morley J, Taddeo M, Floridi L (2020) Digital psychiatry: risks and opportunities for public health and wellbeing. *IEEE-TTS* 1(1):21–33
56. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M et al (2021) The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 20(3):318–335
57. Usmani UA, Watada J, Jaafar J, Aziz IA, Roy A (2021) A reinforcement learning based adaptive ROI generation for video object segmentation. *IEEE Access* 9:161959–161977
58. Beli M, Bobi V, Badža M, Šolaja N, Đuri-Jovii M, Kostić VS (2019) Artificial intelligence for assisting diagnostics and assessment of Parkinson’s disease—a review. *Clin Neurol Neurosurg* 184:105442
59. Usmani UA, Roy, A, Watada J, Jaafar J, Aziz IA (2022) Enhanced reinforcement learning model for extraction of objects in complex imaging. In: *Intelligent computing*, pp 946–964. Springer, Cham

60. Greenwald MF, Danford ID, Shahrawat M, Ostmo S, Brown J et al (2020) Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. *JAAPOS* 24(3):160–162
61. Greenwald MF, Danford ID, Shahrawat M, Ostmo S, Brown J, Kalpathy-Cramer J et al (2020) Evaluation of artificial intelligence-based telemedicine screening for retinopathy of prematurity. *JAAPOS* 24(3):160–162
62. Albert L, Capel I, García-Sáez G, Martín-Redondo P, Hernando ME et al (2020) Managing gestational diabetes mellitus using a smartphone application with artificial intelligence (SineDie) during the COVID-19 pandemic: much more than just telemedicine. *Diabetes Res Clin Pract* 169:108396
63. Elsner P, Bauer A, Diepgen TL, Drexler H, Fartasch M et al (2018) Position paper: telemedicine in occupational dermatology-current status and perspectives. *JDDG: J der Deutschen Dermatologischen Gesellschaft* 16(8):969–974
64. Yuan KC, Tsai LW, Lee KH, Cheng YW, Hsu SC, Lo YS, Chen RJ (2020) The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *Int J Med Inform* 141:104176
65. Lauritsen SM, Kalør ME, Kongsgaard EL, Lauritsen KM, Jørgensen MJ, Lange J, Thieson B (2020) Early detection of sepsis utilizing deep learning on electronic health record event sequences. *Artif Intell Med* 104:101820
66. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE (2018) Patient2vec: a personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6:65333–65346
67. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, Denny JC, Wei WQ (2019) Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep* 9(1):1–10
68. Shi S, He D, Li L, Kumar N, Khan MK, Choo KKR (2020) Applications of blockchain in ensuring the security and privacy of electronic health record systems: a survey. *Comput Secur* 97:101966
69. Basile AO, Yahi A, Tatonetti NP (2019) Artificial intelligence for drug toxicity and safety. *Trends Pharmacol Sci* 40(9):624–635
70. Mak KK, Pichika MR (2019) Artificial intelligence in drug development: present status and future prospects. *Drug Discov Today* 24(3):773–780
71. Zhavoronkov A, Vanhaelen Q, Oprea TI (2020) Will artificial intelligence for drug discovery impact clinical pharmacology? *Clin Pharmacol Ther* 107(4):780–785
72. Bender A, Cortes-Ciriano I (2021) Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov Today* 26(4):1040–1052
73. Hassanzadeh P, Atyabi F, Dinarvand R (2019) The significance of artificial intelligence in drug delivery system design. *Adv Drug Deliv Rev* 151:169–190
74. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers* 25(3):1315–1360
75. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F (2020) Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit Health* 2(12):e667–e676
76. Gruson D, Helleputte T, Rousseau P, Gruson D (2019) Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. *Clin Biochem* 69:1–7
77. Santti U, Happonen A, Auvinen H (2020) Digitalization boosted recycling: gamification as an inspiration for young adults to do enhanced waste sorting. In: *AIP conference proceedings*, vol 2233, Iss 1, pp 1–12. <https://doi.org/10.1063/5.0001547>
78. Happonen A, Santti U, Auvinen H, Räsänen T, Eskelinen T (2020) Digital age business model innovation for sustainability in University Industry Collaboration Model. In: *E3S web of conferences*, vol 211, Article 04005, pp 1–11. <https://doi.org/10.1051/e3sconf/202021104005>
79. Happonen A, Ghoreishi M (2022) A mapping study of the current literature on digitalization and industry 4.0 technologies utilization for sustainability and circular economy in textile

- industries. *Lecture Notes in Networks and Systems*, vol 217, Chapter 63, pp 697–711. [https://doi.org/10.1007/978-981-16-2102-4\\_63](https://doi.org/10.1007/978-981-16-2102-4_63)
80. Ghoreishi M, Happonen A (2022) The case of fabric and textile industry: the emerging role of digitalization, internet-of-things and industry 4.0 for circularity. *Lecture Notes in Networks and Systems*, vol 216, pp 189–200. [https://doi.org/10.1007/978-981-16-1781-2\\_18](https://doi.org/10.1007/978-981-16-1781-2_18)
  81. Vatosios A, Happonen A (2022) Transforming HR and improving talent profiling with qualitative analysis digitalization on candidates for career and team development efforts. In: *Intelligent computing*. *Lecture Notes in Networks and Systems*, vol 283, Chapter 78, Springer, pp 1149–1166. [https://doi.org/10.1007/978-3-030-80119-9\\_78](https://doi.org/10.1007/978-3-030-80119-9_78)
  82. Vaddepalli K, Palacin V, Porras J, Happonen A (2023) Taxonomy of data quality metrics in digital citizen science. *Lecture Notes in Networks and Systems*, vol 578, pp 391–410. [https://doi.org/10.1007/978-981-19-7660-5\\_34](https://doi.org/10.1007/978-981-19-7660-5_34)
  83. Zhang J, Zhang ZM (2023) Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decis Making* 23(1):1–15
  84. Vayena E, Blasimme A, Cohen IG (2018) Machine learning in medicine: addressing ethical challenges. *PLoS Med* 15(11):e1002689

# Maintaining Performance with Less Data: Understanding Useful Data



Dominic Sanderson  and Tatiana Kalganova

**Abstract** As deep learning tasks become more popular, their computational complexity increases, leading to more intricate algorithms and models which have longer runtimes and require more input data. The result is a greater cost on time, hardware, and environmental resources. We propose a novel method for training a neural network for image classification to reduce input data dynamically, in order to reduce the costs of training a neural network model. By using data reduction techniques, we reduce the amount of work performed and therefore the environmental impact of AI techniques, and with dynamic data reduction, we show that accuracy may be maintained while reducing runtime by up to 50% and reducing carbon emission proportionally and gives a way forward for the better implementation of useful data.

**Keywords** Data reduction · Data augmentation · Data step · Data increment · Data cut · Useful data

## 1 Introduction

When creating a deep learning solution, there are two main factors that determine its success: first, the model used, and second, the data used to train with. The model, often a neural network, may be programmed, optimized, and experimented with to reach an acceptable performance, but the data used to train it is argued to be the determining factor [1]. If inappropriate data is used to train and test a model, the model will perform poorly when deployed in real-world situations and can be regarded as a bottleneck to the system's performance [2]. As such, consideration must be given to ensure not only an appropriate quantity of data is available but also data of suitable quality [2]. It has been shown that increasing the data available to a model improves average accuracy [3]. Because of this, when creating new models for different tasks, researchers and data scientists alike will give as much data as possible to train the

---

D. Sanderson (✉) · T. Kalganova  
Brunel University London, London, UK  
e-mail: [dominic.sanderson@brunel.ac.uk](mailto:dominic.sanderson@brunel.ac.uk)

model, wherever possible, without considering if it is truly necessary. This direction is moving rapidly to a state that is environmentally unsustainable and technically unachievable for those outside of big tech companies [4].

In this paper, we analyze three novel methods to dynamically allocate the data used to train a neural network model, for image classification tasks.

The article is organized as follows: In Sect. 2, we review related methods to reduce and increase the data used for training, as well as existing methods to use dynamically changing training data. In Sect. 3, we describe the methods and environments used for our experimentation; in Sect. 4, we describe the experimental techniques used in our investigation of dynamic data use. In Sect. 5, we discuss the results of the experimentation and conclude our work in Sect. 6.

## 2 Previous Work

### 2.1 *Increasing Data Use*

It is well established that a deep learning model will have increased performance if provided with more training data. Acquiring additional data may be achieved by manually gathering data, a process commonly performed either as an individual or part of a group/company, by using freely available data licensed as open-source, or via methods such as crowdfunding [2, 4]. Alternatively, data augmentation may be used. Data augmentation is a powerful technique that increases the amount of data available to a model [5]. Instead of collecting more data samples, currently, existing data samples may be modified to give new, unique datapoints for training. This is preferable over manual data collection, as annotated data is relatively scarce and can be difficult to obtain, as well as requiring expert knowledge in order to label and validate the data [2, 6]. Table 1 shows the current state-of-the-art models for image classification across six datasets. Each model uses some form of data augmentation to artificially increase the amount and variety of data available for training, which gives the model the improved accuracy that puts the model at the top of their leaderboard.

