Lecture Notes in Networks and Systems 1440

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

Proceedings of Tenth International Congress on Information and Communication Technology

ICICT 2025, London, Volume 1

OPEN ACCESS

Springer

Lecture Notes in Networks and Systems

Volume 1440

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, EI Compendex, 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).

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

Proceedings of Tenth International Congress on Information and Communication Technology

ICICT 2025, London, Volume 1

Springer

Editors
Xin-She Yang
Middlesex University
London, UK

Nilanjan Dey
Techno International New Town
Kolkata, West Bengal, India

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

Amit Joshi
Global Knowledge Research Foundation
Ahmedabad, Gujarat, India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-96-9708-3 ISBN 978-981-96-9709-0 (eBook)
https://doi.org/10.1007/978-981-96-9709-0

This work was supported by G R SCHOLASTIC LLP.

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 Tenth International Congress on Information and Communication Technology will be held during 18–21 February 2025 in a hybrid mode, Physical at London, UK and Digital Platform: Zoom. ICICT 2025 organised by Global Knowledge Research Foundation and Managed by G R Scholastic LLP. The associated partners were Springer and Springer Nature. The conference will provide a useful and wide platform both for display of the latest research and for 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 2200 papers were received for this conference from across 115 countries among which around 433 papers were accepted and will be presented physically at London and digital platform Zoom during the four days. Due to the overwhelming response, we had to drop many papers in the hierarchy of the quality. Total 65 technical sessions will be organised 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 ten volumes of proceedings by Springer LNNS Series. Over the years, this congress has been organised and conceptualised with 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 acknowledgements are extended to the team of Global Knowledge Research Foundation for their valuable efforts and support.

I look forward to welcoming you to the 11th Edition of this ICICT Congress 2026.

London, UK — Xin-She Yang
England, UK — R. Simon Sherratt
Kolkata, India — Nilanjan Dey
Ahmedabad, India — Amit Joshi

Contents

About the Editors

Xin-She Yang obtained his D.Phil. in Applied Mathematics from the University of Oxford. He then worked at Cambridge University and then later at the National Physical Laboratory (UK) as Senior Research Scientist. Now he is Reader at Middlesex University London, and Co-editor of the *Springer Tracts in Nature-Inspired Computing*. He is also an elected Fellow of the Institute of Mathematics and its Applications (FIMA), UK. He was the IEEE Computational Intelligence Society (CIS) chair for the Task Force on Business Intelligence and Knowledge Management (2015–2020). He has published more than 50 books and more than 400 peer-reviewed research papers with more than 90,000 citations, and he has been on the prestigious list of most influential researchers or highly cited researchers (Web of Sciences) every year since 2016.

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. Professor Sherratt received the 1st place IEEE Chester Sall Memorial Award in 2006, the 2nd place in 2016, and the 3rd place in 2017.

Nilanjan Dey is an associate professor in the Department of Computer Science and Engineering, Techno International New Town, Kolkata, India. He is a visiting fellow of the University of Reading, UK. He also holds a position of Adjunct Professor at Ton Duc Thang University, Ho Chi Minh City, Vietnam. Previously, he held an honorary position of Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012–2015). He was awarded his Ph.D. from Jadavpur University in 2015. He is Editor-in-Chief of the *International Journal of Ambient Computing and Intelligence*, IGI Global, USA. He is Series Co-editor of *Springer Tracts in Nature-Inspired Computing* (Springer Nature), *Data-Intensive Research* (Springer Nature), and *Advances in Ubiquitous Sensing Applications for Healthcare* (Elsevier). He is an associate editor of *IET Image Processing* and an editorial board member of *Complex and Intelligent Systems*, Springer Nature, *Applied Soft Computing*, Elsevier, etc. He

is having 35 authored books and over 300 publications in medical imaging, machine learning, computer-aided diagnosis, data mining, etc. He is Fellow of IETE and Senior Member of IEEE.