### 2.2 *Reducing Data Use*

While it is popular to use methods of increasing data, there are uses for data reduction techniques. Such practices are uncommon, as basic linear random data exclusion causes an exponential decrease in system performance [14]. Not only is the average accuracy decreased, but the standard deviation of accuracy increases, showing that the training outputs are inconsistent and difficult to verify [14]. Contrary to this, some uses of data reduction techniques give an increase to a neural network model's performance. In fields where labeled data is sparse, for example, medical imaging,

**Table 1** Top performing image classification models and their data augmentation methods

Dataset	No. training images	Model	Top accuracy (%)	Extra data/ augmentation used
MNIST	60,000	Homogeneous ensemble with simple CNN [7]	99.91	Random rotation/translation
CIFAR-10	50,000	ViT-H/14 [8]	99.50	Random horizontal flip, square crop
CIFAR-100	50,000	EffNet-L2 (SAM) [9]	96.08	Horizontal flip, padding, random crop, AutoAugment [10]
SmallNorb	24,300	Efficient-CapsNet [11]	98.77	Downsampling, random $32 \times 32$ crop, random brightness
FlowerNet-102	1020	CCT-14/7 $\times$ 2 [12]	99.72	RandAugment [5], AutoAugment [10]
iNaturalist 2018	675,170	MetaFormer [13]	88.70	RandAugment [5], AutoAugment [10]

the data often suffers from class imbalance. In such cases, undersampling techniques may be employed to reduce the amount of data of the majority class, alleviating the effects of class imbalance [15]. Less success is to be gained with random exclusion, but methods such as cluster centroids [16] and Tomek links [17] select the most appropriate data to remove; they are often used with oversampling techniques to give an increase in model performance [18]. Undersampling is a technique used primarily to resolve class imbalance and is therefore unsuitable for uniformly balanced datasets. However, for balanced datasets, methods exist for identifying the ‘usefulness’ of data, which allows less useful data to be removed from training. Dimensionality reduction with algorithms such as uniform manifold approximation projection (UMAP) [19] allows a dataset to be observable in as little as 2 or 3 dimensions. By applying the UMAP algorithm to a dataset, the centroid of each class (where all its datapoints revolve) can be calculated, and data may then be removed based on its distance from its centroid [20]. Surprisingly, by removing data closest to its class’s centroid, accuracy is increased, despite fewer training evaluations being performed [20].

**2.3 Variable Data Use**

Varying data usage is the process of using different data throughout training. It is a novel concept; the convention is that a model is trained with the same data over multiple iterations, to allow the model’s weights and biases to converge to a stable



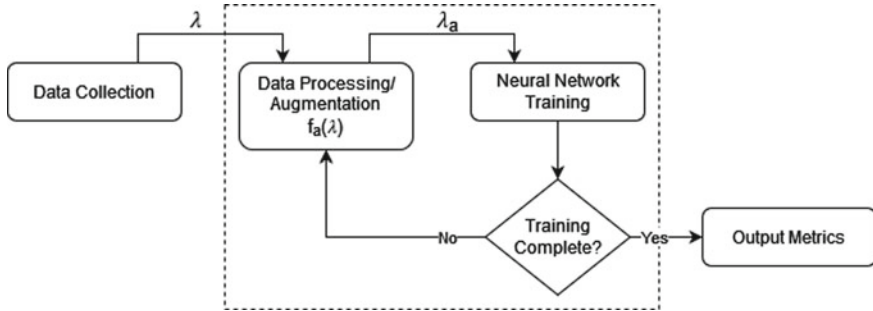
state. It would seem contradictory to add a change to a system that would disturb this convergence; however, there are cases where its use has improved performance, discussed below.

Real-time or online augmentation creates unique data each epoch. Compared to offline augmentation, where the data is augmented once (before the model is run), online augmentation performs augmentation before each epoch during training [21]. Augmentation is commonly used to combat issues such as overfitting and shortage of data [22], but for datasets with well-balanced data, it is also used to achieve higher levels of generalisability [23]. Transfer learning is used to improve a model from one domain by transferring information from a related domain [24]. It may also be used to transfer knowledge within the same domain, to reinforce the model when new data becomes available [25].

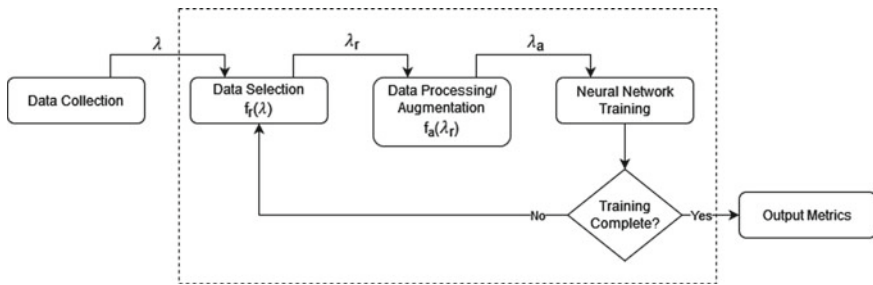
Foremost, transfer learning is a technique used to reduce the amount of work needed to train a model [25]; a model may be trained with one set of data, and the outputs recorded; then, when more data is made available, training can recommence, but with the previous outputs used as a starting point. This means that the model need not be retrained with the original data each time more data is added. In a way, this is varying data usage; throughout the course of training, different data is used. The unique property of transfer learning is that when a model is introduced to new training data, the old training data is not also used. This can often cause catastrophic forgetting [26]. There has been no investigation of using less data at the start of training and increasing data as training progresses. Unlike transfer learning, we have all the data needed at the start of training, but we choose to withhold some of it as training is performed. This is to show the effect of varying data use in a controlled manner.

The experiments described in this paper investigate some extreme cases of reducing data use at earlier stages and show the effect of this on model performance. Our contributions are as follows:

- Three novel methods of dynamically introducing data to a model are described, which each reduce the amount of evaluations performed:
  - Data step
  - Data increment
  - Data cut
- Testing on three datasets is shown, demonstrating the positive and negative effects of these new methods on network accuracy and runtime.
- Evidence is given that these methods may be used reduce the resources required for training while maintaining or improving on the performance of the training output with a runtime reduction of over 50%.



**Fig. 1** Common model structure with online augmentation.  $\lambda$  = full dataset,  $\lambda_a$  = augmented dataset.  $f_a()$  = augmentation function



**Fig. 2** Proposed model structure with dynamic data selection.  $\lambda$  = full dataset,  $\lambda_r$  = reduced dataset,  $\lambda_a$  = augmented dataset.  $f_r()$  = reduction function,  $f_a()$  = augmentation function

### 3 Methods

Traditionally, a deep learning task has three stages: data collection [2], data processing [27, 28], and the training of the neural network model. These steps are most often performed sequentially, but are sometimes somewhat interlinked; for example, the data collection may be real-time image or text capture, or the model may use online augmentation, where data is augmented differently each training loop. Figure 1 shows the data flow of an image classification model with online augmentation. After each training loop, the model returns to the augmentation stage. Figure 2 shows the model structure proposed in this paper, with a data selection stage. This allows the model to select data dynamically, at each loop of training. This in turn enables the various data reduction techniques described further in this paper.

This paper introduces and investigates three sets of experiments where, each epoch, data is selected for training during the data selection stage. The result is a model that does not use all data for each epoch. These methods are

- Data step
- Data increment
- Data cut

**Table 2** Used dataset properties

Dataset	# Training images	# Test images	Image size (pixels)	# Color channels
MNIST	60,000	10,000	$28 \times 28$	1
CIFAR-10	50,000	10,000	$32 \times 32$	3
smallNorb	24,300	24,300	$96 \times 96$	1

**3.1 Datasets**

Three datasets were used for experimentation: the handwritten digits dataset MNIST [29], CIFAR-10 [30], and the model toy dataset smallNORB [31]. Table 2 shows the properties of each dataset.

Multiple variations of experiments were performed, to observe the effect of varying amounts of data reduction on model performance. As such, datasets with small image sizes are ideal, as they required less time to train.

**3.2 Hardware**

Multiple hardware were used to perform training, to make the most of resources available for this research. Training for each dataset was carried out on only one machine, to ensure metrics were consistent across a single dataset. The average CO<sub>2</sub> emission for each GPU is also given, which is used for emission reduction calculations [32]. For training on MNIST:

- CPU: AMD Ryzen 9 3950X 16-Core
- GPU: nVidia GeForce RTX 2070 32GB
- RAM: 64GB
- CO<sub>2</sub> emission per hour: 0.066kg

For training on CIFAR-10:

- CPU: Intel i7 10700
- GPU: RTX 4000
- RAM: 64GB
- CO<sub>2</sub> emission per hour: 0.0922kg

For training on smallNorb:

- CPU: AMD Ryzen 9 3950X 16-Core
- GPU: nVidia GeForce RTX 2070 32GB
- RAM: 64GB
- CO<sub>2</sub> emission per hour: 0.066kg

### 3.3 *Performance Metrics*

These experiments are to observe the effect of reducing data used each epoch of training. To quantify this, the results of the experiments show the total number of evaluations performed; each evaluation is a single image used for training. The runtime of each experiment is also recorded. The successfulness of the model is quantified with the top1 accuracy achieved. The runtime and accuracy for each experiment are directly compared with the baseline for each dataset, to show the increase or decrease in performance in accuracy and runtime. As the model is otherwise unchanged when run with data reduction techniques, the CO<sub>2</sub> reduction of each experiment is directly correlated to the runtime. Therefore, CO<sub>2</sub> reduction is equal to the runtime reduction.

### 3.4 *Network Architecture*

The model used for testing is a convolutional neural network with a simple monolithic architecture. It uses 9 convolutional layers and a primary and secondary capsule layer. The capsule layer uses homogeneous vector capsules, which replace the fully connected layer. The model is based on [23], which shows relatively high performance despite its few network layers. Tests were run for 300 epochs with a batch size of 120. Optimization was performed using the Adam optimizer with an initial learning rate of 0.999, with an exponential decay rate of 0.005 per epoch. These values provide consistent initial settings across all experiments for all datasets.

The datapoints to exclude for each of the methods were selected at random. The experiments were performed to gather an understanding of the effect of the novel dynamic data reduction methods with the simplest method of data exclusion, namely random data exclusion.

## 4 Experimentation

### 4.1 *Benchmark*

Table 3 shows the benchmark runtime and accuracy for MNIST, CIFAR-10, and smallNorb. These results will be used to compare the data reduction experiments, to show the reduction in runtime and effect on performance. The number of total image evaluations is the number of training images evaluated multiplied by the number of training epochs.

**Table 3** Benchmark results for MNIST, CIFAR-10, and smallNorb datasets

Dataset	Dataset size	#Training epochs	#Total image evaluations	Runtime (hours)	CO <sub>2</sub> emission (kg)	Test accuracy		
						Best (%)	Average (%)	Standard deviation (%)
MNIST	50,000	300	18,000,000	02:01	0.134	99.719	99.701	0.011
CIFAR-10	49,920	300	14,976,000	02:54	0.268	88.825	88.689	0.096
smallNorb	24,240	300	7,272,000	01:27	0.097	93.152	92.984	0.135

4.2 Data Step

The simplest method of dynamically selecting data is to ‘step up’ the data usage at a given point during training. A fraction of the dataset is used for a given number of training loops, after which, the full dataset is used for training. This splits the training process into two sections; Section one (S1), which uses less data, and Section two (S2), which uses the full dataset. Below are definitions for the two sections:

- The number of epochs run in Sect. 1:  $S_1^E$
- The amount of data used in Sect. 1:  $S_1^D$
- The number of epochs run in Sect. 2:  $S_2^E$
- The amount of data used in Sect. 2:  $S_2^D$

With these definitions, we can derive the formula for the total number of evaluations performed:

$$n_o \text{ total evaluations} = S_1^D * S_1^E + S_2^D * S_2^E$$

This data split causes a ‘step up’ in data usage, and model will only train with a fraction of the data for some time. Figures 3, 4, and 5 show how this split is applied between these sections. The hypothesis is that due to less data being processed in Section S1, there will be a reduction in runtime. After the step, the full dataset is used; this is to ensure that all features are made available at some point in training, although not at every epoch. This is to aid the reduction of overfitting, which is common when too little data is used. Experimentation for the data step method is split into three parts, with three experiments each.

**Starting with 25% of the dataset** Three experiments were performed, and each experiment is comprised of two sections. Each section corresponds to the amount of data used in a single epoch. Epochs in in  $S_1^E$  use  $S_1^D$ , and epochs in  $S_2^E$  use the full dataset. Each of the experiments switch between these sections at a different epoch, shown in Table 4. The values of  $S^D$  vary between datasets as they each have a different amount of training images.

**Table 4** Part 1 initial data and epoch distribution

Experiment	$S_1^D$ (%)	$S_2^D$ (%)	$S_1^E$ (%)	$S_2^E$ (%)
E1	25	100	25	75
E2	25	100	50	50
E3	25	100	75	25

Figure 3 shows graphs of the data use, for the MNIST dataset with 60,000 training images. The total area under the line represents the total number of evaluations (Table 5).

This first phase of experiments sees the largest decrease in evaluations performed compared to each benchmark. In particular, experiment E3 has the largest decrease out of all the data step experiments. As such, if it is assumed that modifying number of evaluations performed would affect the outputs of the program, then these experiments would differ the most from the benchmarks. The results of this first phase of experiments have shown that when fewer evaluations are performed, runtime is reduced. This was expected. More importantly is the effect this has the accuracy:

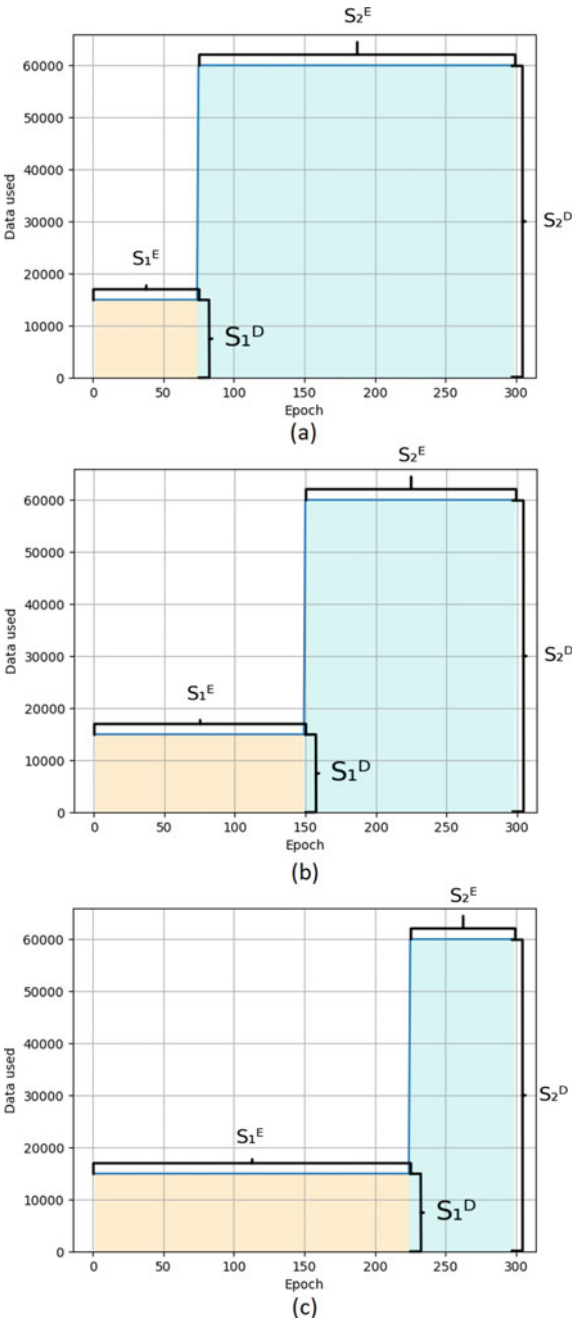
- With the MNIST dataset, there has been a slight increase in average and best accuracy: E1, E2, and E3 show an average accuracy increase of 0.01%, 0.024% and 0.008%, respectively.
- For the CIFAR-10 dataset, there is a much more noticeable decrease in accuracies. E1, E2, and E3 show average accuracy decrease of 0.507%, 1.141%, and 2.346%, respectively.
- For the smallNorb dataset, there is also a decrease in accuracy: E1, E2, and E3 show average accuracy decrease of 0.043%, 0.425%, and 0.962%, respectively.

Based on the previous research on reducing data used for training, the results on CIFAR-10 and smallNorb datasets are perhaps to be expected. It is curious that the MNIST shows an increase in performance, albeit a slight one. What is interesting still is that for both MNIST and smallNorb datasets, the standard deviation of their accuracies is greater than their benchmarks, but for CIFAR-10, the standard deviation is lower. There seems to be no pattern, and all three datasets show unique characteristics.

**Starting with 50% of the dataset** The next three experiments under the data step method follow the same form as in part one, but the initial data  $S_1^D$  is 50% of the full dataset. This is shown in Table 6, and Fig. 4 shows graphs that demonstrate this with the MNIST dataset. This set of experiments is to observe the effect of a different amount of data reduction. Experimental results are shown in Table 7.

The results of this second phase of experiments mirrors those of the previous phase of experiments. We observe that with fewer evaluations performed, runtime is reduced; though as more evaluations are performed compared to the previous phase, the runtime is decreased less. The accuracies show the following patterns:

**Fig. 3** Data step experiments starting with 25% data. **a** Step at 25% through training, **b** Step at 50% through training, **c** Step at 75% through training



**Table 5** Data step part 1 experimental results

Test type	# Image evaluations	Runtime (hours)	Test accuracy			Percentage difference		
			Best (%)	Average (%)	Standard deviation	Runtime (%)	Best accuracy (%)	Average accuracy (%)
MNIST								
Benchmark	18,000,000	02:01	99.719	99.701	1.10E-04			
E1	14,678,880	01:43	99.739	99.711	2.79E-04	-15.148	0.020	0.010
E2	11,312,880	01:22	<b>99.749</b>	<b>99.725</b>	<b>1.82E-04</b>	-32.627	<b>0.030</b>	<b>0.024</b>
E3	7,946,880	00:58	<b>99.749</b>	99.709	2.35E-04	<b>-51.742</b>	<b>0.030</b>	0.008
CIFAR-10								
Benchmark	14,976,000	02:54	88.825	88.689	9.59E-04			
E1	12,214,320	02:25	<b>88.675</b>	<b>88.239</b>	2.78E-03	-16.467	-0.170	<b>-0.507</b>
E2	9,415,320	01:54	87.801	87.677	<b>1.49E-03</b>	-34.293	<b>-1.153</b>	-1.141
E3	6,616,320	01:22	86.878	86.608	2.55E-03	<b>-52.555</b>	-2.193	-2.346
smallNORB								
Benchmark	7,272,000	01:27	93.152	92.984	1.35E-03			
E1	5,931,120	01:15	<b>93.172</b>	<b>92.944</b>	<b>2.22E-03</b>	-13.922	<b>0.022</b>	<b>-0.043</b>
E2	4,572,120	01:04	93.003	92.588	3.70E-03	-26.674	-0.159	-0.425
E3	3,213,120	00:51	92.628	92.089	4.59E-03	<b>-41.191</b>	-0.562	-0.962