Amit Joshi is currently serving as the director of the esteemed Global Knowledge Research Foundation, India, this distinguished individual is an accomplished entrepreneur and researcher. His academic journey includes obtaining a B.Tech. degree in Information Technology, an M.Tech., in Computer Science and Engineering, and a Ph.D. focusing on the intricate fields of Cloud Computing and Cryptography in Medical Imaging, experiencing his vision with his rich experience spanning approximately 15 years and realising the need of globalisation in education and business. His present-day interests are primarily oriented towards the critical examination of government strategies and global forum requirements across education and business sectors. His active affiliations include esteemed professional societies such as ACM, IEEE, CSI, AMIE, IACSIT-Singapore, IDES, ACEEE, NPA, and more. His past responsibilities have included chairing the Computer Society of India (CSI) Udaipur Chapter and serving as Secretary and Chairman for the Association of Computing Machinery (ACM) Udaipur Professional Chapter. He was also the International Young ICT Chair for International Federation for Information Processing (IFIP), Austria formed through UNESCO in 1960. He has presented and authored over 50 papers in reputable national and international journals and conferences, specifically those organised by IEEE and ACM. His editorial contributions include editing over 100 books published by renowned publishers such as Springer, T&F, and ACM. He has also organised over 100 national and international conference delegations and workshops in over 20 countries including the USA, Canada, major of Europe, Southeast Asia through various societies, and international organisations. Apart from his academic pursuits, he is also actively involved in the global business and industrial community. He serves as the director of the Knowledge Chamber of Commerce and Industry, where he concentrates on establishing effective relationships among bureaucrats, industry associations, academic leaders, and regulatory authorities to address common research-related issues across sectors. He has organised over 50 industry forum events, facilitating communication with state and federal establishments, as well as corporate and academic bodies, to promote collaboration between industry and government sectors. Also, he has attended various high-profile events at UNESCO, International Telecommunication Union (ITU), Geneva, and with various embassies, consulates, and federal governments. Also, one of his primary focuses also is on building academic collaborations and promoting the 'ZAPAL' initiative, a future-oriented project involving abroad education powered by his profit venture, G R Scholastic LLP. Overall, he is a forward-thinking individual who is dedicated to creating networks across various sectors, particularly in education and industry.

Optimizing V2X Communications for 6G: A Summary of Techniques and AI Methods

Ali Belgacem and Abbas Bradai

Abstract This summary research paper provides a comprehensive overview of Vehicle-to-Everything (V2X) communications, including various communication types and the roles of base stations. It covers resource allocation techniques and beamforming for high-quality connectivity and addresses energy efficiency optimization metrics. The paper also discusses artificial intelligence methods and their integration to optimize these systems and enhance performance. This research serves as a valuable guide for those aiming to contribute to advancements in 6G technologies for efficient vehicular communications.

Keywords 6G · Vehicle to everything (V2X) · Artificial intelligence (AI) · Resource allocation · Beamforming · Base station · Energy efficiency

1 Introduction

6G is significantly ahead of 5G by using the terahertz, millimeter wave and sub-6GHz bands, along with satellite integration and access to dynamic spectrum for improved connectivity. By focusing on artificial intelligence, 6G technology aims to improve network operations, decision-making, and user experiences while saving more energy [1]. However, the transition to 6G may increase communications complexities and power requirements, especially regarding continuous connectivity in V2X cellular technology, requiring efficient power management, especially with the advent of electric vehicles. As such, key technologies such as Orthogonal Frequency Division Multiple Access (OFDMA) and the 3rd Generation Partnership

A. Belgacem (✉)
XLIM Research Institute, Poitiers University, Poitiers, France
e-mail: ali.belgacem@univ-poitiers.fr

A. Bradai
LEAT Lab, Cote d'Azur University, Nice, France
e-mail: abbas.bradai@univ-cotedazur.fr

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_1

Project (3GPP) will enhance global interoperability and integrate new innovations, propelling 6G networks toward superior performance and capabilities [2].

The integration of resource allocation and AI-based methods is essential for enhancing the efficiency and effectiveness of V2X communication systems. By optimizing resource utilization and energy consumption, AI-driven approaches can significantly improve the reliability, scalability, and sustainability of V2X networks, paving the way for safer and more efficient transportation systems of the future. On the other hand, joint beamforming involves coordinating beamforming techniques for transmission and reception among multiple vehicles and base stations. This leads to enhanced reliability, coverage, and throughput while minimizing interference and energy consumption, thereby contributing to the overall effectiveness of V2X communication networks [3].