**Table 6** Part 2 initial data and epoch distribution

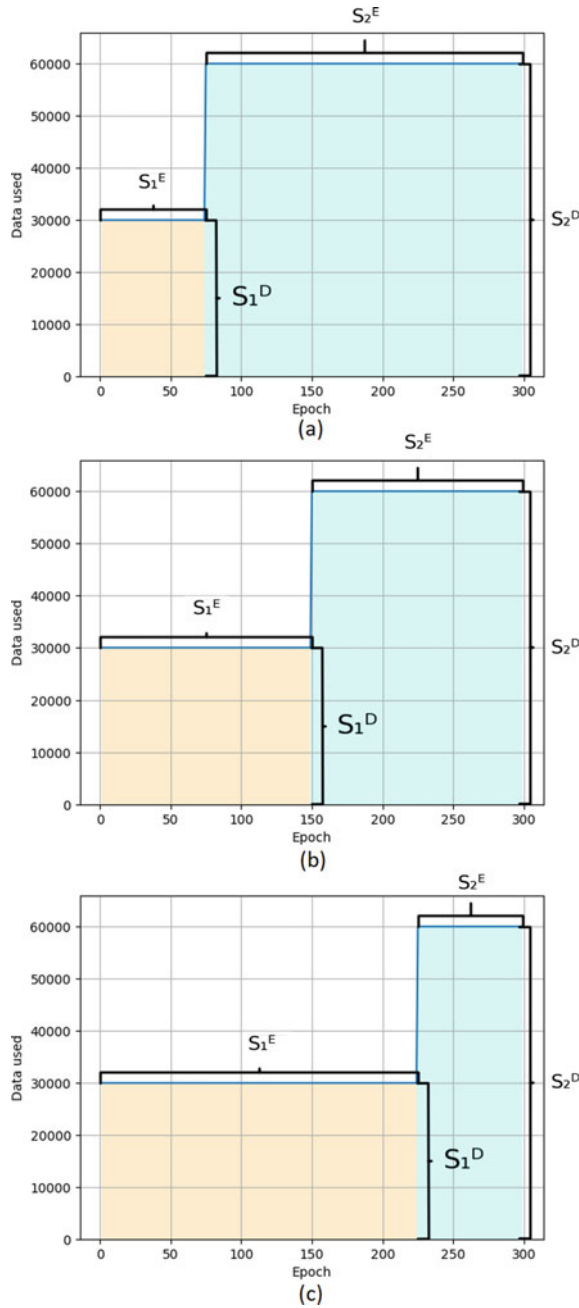
Experiment	$S_1^D$ (%)	$S_2^D$ (%)	$S_1^E$ (%)	$S_2^E$ (%)
E1	50	100	25	75
E2	50	100	50	50
E3	50	100	75	25

- With the MNIST dataset, there is a slight increase in average and best accuracy for all but one experiments: E4 shows an average accuracy increase of 0.014%, while E5 and E6 show average accuracy decrease of 0.002% and 0.006%, respectively.
- For the CIFAR-10 dataset, there is still a decrease in accuracies across all experiments, but not as steep a decline as in the previous experiments. E4, E5, and E6 show average accuracy decrease of 0.457%, 0.779% and 1.103%, respectively.
- For the smallNorb dataset, there is also an accuracy decrease, but less extreme than in the previous phase. E4, E5, and E6 show average accuracy decrease of 0.006%, 0.520%, and 0.084%, respectively. The average accuracy of experiment E5 is also much higher than the others and also has a higher standard deviation, which shows that the metrics are less stable.

The expectation is that as these experiments use more data than the last set, the accuracies should be improved. However, experiments E5 and E6 for MNIST and E5 for smallNorb show worse accuracies. This may show that dynamic data reduction in this particular way hinders the accuracy of the model, for these particular datasets.



**Fig. 4** Data step experiments starting with 50% data. **a** Step at 25% through training, **b** Step at 50% through training, **c** Step at 75% through training



**Table 7** Data step part 2 experimentation results

Test type	# Image evaluations	Runtime (hours)	Test accuracy			Percentage difference		
			Best (%)	Average (%)	Standard deviation	Runtime (%)	Best accuracy (%)	Average accuracy (%)
MNIST								
Benchmark	18,000,000	02:01	99.719	99.701	1.10E-04			
E4	15,788,880	01:53	<b>99.759</b>	<b>99.715</b>	2.52E-04	-6.638	<b>0.040</b>	<b>0.014</b>
E5	13,547,880	01:39	99.719	99.699	<b>1.59E-04</b>	-18.038	0.000	-0.002
E6	11,306,880	01:24	99.729	99.695	2.98E-04	<b>-30.597</b>	0.010	-0.006
CIFAR-10								
Benchmark	14,976,000	02:54	88.825	88.689	9.59E-04			
E4	13,137,840	02:35	<b>88.464</b>	<b>88.283</b>	<b>1.78E-03</b>	-10.458	<b>-0.407</b>	<b>-0.457</b>
E5	11,274,840	02:14	88.434	87.998	3.01E-03	-22.581	-0.441	-0.779
E6	9,411,840	01:55	88.243	87.711	3.00E-03	<b>-33.831</b>	-0.656	-1.103
smallNORB								
Benchmark	7,272,000	01:27	93.152	92.984	1.35E-03			
E4	6,384,000	01:20	<b>93.564</b>	<b>92.978</b>	4.45E-03	-8.475	<b>0.443</b>	<b>-0.006</b>
E5	5,484,000	01:11	93.263	92.500	6.19E-03	-18.317	0.120	-0.520
E6	4,584,000	01:04	93.544	92.905	<b>4.39E-03</b>	<b>-26.675</b>	0.421	-0.084

**Table 8** Part 3 initial data and epoch distribution

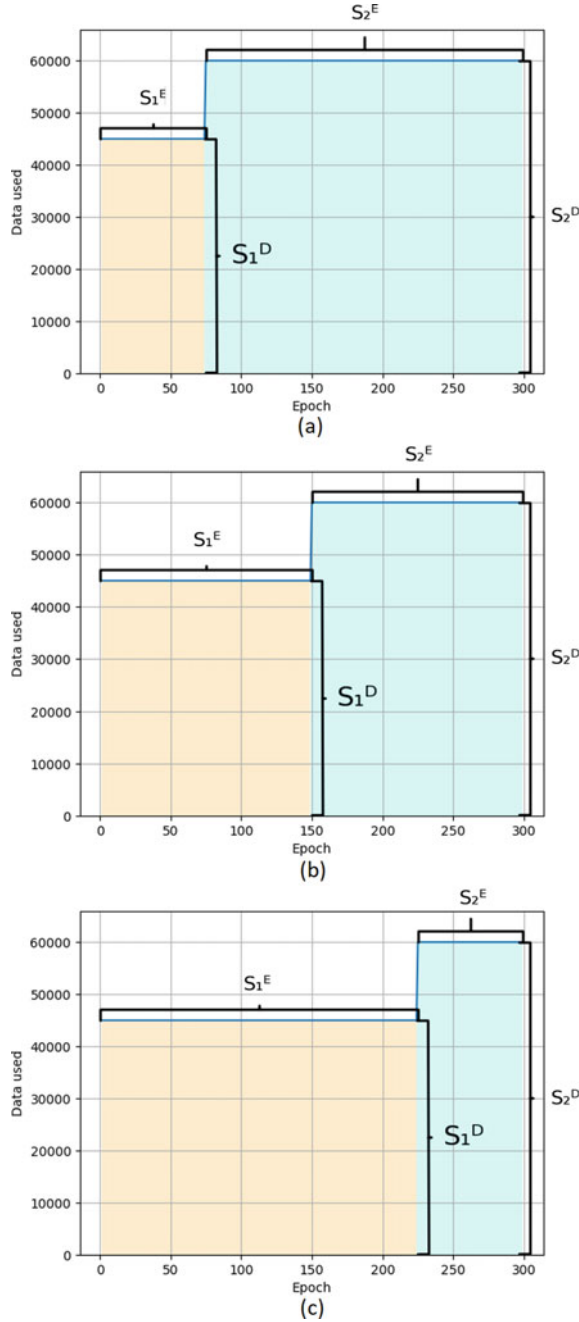
Experiment	$S_1^D$ (%)	$S_2^D$ (%)	$S_1^E$ (%)	$S_2^E$ (%)
E1	75	100	25	75
E2	75	100	50	50
E3	75	100	75	25

**Starting with 75% of the dataset** The last lot of three experiments were performed, also with two sections each. Epochs in  $S_1^E$  use  $S_1^D$ , and epochs in  $S_1^E$  use the full dataset  $S_2^D$ . In the same manner as the previous two lots of experiments, each of the experiments switch between these sections at a different epoch, shown in Table 8. Figure 5 shows graphs that demonstrate this. The aim of these experiments is to observe the effect of a smaller data reduction than in the previous tests. Table 9 shows the results of the experimentation.

This final phase of testing uses the least data reduction, and as such, the decrease in both runtime and average accuracy is the smallest. The following patterns are observed:

- With the MNIST dataset, there is an increase in accuracy for all experiments: E7, E8, and E9 show an average accuracy increase of 0.018%, 0.016%, and 0.012%, respectively. Experiment E7 shows a decrease in standard deviation of accuracies, compared to the benchmark.

**Fig. 5** Data step experiments starting with 75% data. **a** Step at 25% through training, **b** Step at 50% through training, **c** Step at 75% through training



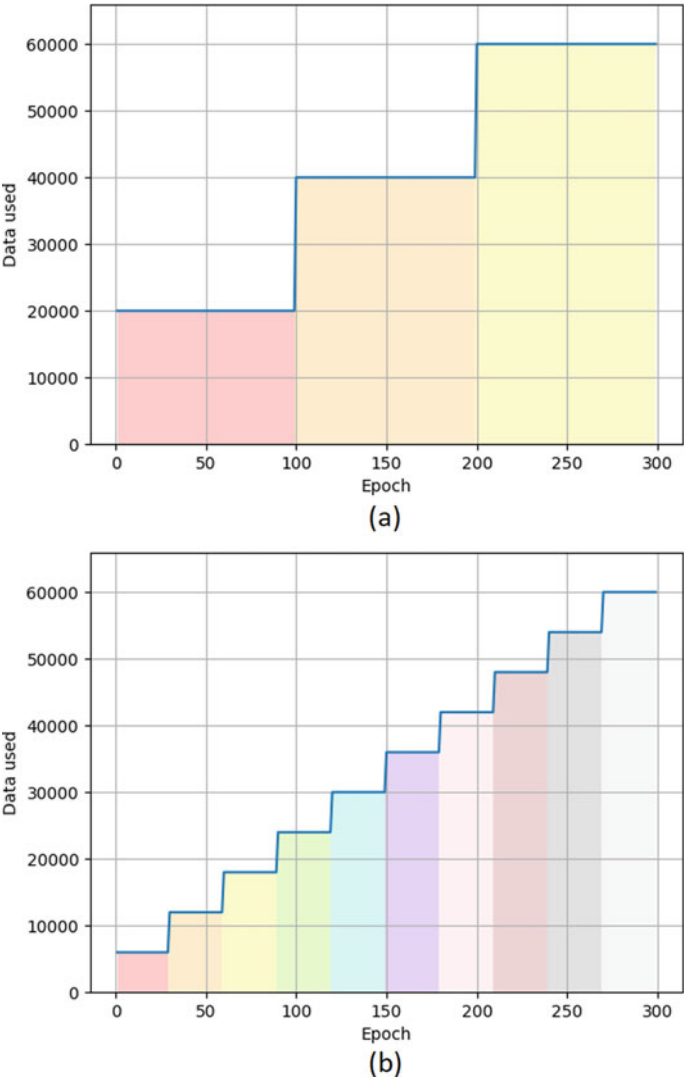
**Table 9** Data step part 3 experimentation results

Test type	# Image evaluations	Runtime (hours)	Test accuracy			Percentage difference		
			Best (%)	Average (%)	Standard deviation	Runtime (%)	Best accuracy (%)	Average accuracy (%)
MNIST								
Benchmark	18,000,000	02:01	99.719	99.701	1.10E-04			
E7	16,898,880	01:55	<b>99.729</b>	<b>99.719</b>	<b>1.00E-04</b>	-5.412	<b>0.010</b>	<b>0.018</b>
E8	15,782,880	01:54	<b>99.729</b>	99.717	1.31E-04	-5.713	<b>0.010</b>	0.016
E9	14,666,880	01:47	<b>99.729</b>	99.713	1.52E-04	<b>-11.948</b>	<b>0.010</b>	0.012
CIFAR-10								
Benchmark	14,976,000	02:54	88.825	88.689	9.59E-04			
E7	14,061,360	02:47	<b>88.675</b>	<b>88.625</b>	5.59E-04	-4.027	<b>-0.170</b>	<b>-0.072</b>
E8	13,134,360	02:36	88.655	88.402	2.98E-03	-9.888	-0.192	-0.324
E9	12,207,360	02:25	88.353	88.239	<b>1.14E-03</b>	<b>-16.565</b>	-0.531	-0.507
smallNORB								
Benchmark	7,272,000	01:27	93.152	92.984	1.35E-03			
E7	6,828,000	01:23	<b>93.705</b>	<b>93.059</b>	5.45E-03	-5.036	<b>0.593</b>	<b>0.081</b>
E8	6,378,000	01:19	93.581	92.926	5.19E-03	-9.528	0.461	-0.062
E9	5,928,000	01:16	93.300	92.751	<b>3.85E-03</b>	<b>-13.423</b>	0.159	-0.250

- With the CIFAR-10 dataset, the results follow the same patterns as with the results of the last groups of experiments—the accuracy decrease directly correlates to the reduction in data used, although this correlation is non-linear. E7, E8, and E9 show an average accuracy decrease of 0.072%, 0.324%, and 0.507%, respectively.
- With the smallNorb dataset, E7 shows an average accuracy increase of 0.081%, and E8 and E9 show an average accuracy decrease of 0.062% and 0.250%, respectively. There is an increase in performance with experiment E7; this is the only experiment on the smallNorb dataset to have an increase in performance. The average accuracies of experiments E8 and E9 are decreased, which shows how much of an effect such a slight difference in number of evaluations performed has on the model.

4.3 Data Increment

Like the methods described in the step section, the incrementing method uses a fraction of the data at the start of training. At a previously defined interval, more of the data is added incrementally, causing multiple smaller steps in data usage throughout the training. As such, the number of sections present is equal to the number of increments.



**Fig. 6** Data increment experiments, with **a** 33% data increments, **b** 10% data increments

Adding the data this way will decrease the time spent training for the earlier epochs, while ensuring that all training data is eventually used. Figure 6 shows example data usage graphs with varying increment intervals. Table 10 shows the results of the experimentation.

Using the data increment method causes much less data to be used throughout training. The most extreme case of this is with steps of 0.33%, where roughly half of the evaluations are performed. This makes the data increment method a more extreme data reduction method than the data step method. For each dataset, the observations are as follows:

**Table 10** Data increment experimentation results

Test type	# Image evaluations	Runtime (hours)	Test accuracy			Percentage difference		
			Best (%)	Average (%)	Standard deviation	Runtime (%)	Best accuracy (%)	Average accuracy (%)
MNIST								
Benchmark	18,000,000	02:01	99.719	99.701	1.10E-04			
33%	11,992,080	01:23	<b>99.739</b>	<b>99.703</b>	2.89E-04	-31.790	<b>0.020</b>	<b>0.000</b>
25%	11,241,120	01:20	99.719	99.699	<b>1.42E-04</b>	-34.210	0.000	<b>0.000</b>
10%	9,903,720	01:09	99.709	99.687	2.50E-04	-42.630	-0.010	-0.010
5%	9,456,720	01:07	99.729	99.701	1.93E-04	-44.820	0.010	<b>0.000</b>
1%	9,099,120	01:08	99.709	99.677	2.18E-04	-44.090	-0.010	-0.020
0.33%	9,041,880	<b>01:04</b>	99.719	99.695	1.96E-04	<b>-47.310</b>	0.000	-0.010
CIFAR-10								
Benchmark	14,976,000	02:54	88.825	88.689	9.59E-04			
33%	9,981,360	02:00	<b>88.153</b>	<b>87.723</b>	2.69E-03	-30.920	<b>-0.760</b>	<b>-1.090</b>
25%	9,361,560	01:52	87.771	87.582	2.18E-03	-35.130	-1.190	-1.250
10%	8,253,000	01:42	87.550	87.317	1.52E-03	-41.380	-1.440	-1.550
5%	7,877,520	01:36	87.490	87.275	1.92E-03	-44.530	-1.500	-1.590
1%	7,581,360	<b>01:33</b>	87.299	87.106	<b>1.25E-03</b>	-46.370	-1.720	-1.780
0.33%	7,531,560	<b>01:33</b>	87.560	87.225	2.06E-03	<b>-46.560</b>	-1.420	-1.650
smallNORB								
Benchmark	7,272,000	01:27	93.152	92.984	1.35E-03			
33%	4,840,200	01:07	<b>92.954</b>	<b>92.482</b>	3.61E-03	-23.040	<b>-0.210</b>	<b>-0.540</b>
25%	4,536,240	01:03	92.628	92.463	2.21E-03	-28.180	-0.560	-0.560
10%	3,999,840	01:02	92.768	92.373	3.51E-03	-28.850	-0.410	-0.660
5%	3,821,040	00:59	92.735	92.291	3.45E-03	-32.180	-0.450	-0.740
1%	3,676,560	00:57	92.562	92.115	4.91E-03	-34.520	-0.630	-0.930
0.33%	3,651,600	<b>00:56</b>	92.525	92.248	<b>2.39E-03</b>	<b>-35.700</b>	-0.670	-0.790

- For the MNIST dataset, average accuracy is maintained with 33 and 25% increments but is never improved upon the baseline. For both of these cases, runtime is shortened by at least 30%.
- For the CIFAR-10 dataset, all average accuracies are worse by at least 1%. Like MNIST, runtime is shortened by a minimum of 30%.
- For the smallNorb dataset, as the dataset has much fewer training images, both the runtime and accuracy reduction are not as large as with other datasets. For all experiments, average accuracy is reduced by less than 1%, while runtime is reduced by between 20 and 40%.

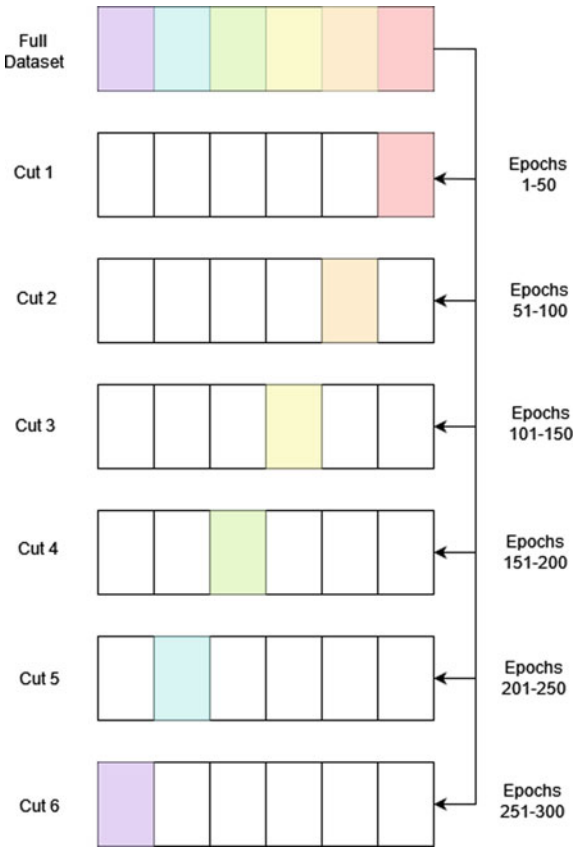
Based on these results, the data increment method seems to have a more detrimental effect on network performance. It is also worth noting that the standard deviation of average accuracies are worse than the benchmark across all experiments, meaning the networks are less reliable at getting consistent results.