Existing surveys in V2X communication have explored resource allocation [4], beamforming [5], and energy consumption [6]. However, these topics are often studied separately, lacking research on their complex relationships. Additionally, recent advancements in deep reinforcement learning and other methodologies are not adequately represented in existing works, hindering a comprehensive understanding of energy efficiency optimization in V2X systems. To address this gap, more rigorous analysis and consideration of real-world scenarios are needed. The main contribution of this research is to provide a comprehensive summary by combining three research trends in the context of V2X: artificial intelligence, resource allocation, and beamforming, with a focus on efficient base station energy consumption. By synthesizing current research, identifying challenges, and proposing future directions, this paper serves as a valuable resource for researchers and practitioners in V2X communication.

The remainder of the paper is organized as follows: Sect. 2 covers the basics of V2X communication. Sections 3, 4, and 5 provide a literature summary on resource allocation, beamforming, and energy efficiency, respectively. Section 6 discusses AI method applications in V2X networks. Following this, Sect. 7 presents challenges and future directions of the studied field. Finally, Sect. 8 concludes the paper.

2 The Basics of V2X Communication

Overall, V2X connectivity holds great potential to enhance traffic flow and enable innovative transportation services in both urban and rural areas (Fig. 1). Here are the basics of V2X communication.

2.1 V2X Communication Types

V2X communication projects aim to provide intelligent vehicle communication, covering various aspects such as Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure

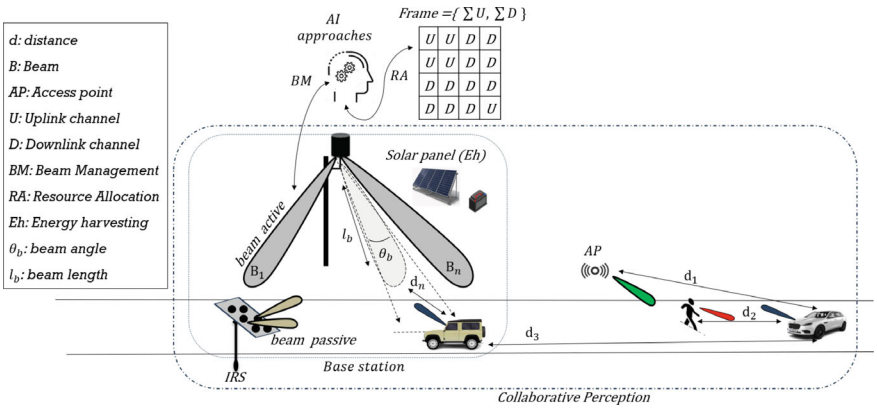


Fig. 1 V2X Resource allocation and beamforming with AI

(V2I), Vehicle-to-Pedestrian (V2P), Vehicle-to-Device (V2D), Vehicle-to-Grid (V2G), and Vehicle-to-Network (V2N). This communication ensures safety in autonomous driving by enabling vehicles to be aware of their surroundings and other road users through Collaborative Perception (CP) technology, which allows interaction between multiple entities near vehicles. With the upcoming 6G technology, V2X communication relies on Dedicated Short Range Communications (DSRC) and cellular networks, particularly utilizing mmWave channel propagation [3]. Therefore, cellular V2X (C-V2X) emerges as a crucial type within this context [7].

2.2 Base Station Functions

In V2X networks, the Medium Access Control (MAC) layer of base stations performs critical functions and requires specific power requirements [8]. Here's an overview of the main of them:

- **Manage channel access**: Modern 6G base stations are expected to use orthogonal frequency division multiplexing (OFDM) as a critical technology to access V2X networks. This enables high data rates, reliable communications, and efficient use of the radio spectrum [9].
- **Frame Structure Management**: The base station defines the structure of communication frames, incorporating synchronization headers, data payload, error checking, and acknowledgment mechanisms to ensure that transmitted data is organized into coherent frames for efficient transmission and reception [10].
- **Transmit and receive signals**: The base station contributes to energy consumption through the transmission and reception of radio signals, crucial for facilitating communication between vehicles [8]. Transmitting data frames at higher power levels and receiving acknowledgment signals are particularly energy-intensive, especially in scenarios with dense traffic or extended communication ranges.