4.4 Data Cut

The data cut method splits a dataset into multiple cuts of equal size. Each cut is then used for an equal number of training epochs. The more splits a deck is cut into, the less data is present in each split. As such, the more splits there are, the fewer evaluations will be performed. Figure 7 shows how a dataset may be divided into 6 cuts, for training over 300 epochs.

This approach is similar to transfer learning techniques, in which a network that has already been trained on some data is introduced to a new set of training data. The new input data contains the same classes as it had already been trained with, so the network does not need to adapt to new classes. However, while the previous methods in this paper introduced new training data to accompany previously used data, this method does not include the previously used data when new data is introduced - it continues training on only new data. Normally, when introducing new data via transfer learning, a fresh network is deployed, and given the weights and biases of

**Fig. 7** Dataset is split into 6 sections (cuts). A single split is used for training every 50 epochs out of 300



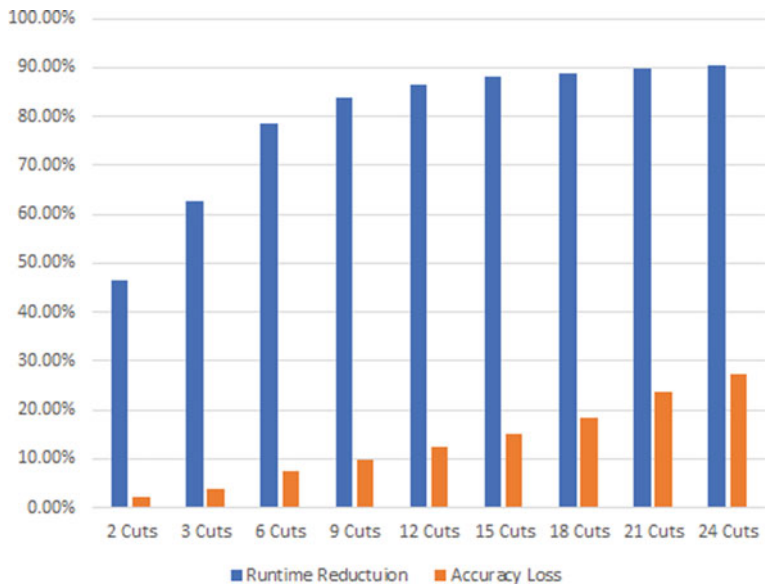
**Table 11** Data cut experimentation results

Test type	# Image evaluations	Runtime (hours)	Test accuracy			Percentage difference		
			Best (%)	Average (%)	Standard deviation	Runtime (%)	Best accuracy (%)	Average accuracy (%)
MNIST								
Benchmark	18,000,000	02:01	99.719	99.701	1.10E-04			
2 Splits	9,000,000	01:07	<b>99.719</b>	<b>99.699</b>	<b>1.23E-04</b>	-44.272	<b>0.000</b>	<b>-0.002</b>
3 Splits	5,976,000	00:44	99.679	99.655	1.68E-04	-63.331	-0.040	-0.046
6 Splits	2,988,000	00:25	99.659	99.596	4.10E-04	-78.800	-0.060	-0.105
9 Splits	1,980,000	00:19	99.598	99.526	6.82E-04	-84.016	-0.121	-0.175
12 Splits	1,476,000	00:16	99.508	99.337	1.41E-03	-86.604	-0.211	-0.365
15 Splits	1,188,000	00:14	99.367	99.102	2.85E-03	-88.074	-0.352	-0.600
18 Splits	972,000	00:13	99.297	98.855	3.86E-03	-89.124	-0.423	-0.848
21 Splits	828,000	00:12	99.237	97.476	1.83E-02	-89.909	-0.483	-2.232
24 Splits	720,000	00:11	96.365	95.645	7.33E-03	<b>-90.501</b>	-3.363	-4.068
CIFAR-10								
Benchmark	14,976,000	02:54	88.825	88.689	9.59E-04			
2 Piles	7,488,000	01:33	<b>86.777</b>	<b>86.528</b>	<b>1.54E-03</b>	-46.518	<b>-2.306</b>	<b>-2.436</b>
3 Splits	4,968,000	01:04	85.492	85.143	3.20E-03	-62.692	-3.753	-3.998
6 Splits	2,484,000	00:37	82.269	81.984	3.27E-03	-78.491	-7.381	-7.560
9 Splits	1,656,000	00:28	80.110	79.673	2.65E-03	-83.690	-9.811	-10.166
12 Splits	1,224,000	00:23	77.791	77.072	6.68E-03	-86.431	-12.422	-13.098
15 Splits	972,000	00:20	75.552	72.888	2.02E-02	-88.013	-14.943	-17.816
18 Splits	828,000	00:19	72.440	69.066	2.70E-02	-88.971	-18.447	-22.125
21 Splits	684,000	00:17	67.731	66.030	1.94E-02	-89.880	-23.748	-25.548
24 Splits	612,000	00:16	64.518	60.600	4.62E-02	<b>-90.370</b>	-27.365	-31.671
smallNORB								
Benchmark	7,272,000	01:27	93.152	92.984	1.35E-03			
2 Splits	3,636,000	00:59	<b>93.106</b>	<b>92.744</b>	3.37E-03	-32.490	<b>-0.049</b>	<b>-0.257</b>
3 Splits	2,412,000	00:51	92.880	92.616	<b>2.36E-03</b>	-41.598	-0.292	-0.396
6 Splits	1,188,000	00:42	92.946	92.104	5.30E-03	-52.148	-0.221	-0.946
9 Splits	792,000	00:38	91.840	91.408	3.93E-03	-55.659	-1.408	-1.694
12 Splits	576,000	00:36	90.924	90.012	7.59E-03	-58.625	-2.392	-3.195
15 Splits	468,000	00:36	89.975	87.791	2.33E-02	-58.774	-3.410	-5.584
18 Splits	396,000	00:36	88.461	87.139	1.05E-02	-58.445	-5.035	-6.286
21 Splits	324,000	00:36	86.865	80.511	8.85E-02	-58.097	-6.749	-13.414
24 Splits	288,000	00:37	86.894	80.776	7.48E-02	<b>-57.390</b>	-6.718	-13.128

previously trained networks. The data cut method aims to give the network different data at different times, while preserving the state of the network throughout a single training cycle. Table 11 shows the results of the experimentation.

Of the three experimental methods described in this paper, the data cut method cuts the number of evaluations performed down the most. It becomes apparent quickly that the more splits are used, the more the accuracy is degraded. Results on the





**Fig. 8** Runtime reduction and accuracy loss with data cut method, on CIFAR-10 dataset

MNIST dataset show that by cutting the dataset into two splits, we almost cut the runtime in half while maintaining the benchmark accuracy. With 24 splits, there is a large drop in performance. This shows a ‘cut-off point’, where too few evaluations are performed, and the model doesn’t have enough iterations to allow the weights and biases to converge, causing more classifications to fail. Figure 8 shows the percentage difference in runtime and accuracy of the CIFAR-10 dataset, compared to its benchmark. The results of this dataset show the biggest percentage differences. The graph shows that by introducing more data splits, the runtime decreases non-linearly, while the accuracy decreases more steadily. Therefore, the most appropriate number of data splits must be chosen to benefit the most from this dataset reduction technique.

5 Discussion

The runtime of the experiment correlates directly with the total number of evaluations performed. This is as expected, as the fewer evaluations are performed, the fewer calculations are performed by the neural network.

For the data step method, the most notable results are those of the MNIST dataset, and E7, E8, and E9 of smallNorb, as they show an increase in accuracy despite the fewer evaluations performed. An future research direction would be to investigate why there is a performance increase in these cases, and how the detrimental effect on stability may be rectified. Also, further research into why CIFAR-10 does not show any increase could be explored.

As with the data step method, the data increment method shows that reducing the number of evaluations performed decreases the runtime. However, despite both methods performing the same fundamental task of data reduction, they produce different results for roughly the same number of evaluations performed. For example, comparing experiments E2 of the data step method and data increments of 25%, we can observe the accuracies when roughly the same number of evaluations are performed.

The number of evaluations for the two experiments are

- E2 for MNIST dataset uses 11,312,880 evaluations, data increments of 25% use 11,241,120 evaluations, a difference of 0.634%.
- E2 for CIFAR-10 dataset uses 9,415,320 evaluations, data increments of 25% use 9,361,560 evaluations, a difference of 0.571%.
- E2 for smallNorb dataset uses 4,572,120 evaluations, data increments of 25% use 4,536,240 evaluations, a difference of 0.785%.

However, despite the marginal difference in number of evaluations performed, the difference in accuracy between these two experiments is

- E2 for MNIST dataset gives an accuracy increase of 0.024%, data increments of 25% give an accuracy difference of 0%.
- E2 for CIFAR-10 dataset gives an accuracy decrease of 1.141%, data increments of 25% give an accuracy decrease of 1.250%.
- E2 for smallNorb dataset gives an accuracy decrease of 0.425%, data increments of 25% give an accuracy difference of 0.560%.

For all cases, the data increment method shows worse accuracy. While it is true that the data increment method has fewer evaluations performed, the difference in data (less than 1% for all datasets) amounts to less than 3 epochs of training. This amount is negligible as the top accuracy of a model settles at much earlier epochs. The results of the experimentation show that decreasing the number of evaluations with the increment method has a more detrimental effect on the accuracy of the model.

Finally, the data cut method shows the largest reduction in evaluations performed. Below is a comparison between experiments data increment of 0.33% and data cut with 2 splits:

- Data increment for MNIST dataset uses 9,041,880 evaluations, data cut with 2 splits use 9,000,000 evaluations, a difference of 0.463%.
- Data increment for CIFAR-10 dataset uses 7,531,560 evaluations, data cut with 2 splits use 7,488,000 evaluations, a difference of 0.578%.
- Data increment for smallNorb dataset uses 3,651,600 evaluations, data cut with 2 splits use 3,636,000 evaluations, a difference of 0.427%.

Comparing the accuracies between these two experiments, we observe

- Data increment for MNIST gives an accuracy decrease of 0.01%, data cut with 2 splits gives an accuracy decrease of 0.02%.

- Data increment for CIFAR-10 gives an accuracy decrease of 1.65%, data cut with 2 splits gives an accuracy decrease of 2.436%.
- Data increment for smallNorb gives an accuracy decrease of 0.79%, data cut with 2 splits gives an accuracy decrease of 0.257%.

Accuracy for all observed data cut experiments is worse than those of the data increment method. We can conclude that, as the data increment method is inferior to the data step method, the data cut method is least effective at maintaining accuracy. However, if the need is to reduce the runtime by as much as possible without hindering accuracy too greatly, a data cut of 9 splits appears to be the best compromise between runtime and accuracy.

## 6 Conclusion

The results of this paper have shown that, contrary to the norm, reducing the data used for training has in some cases improved the performance of the model. It proves that not all data may be necessary for training, and in fact, some data may hinder it.

The approaches used are somewhat brutish, excluding data randomly without consideration of the value of the datapoints removed. Other works cited have shown algorithmic approaches to select which data to remove to enhance performance, which is a next step for varying data usage. Still, having shown that even random exclusion has improved results, it would seem that varying data use is to be explored in further detail.

## References

1. Motamedi M, Sakharnykh N, Kaldewey T (2021) A data-centric approach for training deep neural networks with less data. arXiv preprint. [arXiv:2110.03613](https://arxiv.org/abs/2110.03613)
2. Roh Y, Heo G, Whang SE (2019) A survey on data collection for machine learning: a big data-AI integration perspective. *IEEE Trans Knowl Data Eng* 33(4):1328–1347
3. Linjordet T, Balog K (2019) Impact of training dataset size on neural answer selection models. In: *European conference on information retrieval*, pp 828–835. Springer
4. Thompson NC, Greenewald K, Lee K, Manso GF (2020) The computational limits of deep learning. arXiv preprint [arXiv:2007.05558](https://arxiv.org/abs/2007.05558)
5. Cubuk ED, Zoph B, Shlens J, Le QV (2020) Randaugment: practical automated data augmentation with a reduced search space. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp 702–703
6. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH (2012) Predicting sample size required for classification performance. *BMC Med Inform Decis Making* 12(1):1–10
7. An S, Lee M, Park S, Yang H, So J (2020) An ensemble of simple convolutional neural network models for MNIST digit recognition. arXiv preprint [arXiv:2008.10400](https://arxiv.org/abs/2008.10400)
8. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)

9. Foret P, Kleiner A, Mobahi H, Neyshabur B (2020) Sharpness-aware minimization for efficiently improving generalization. arXiv preprint [arXiv:2010.01412](https://arxiv.org/abs/2010.01412)
10. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV (2019) Autoaugment: learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 113–123
11. Mazzia V, Salvetti F, Chiaberge M (2021) Efficient-capsnet: Capsule network with self-attention routing. Sci Rep 11(1):1–13
12. Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L (2021) Cvt: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 22–31
13. Diao Q, Jiang Y, Wen B, Sun J, Yuan Z (2022) Metaformer: a unified meta framework for fine-grained recognition. arXiv preprint [arXiv:2203.02751](https://arxiv.org/abs/2203.02751)
14. Cho J, Lee K, Shin E, Choy G, Do S (2015) How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv preprint [arXiv:1511.06348](https://arxiv.org/abs/1511.06348)
15. Liu X-Y, Wu J, Zhou Z-H (2008) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B (Cybern) 39(2):539–550
16. Rahman MM, Davis D (2013) Cluster based under-sampling for unbalanced cardiovascular data. In: Proceedings of the world congress on engineering, vol 3, pp 3–5
17. Tomek I (1976) A generalization of the k-NN rule. IEEE Trans Syst Man Cybern 2:121–126
18. Rahman MM, Davis DN (2013) Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput 3(2):224
19. McInnes L, Healy J, Melville J (2018) Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint [arXiv:1802.03426](https://arxiv.org/abs/1802.03426)
20. Byerly A, Kalganova T (2022) Towards an analytical definition of sufficient data. arXiv preprint [arXiv:2202.03238](https://arxiv.org/abs/2202.03238)
21. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):1–48
22. Rai R, Sisodia DS (2021) Real-time data augmentation based transfer learning model for breast cancer diagnosis using histopathological images. In: Advances in biomedical engineering and technology, pp 473–488. Springer
23. Byerly A, Kalganova T, Dear I (2021) No routing needed between capsules. Neurocomputing 463:545–553
24. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big data 3(1):1–40
25. Chen L, Moschitti A (2019) Transfer learning for sequence labeling using source model and target data. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 6260–6267
26. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S (2019) Continual lifelong learning with neural networks: a review. Neural Networks 113:54–71
27. Kong F, Henao R (2022) Efficient classification of very large images with tiny objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2384–2394
28. Kuo CW, Ashmore JD, Huggins D, Kira Z (2019) Data-efficient graph embedding learning for PCB component detection. In: 2019 IEEE winter conference on applications of computer vision (WACV), pp 551–560. IEEE
29. LeCun Y (1998) The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>
30. Krizhevsky A, Hinton G et al (2009) Learning multiple layers of features from tiny images
31. Huang FJ, LeCun Y (2009) The small nobb dataset, v1. 0, 2005
32. Lacoste A, Luccioni A, Schmidt V, Dandres T (2019) Quantifying the carbon emissions of machine learning. arXiv preprint [arXiv:1910.09700](https://arxiv.org/abs/1910.09700)

# Author Index

## A

Acosta, Mario C., 129  
Agarwal, Samarth, 451  
Ahamed, Saahira Banu, 231  
Ahmad, Gofur, 1003  
Ahmed, B. M., 87  
Aini, Qurratul, 1003  
Alamer, Latifah, 231  
Alfimtsev, A. N., 269  
Allani, Mohamed Amine, 715  
Allmendinger, Jessika, 641  
Alqahtani, Iman Mohammad, 231  
Ammar, Hossam Hassan, 189  
Amrurrobbi, Azka Abdi, 1035  
Andaluz, Víctor H., 383  
Andrade de, Francisco Pacheco, 535  
Androcec, Darko, 205  
An, Jing, 939  
Antonopoulos, Antonis, 467  
Arias-Flores, Hugo, 811  
Arias, Rino, 439  
Asami, Koji, 139  
Asano, Haruki, 521  
Avdalyan, Narek, 1  
Azeez, Stephen, 949

## B

Baharuddin, Tawakkal, 1027  
Bajaj, Nikesh, 655  
Balayev, Rasul, 321  
Barrena García, Cristina, 109  
Basu, Gaurav, 451  
Batraga, Anda, 785  
Bernardes, Marciele Berger, 535

Bocklisch, Franziska, 33  
Boonwatthanakul, Chanoknart, 613  
Borja-Galeas, Carlos, 811  
Brandtner, Patrick, 641  
Brenes-García, Luis F., 545  
Butakova, Maria A., 429

## C

Cabacungan, Paul M., 745  
Cai, Shuyao, 683  
Cantú-Ortiz, Francisco J., 397, 499, 545, 855  
Cao, Reymond P., 745  
Carmona-Herrera, Rodrigo, 397  
Castro, John W., 439  
Ceballos-Cancino, Héctor, G., 397, 499, 545, 855  
Chasi-Pesantez, Paul A., 489  
Cheng, Yun-Quan, 569  
Chin, Ji-Jian, 569  
Clavijo, Luis, 73  
Colace, Francesco, 1075  
Conte, Dajana, 1075  
Cortizo, Lucas, 535  
Cruz Aida La, De, 383

## D

Dac, Hoan Nguyen, 417  
Dajao, Edgardo S., 601  
Dakić, Pavle, 477  
Damiani, Ernesto, 11  
D'Arienzo, Maria Pia, 1075  
Dawson, Matthew, 129

Delgado, Johnny, 73  
 Dobkacz, L. Y., 269  
 Dubovik, Irina, 23  
 Duy, Huynh Anh, 255

## E

Edison, Tapia, 383  
 Edisson, Coronel-Villavicencio, 1047  
 Elatrachi, Mounir, 879  
 Espinoza-Saquicela, Darío, 843

## F

Farat, Oleksandra, 339  
 Fatmawati, Indah, 1015  
 Felipe, Juan, 705  
 Fernando, Guerrero-Vasquez, 971  
 Flores-Siguenza, Pablo, 843  
 Fujino, Yuichi, 349  
 Fukuda, Shuichi, 373  
 Fuller, Ged, 655