3 Resource Allocation Techniques

In resource allocation for V2X networks, three primary resources are crucial: power, channel, and spectrum resources. They are mentioned in [4, 11]. Resources are optimized in various ways, among them:

Dynamic Resource Allocation Adaptive power control, spectrum management, and Time Division Multiple Access (TDMA) are crucial techniques for optimizing energy efficiency and communication reliability. Base stations dynamically adjust transmission power levels based on distance and signal quality, conserving energy by reducing power for nearby vehicles while maintaining reliable communication [12]. Spectrum management ensures efficient allocation of frequency bands, minimizing interference and maximizing spectrum utilization through techniques like spectrum sensing and cognitive radio.

Traffic-Aware Scheduling Base stations implement several strategies to enhance energy efficiency and prioritize safety-critical traffic. Firstly, prioritization of safety-critical messages ensures timely delivery of collision warnings and emergency alerts over non-critical traffic, reducing the need for retransmissions and conserving energy. Additionally, dynamic bandwidth allocation allows base stations to adaptively allocate bandwidth to V2X applications based on their quality of service requirements and traffic loads [13]. During low-traffic periods, base stations can allocate unused bandwidth to energy-efficient modes or reduce the number of active transmission antennas, further optimizing energy consumption within the network.

Cooperative Communication In V2X systems, base stations employ relay-based and group-based communication techniques to enhance energy efficiency and extend communication capabilities. Relay-based communication involves nearby vehicles acting as relay nodes, extending the communication range and reducing the transmission power required for direct communication between distant vehicles. By strategically leveraging relay nodes, base stations can achieve significant energy savings while maintaining reliable communication. Additionally, group-based communication organizes V2X devices into communication groups based on proximity or traffic patterns. This approach enables base stations to transmit data simultaneously to multiple vehicles within the same group, reducing overall energy consumption compared to individual point-to-point transmissions and improving network efficiency [7].

Joint resource allocation and beamforming optimization Intelligent Reflecting Surface (IRS) technology significantly enhances spectral and energy efficiency in Vehicle-to-Everything (V2X) communication by improving beamforming and resource allocation. IRS optimizes beamforming by dynamically adjusting reflection coefficients, leading to higher signal gains. Integrating IRS into resource allocation involves optimizing and managing channel access, resulting in improved overall performance [14].

4 Beamforming

Beamforming is essential in modern V2X communication for enhancing performance and efficiency by enabling directional signal transmission and reception. Several beam types have been addressed in the literature, as mentioned below:

- **Narrow beam**: It refers to the use of directional antennas to focus the transmission and reception of signals into a narrow, targeted beam rather than broadcasting signals in all directions, as detailed in paper [15].
- **Passive beam**: Passive beamforming involves the design and arrangement of antennas to naturally direct signals in specific directions. This is achieved using fixed antenna arrays, reflectors, and lens antennas [2]. The physical shape and orientation of the antenna elements cause the signals to combine constructively in targeted directions, enhancing signal strength without the need for electronic adjustments.
- **Active beam**: This type uses electronically controlled antennas to dynamically direct signal beams, improving connectivity, reducing interference, and improving signal quality [8]. The researchers in [2] suggested a combination of active and passive beamforming techniques to significantly enhance communication throughput and reliability. Active beamforming dynamically steers the signal to adapt to changing conditions, while passive beamforming optimizes signal directionality through fixed antenna configurations.

Effective energy consumption requires different management strategies for various types of beams, as illustrated below:

- **Beam selection**: Dynamic beamforming strategies can achieve reliable communication while minimizing energy consumption, but the selection process must balance connectivity needs with energy demands [16]. Energy-efficient algorithms aim to maximize communication quality while minimizing energy expended on beamforming, enhancing V2X network sustainability and effectiveness. Factors like vehicle density, mobility patterns, channel conditions, and traffic dynamics influence beam selection.
- **Beam tracking**: Beam tracking is the dynamic adjustment of wireless communication beams between vehicles and surrounding infrastructure to optimize signal strength and minimize energy consumption. Adjusting beamforming parameters in response to changing environmental conditions, reduces unnecessary energy usage associated with maintaining connections. However, it is essential to take into account the channel state information (CSI) for effective beam tracking [5].
- **Beam alignment**: It refers to the process of precisely aligning directional antennas between vehicles and roadside infrastructure to establish and maintain reliable communication links while minimizing energy usage. The alignment of the beam is influenced by factors such as the size of the beamwidth and the quality of the channel state information [8].
- **Beam recovery**: It refers to the process of regaining or reestablishing a communication beam after a disruption or interruption. This method of beam management

enables the detection of potential link failures and the identification of alternative beam pairs for mmWave communication. Thus, it is necessary to minimize the impact of blockage and the duration of link failure [17].

- **Beam switching**: It involves transitioning between different communication beams to sustain connectivity and optimize communication performance, all while minimizing unnecessary energy consumption. This process utilizes predictions of a vehicle's next position to facilitate smooth transitions to upcoming beams. Notably, larger beam widths result in fewer instances of beam switching [8].
- **Beam training**: It refers to the process of optimizing beamforming parameters (beam direction, width, and power) to establish efficient communication links between vehicles and base stations [5, 8].

5 Energy Efficiency in V2X Communications

The following optimization metrics ensure effective energy use, often resulting in reduced energy consumption. They include:

- **Interference**: Interference optimization for energy efficiency for V2X networks can be achieved in various ways: Interference prediction, interference minimization, multiple access interference, and interference mitigation, as shown in [18], respectively. Generally, interference improvement relies on maximizing the signal-to-noise ratio (SNR).
- **Offloading**: It involves shifting computational and data processing tasks from vehicles to external infrastructure, such as roadside units or cloud servers. This approach can reduce the computational burden on vehicles, enhance network resource utilization, and employ intelligent strategies for dynamic task allocation and efficient data management [19].
- **Throughput**: Throughput measures the data transmitted between source and destination within a timeframe, often in bits per second (bps). Effective throughput accounts for overhead, retransmissions, and protocol inefficiencies, representing the actual data rate achieved. For example, reference [20] proposes maximizing total throughput in NR-V2X networks across subcarriers while considering available power and minimum transmission rate constraints.
- **Energy Harvesting Efficiency**: Energy harvesting involves converting ambient energy sources like solar, kinetic, or electromagnetic energy into usable electrical power to support V2X communication systems. Improving energy harvesting efficiency is crucial for extending the operational lifespan of V2X devices, reducing dependence on traditional power sources, and promoting sustainability. In reference [21], researchers utilized energy harvesting to achieve long-term energy efficiency maximization by employing power splitting to delicately divide the harvested energy.

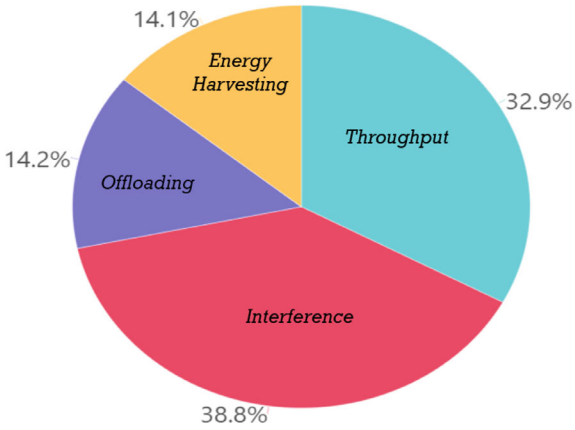


Fig. 2 Percentage of energy efficiency metric optimization usage

Fig. 2 shows the percentage use of the different improvement metrics mentioned above from 2020 to mid-2024 (according to Google Scholar).

6 Utilized AI Approaches

Most investigations into the issue studied in this paper classify it as a non-convex problem [22]. Thus, by employing the AI approaches, V2X networks can achieve more efficient and intelligent beam management and resource allocation, leading to improved communication performance and energy efficiency. In recent years, researchers have begun applying various types of AI methods to this research trend [23]. According to our statistics from Google Scholar for the period from 2020 to mid-2024 (Fig. 3), it is notable that deep reinforcement learning is used by a significant portion of the scientific research community, and advanced reinforcement learning techniques are emerging as new propositions. These are promising solutions. Therefore, we focused our classification on these latter types.