## G

Gabel, Thomas, 715  
 Galobardes, Eduardo Cesar, 129  
 García-Pando, Carlos Pérez, 129  
 Garrido Madrigal, Aitor, 109  
 Garzón, Christian, 705  
 Gayathri, Obulreddigari, 451  
 Gelera, Aries M., 591  
 Gertsy, Alexander, 329  
 Godehardt, Eicke, 715  
 Gonzales, Alyssa Joi A., 591  
 Gonzales, Cristina F., 745  
 Gonzalez, Ricardo Romero, 867  
 Granada, Salvador P., 745  
 Guerrero-Vásquez, Luis F., 489  
 Gupta, Pranjal, 309  
 Gurunarayanan, S., 509

## H

Haar Van Der, Dustin, 1057  
 Han, Xu, 939  
 Hanyurwimfura, Damien, 405  
 Haponen, Ari, 1085  
 Harada, Fumiko, 629  
 Harasym, Lyudmyla, 339  
 Hasuike, Takashi, 667  
 Henna, Shagufta, 949  
 Hernández-Manrique, Victor, 397  
 Hidayah, Nur, 1003

Hinojosa-Ramos, Miriam Vanessa, 895  
 Hladkyi, Oleksandr, 329  
 Hung, Phan Duy, 255  
 Hussin, Ab Razak Che, 819

## I

Ibrahimov, Fuad, 321  
 Iribaram, Suparto, 1027  
 Ishida, Nagito, 139

## J

Jain, Chirag, 309  
 Jia, Yanxia, 683  
 Jorba, Oriol, 129  
 Jubba, Hasse, 1027

## K

Kaibe, Liene, 785  
 Kakade, Manoj Subhash, 451  
 Kalganova, Tatiana, 1105  
 Kalgin, Y. A., 269  
 Kalita, Sampreet, 581  
 Karakostas, Bill, 467  
 Karim, Hezerul Abdul, 569  
 Karuppiyah, Anupama, 451, 509  
 Kar-Weng, Ban, 569  
 Kashima, Tomoko, 667  
 Katayama, Shogo, 139  
 Katsoulakos, Takis, 467  
 Kellerte, Katrina, 785  
 Khakurel, Jayden, 1085  
 Khubrani, Mousa, 231  
 Kim, Song-Kyoo, 11  
 Kizyan, Sergey, 23  
 Klichowicz, Anja, 33  
 Kobayashi, Haruo, 139  
 Koman, Bohdan, 339  
 Komariah, Ade, 993  
 Kontsevoi, Boris, 23  
 Koppad, Deepali B., 695  
 Korablyov, Mykola, 361  
 Koumpian, Elizabeth, 97  
 Kulivnuk, Volodymyr, 329  
 Kumari, Deepa, 309  
 Kuwana, Anna, 139  
 Kuzmin, Ivan, 329

## L

Lacatan, Luisito Lolong, 557  
 LaMont, Logan, 655

Latorre, Roberta Valeria, 979  
 Lecca, Paola, 979  
 Lennon, Patrick, 763  
 Lennon, Ruth G., 763  
 Leon, Marcelo, 895  
 Leon, Paulina, 895  
 Le, Van-Hung, 723  
 Liang, Qiao, 733  
 Libatique, Nathaniel Joseph C., 745  
 Licona-Luque, Juan P., 545  
 Li, Cun, 733  
 Llerena, Lucrecia, 439  
 Lojano-Angamarca, Marcelo D., 489  
 Lombardi, Giulia, 979  
 Lo, Nai-Wei, 11  
 Loukantschewsky, Milen, 245  
 Luis, Serpa-Andrade, 971, 1047  
 Lutskey, Sergey, 361

## M

Maeko, Mapula Elisa, 1057  
 Makris, George C., 429  
 Maldonado-Quezada, José, 895  
 Mandová, Lucia, 833  
 Mankodi, Amit, 923  
 Mansor, Sarina, 569  
 Mata-Quevedo, Jean P., 867  
 Matsuhara, Masafumi, 521  
 Matsumoto, Shimpei, 667  
 Matsumoto, Shuichi, 349  
 Mavrakos, Anargyros, 467  
 Mendez-Vargas, Jorge A., 855  
 Méndez-Vásquez, Rafael B., 489  
 Mendoza-Leal, Gabriela G., 855  
 Mercado, Neil Angelo M., 745  
 Minh Le, Tu, 417  
 Mkansi, Marcia, 51  
 Morgenstern, Tina, 33  
 Moscoso-Martínez, Marlon, 843  
 Mostafa, Aya Essam, 189  
 Mostafa, Mohamed Essam, 189  
 Mujahid, Umar, 961  
 Mysiuk, Iryna, 339  
 Mysiuk, Roman, 339

## N

Nabablit, Karlo Jose E., 601  
 Nadprasert, Phisit, 613  
 Nethra, N., 695  
 Netke, Ashay, 309  
 Nsenga, Jimmy, 405

Nuaimy Al, Louay, 159

## O

Ogore, Marvin, 405  
 Opatovský, Martin, 833  
 Oppus, Carlos M., 745  
 Ordoñez-Ordoñez, Jorge O., 489  
 Oreski, Dijana, 205  
 Ortega, Juan, 73  
 Osorio, Ana, 439  
 Oukarfi, Samira, 879

## P

Pahlevi, Moch Edward Trias, 1035  
 Panda, Subhrakanta, 309  
 Pandey, Rohit, 799  
 Paranthaman, Pratheep Kumar, 655  
 Paul, Mata-Quevedo, 971  
 Petrosyan, Armen, 1  
 Pheeraphan, Nutteerat, 293  
 Porio, Emma E., 745  
 Poteat, Thomas, 655  
 Poulakis, Dimitrios, 429  
 Prieto-Egido, Ignacio, 109  
 Purwaningsih, Titin, 1035  
 Putkhiao, Likkhasit, 613

## Q

Qodir, Zuly, 1027  
 Quang, Diep Pham, 417  
 Quevedo, Sebastian, 73

## R

Raikov, Aleksandr, 911  
 Ramovha, Tshililo, 51  
 Ranjith, V., 451  
 Re, Angela, 979  
 Rizanov, Stefan, 279  
 Roberto, Garcia-Velez, 971, 1047  
 Rocamora, Maria Theresa Joy G., 745  
 Rochmawati, Erna, 993  
 Rodríguez, Nancy, 439  
 Rodríguez, Ramon L., 61  
 Romo-Fuentes, María Fernanda, 499  
 Ruiz, Christian Guzman, 129  
 Rzayeva, Ulviyya, 321

## S

Sabani, Maria E., 429

Saddi, Khim Cathleen M., 745  
 Sakulin, S. A., 269  
 Salkovska, Jelena, 785  
 Samsuryadi, 819  
 Sanderson, Dominic, 1105  
 Sandro, Gonzalez-Gonzalez, 971  
 Santaniello, Domenico, 1075  
 Santiago, Paul Ryan A., 745  
 Sapphayalak, Duangbhorn, 613  
 Sario, Alvin A., 61  
 Sarma, Amarendra K., 581  
 Sato, Ikuma, 349  
 Savvas, Ilias K., 429  
 Saxena, Aman, 309  
 Serpa-Andrade, Luis, 867  
 Serradell, Kim, 129  
 Serrano, Elcid A., 61  
 Shadadi, Ebtesam, 231  
 Shahbaz, Muzammil, 171  
 Shalaby, Raafat, 189  
 Sharma, Pawan, 509  
 Sharma, Raghav, 799  
 Shimakawa, Hiromitsu, 629  
 Shoukralla, E. S., 87  
 Shparaga, Tetiana, 329  
 Siguenza-Guzman, Lorena, 843  
 Sison, Marvin G., 591  
 Sonali, C. S., 695  
 Soria, Carlos, 73  
 Sorio, Claudio, 979  
 Špitálová, Zuzana, 833  
 Sreevastav, Mudigonda, 451  
 Strahonja, Vjeran, 205  
 Suarez, Jane Kristine G., 557  
 Sudsaward, Supanita, 613  
 Suksen, Amornphong, 293  
 Sumaili, Aisha, 231  
 Suma, K. V., 695

## T

Tai, Hung Nguyen, 417  
 Tangonan, Gregory L., 745  
 Tianxiao, Chen, 215  
 Tkachenko, Tetiana, 329  
 Todorov, Dimitar, 279  
 Todorović, Vladimir, 477  
 Tomala, Viviana, 895  
 Topol, Anna W., 97  
 Toprani, Dhvani, 655  
 Torres, Bryan James V., 591

Tran, Binh, 961  
 Troiano, Alfredo, 1075  
 Tsaousis, Theodosios, 467  
 Tyrkalo, Yuriy, 339

## U

Umutoni, Ritha M., 405  
 Usmani, Usman Ahmad, 1085

## V

Valentino, Carmine, 1075  
 Verdugo-Ormaza, Diego, 867  
 Vergara, Jaime, 705  
 Vieth, Julian, 715  
 Vinh, Bui Trong, 255  
 Vishwanath, Sista Kasi, 451  
 Vranić, Valentino, 477

## W

Watada, Junzo, 1085  
 Wedhasmara, Ari, 819  
 Wei, Zhang, 215  
 Widodo, Bambang Eka Cahya, 1035

## X

Xu, Qian, 655

## Y

Yakimov, Peter, 279  
 Yamanaka, Ayaka, 349  
 Yao, Huang, 215  
 Yeun, Chan Yeob, 11  
 Yoo, Paul D., 11  
 Yuasa, Tomoya, 629  
 Yuzevych, Volodymyr, 339

## Z

Zavvos, Stathis, 467  
 Zeini, Ammar, 763  
 Zeng, Miaoting, 939  
 Zhang, Lei, 939  
 Zhang, Sunyi, 683  
 Zhu, Lin, 683  
 Zimmermann, Robert, 641  
 Zuoping, Zhu, 215