6.1 Deep Reinforcement Learning (RL)

The most commonly utilized methods in the literature are as follows:

Deep Q-Learning (DQL) It combines the strengths of reinforcement learning with the powerful symbolic capabilities of deep neural networks, enabling efficient handling of high-dimensional state spaces, learning directly from initial inputs, generalization across states, scaling to complex environments, improving policy learning, dealing with partial observability, and integration with other deep learning techniques [24, 25]. In [26], Deep Q-Learning optimizes joint beamforming by effectively handling complexity and adapting to changing conditions. This ensures optimal

performance and can effectively mitigate blocking effects. DQL maximizes the total communication rate of target vehicles while also ensuring the service quality of each target vehicle.

Multi-Agent RL (MARL) It possesses the capability to capture complex interactions, emergent coordination, division of labor, adaptability to dynamic environments, learning from others, robustness to failures, privacy preservation, and real-world applicability. For example, in [27], it is utilized in various V2I and V2V scenarios within C-V2X cellular communications to improve the joint optimization of the spectrum and power allocation.

Proximal Policy Optimization (PPO) PPO's stability in training neural network policies makes it advantageous for optimizing beamforming strategies in wireless communication systems. In the study [22], the authors utilize the PPO algorithm to obtain the optimal solution to the formulated problem in the context of RIS-assisted 6G-V2X communication networks. By iteratively adjusting policy parameters and leveraging the clipping surrogate method, the algorithm effectively improves network performance.

Multi-armed bandits (MAB) Within beamforming systems, MAB algorithms dynamically select optimal beams or combinations, considering channel conditions, traffic demands, and interference levels, thus improving communication performance and spectral efficiency. The methodology presented in reference [28] employs Contextual Bandit to manage interference while allocating mmWave beams for vehicle service in the network. By leveraging neighboring beam status knowledge, it identifies and avoids potential interfering transmissions, ensuring minimal interference even under heavy traffic loads.

6.2 Advanced RL Techniques

The utilization of advanced RL techniques is still in its early stages in this studied field, with the widespread adoption of federated methods being prominent among various approaches.

Federated Learning (FL) substantially benefits V2X communication and base stations within wireless networks. It allows for distributed optimization and resilience to network failures in base stations, improving scalability, strength, and resource access. In V2X communication, FL supports localized learning and collaborative decision-making among vehicles, improving communication reliability and efficiency. Additionally, FL reduces energy consumption through energy-efficient training and adaptive model updates, promoting sustainability in wireless network operations. The article [29] presents a task offloading and resource allocation algorithm for connected and autonomous vehicle networks, using Federated Reinforcement Learning (FRL). This method, within a cooperation architecture among vehicles, roads, and base stations, aims to reduce task execution delays under different constraints. It

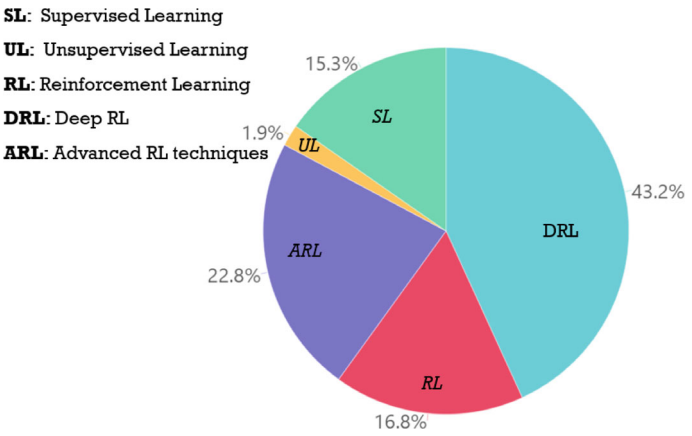


Fig. 3 Percentage of primary AI methods used for energy efficiency

boosts system throughput by optimizing task offloading and resource allocation, cutting down on data transmission delays and communication overhead, and adapting to changing network setups.

6.3 Integrating AI Approaches in V2X Networks

Implementing Deep Reinforcement Learning involves five main steps, as follows:

Data Collection and Preprocessing AI models require extensive data from V2X networks, including vehicle mobility patterns, channel conditions, and network usage statistics. Effective data collection and preprocessing are crucial for accurate model training [28, 29].

Making of the Markov Decision Process (MDP) It is one of the most fundamental components in a reinforcement learning (RL) problem. Typically in V2X scenarios, the base station or vehicle is considered the agent, which is the learner interacting with the environment by taking different actions in various states [25].

Deep learning Network Model This refers to the network architecture (layers, nodes, pipelines) adopted to construct our deep AI model. Various architectures are utilized, including Neural Networks (NNs), Support Vector Machines (SVMs), and others [24, 25].

Model Training and Deployment Trained models can be deployed at base stations for real-time decision-making. Continuous learning and model updating are essential to adapt to evolving network conditions and vehicular environments.

Performance Evaluation The effectiveness of AI models should be evaluated using metrics such as latency, throughput, energy efficiency, and overall network performance. Real-world trials and simulations help in fine-tuning the models for optimal performance.

6.4 Difficulty in Integrating AI Approaches

Integrating AI into V2X-6G systems presents several challenges. One key issue is communication overhead, as frequent updates between vehicles and base stations demand low-latency connections and efficient handling of network congestion and mobility. The diversity of vehicle data complicates model training, while privacy and security risks, such as model poisoning, add further complexity. High-speed environments impose resource constraints that hinder efficient model training. As the vehicle count increases, scalability becomes a problem, complicating update management and model convergence. Compatibility with legacy systems, growing computational demands, and real-time decision-making are also critical challenges.

7 Challenges and Future Directions

Some key research areas poised to expand the capacity of wireless access for V2X communications include advanced modulation and massive multiple access techniques such as mmWave frequency bands, massive MIMO communications, Orthogonal Time-Frequency Space (OTFS) modulation, and Non-Orthogonal Multiple Access (NOMA) [3, 30]. Additionally, computation and network management at scale through mobile edge computing (MEC) and vehicular cloud and fog computing are critical areas of focus [30].

Collaborative efforts and the integration of emerging technologies are essential to proposing modern solutions that address current challenges such as mobility management, interference management, channel estimation, and efficient network slicing [4]. In terms of AI opportunities, the availability and training of datasets, as well as the consideration of performance parameters and computational complexity, are crucial for creating effective and credible models [31]. These advancements will include the allocation of V2X network resources and beam management for efficient power transfer. They will also involve exploring energy harvesting and renewable energy integration to support sustainable V2X ecosystems. Research should focus on developing adaptive algorithms to manage dynamic changes in V2X environments, exploring hybrid approaches that combine AI and traditional methods, and investigating the scalability of these solutions in larger, more complex networks for practical deployment.

On the other hand, while effective in simulations, the proposed methods face challenges in real-world V2X environments, including dynamic traffic, interference,

hardware limitations, latency, and network congestion. Future research should prioritize field tests and pilot studies to evaluate the practical viability of these methods in live V2X scenarios

8 Conclusion

In this paper, we addressed the current state-of-the-art in improving resource allocation and beamforming in 6G-V2X networks, leveraging artificial intelligence techniques to enhance energy efficiency in base stations. We provided a summary of important research in this direction. The integration of AI to improve resource allocation and beamforming in 6G-V2X networks shows great potential for creating energy-efficient base stations. This combination not only enhances vehicular communication performance but also aligns with the broader goal of developing sustainable and efficient next-generation networks. To the best of our knowledge, this is the first study examining the combination of resource allocation, beamforming, and artificial intelligence in V2X networks.

Acknowledgements This research is conducted within the framework of the SAMBAS project (https://sambas-project.com/).

References

1. da Silva Brilhante D, Manjarres JC, Moreira R, de Oliveira Veiga L, de Rezende JF, Müller F, Klautau A, Mendes LL, de Figueiredo FAP (2023) A literature survey on ai-aided beamforming and beam management for 5g and 6g systems. Sensors 23(9):4359
2. Huang Z, Zheng B, Zhang R (2023) Roadside IRS-aided vehicular communication: efficient channel estimation and low-complexity beamforming design. IEEE Trans Wirel Commun
3. Hisabo DS, Olwal TO, Hassen MR (2024) Channel modeling techniques for multiband vehicle to everything communications: challenges and opportunities. Int J Comput Digit Syst 15(1):29–50
4. Nair A, Tanwar S (2024) Resource allocation in v2x communication: state-of-the-art and research challenges. Phys Commun, 102351
5. Yi W, Zhiqing W, Zhiyong F (2024) Beam training and tracking in mmWave communication: a survey. China Commun
6. Sohaib RM, Onireti O, Sambo Y, Swash R, Imran M (2024) Energy efficient resource allocation framework based on dynamic meta-transfer learning for v2x communications. IEEE Trans Netw Serv Manag
7. Abboud K, Omar HA, Zhuang W (2016) Interworking of DSRC and cellular network technologies for v2x communications: a survey. IEEE Trans Veh Technol 65(12):9457–9470
8. Ghafoor KZ, Kong L, Zeadally S, Sadiq AS, Epiphaniou G, Hammoudeh M, Bashir AK, Mumtaz S (2020) Millimeter-wave communication for internet of vehicles: status, challenges, and perspectives. IEEE Internet Things J 7(9):8525–8546
9. Hua Q, Keping Yu, Wen Z, Sato T (2019) A novel base-station selection strategy for cellular vehicle-to-everything (c-v2x) communications. Appl Sci 9(3):556

10. MacHardy Z, Khan A, Obana K, Iwashina S (2018) V2x access technologies: regulation, research, and remaining challenges. IEEE Commun Surv Tutor 20(3):1858–1877
11. Zhang E, Yin S, Ma H (2019) Stackelberg game-based power allocation for v2x communications. Sensors 20(1):58
12. Guo Q, Tang F, Kato N (2023) Resource allocation for aerial assisted digital twin edge mobile network. IEEE J Sel Areas Commun
13. Le TTT, Moh S (2021) Comprehensive survey of radio resource allocation schemes for 5g v2x communications. IEEE Access 9:123117–123133
14. Yang Y, Zhang S, Zhang R (2020) IRS-enhanced OFDMA: joint resource allocation and passive beamforming optimization. IEEE Wirel Commun Lett 9(6):760–764
15. Zhou T, Li C, Zhang W, Ai B, Liu L, Liang Y (2024) Narrow-beam channel measurements and characterization in vehicle-to-infrastructure scenarios for 5g-v2x communications. IEEE Internet Things J
16. Ahmed I, Shahid MK, Khammari H, Masud M (2021) Machine learning based beam selection with low complexity hybrid beamforming design for 5g massive MIMO systems. IEEE Trans Green Commun Netw 5(4):2160–2173
17. Shimizu T, Va V, Bansal G, Heath RW (2018) Millimeter wave v2x communications: use cases and design considerations of beam management. In: 2018 Asia-Pacific microwave conference (APMC). IEEE, pp 183–185
18. Rehman A, Valentini R, Cinque E, Marco P, Santucci F (2023) On the impact of multiple access interference in LTE-v2x and nr-v2x sidelink communications. Sensors 23(10):4901
19. Prathiba SB, Raja G, Anbalagan S, Dev K, Gurumoorthy S, Sankaran AP (2021) Federated learning empowered computation offloading and resource management in 6g-v2x. IEEE Trans Netw Sci Eng 9(5):3234–3243
20. Xiaoqin S, Juanjuan M, Lei L, Tianchen Z (2020) Maximum-throughput sidelink resource allocation for nr-v2x networks with the energy-efficient CSI transmission. IEEE Access 8:73164–73172
21. Song Y, Xiao Y, Chen Y, Li G, Liu J (2022) Deep reinforcement learning enabled energy-efficient resource allocation in energy harvesting aided v2x communication. In: 2022 IEEE 33rd Annual international symposium on personal, indoor and mobile radio communications (PIMRC). IEEE, pp 313–319
22. Saikia P, Pala S, Singh K, Singh SK, Huang W-J (2023) Proximal policy optimization for RIS-assisted full duplex 6g-v2x communications. IEEE Trans Intell Veh
23. Christopoulou M, Barmpounakis S, Koumaras H, Kaloxylos A (2023) Artificial intelligence and machine learning as key enablers for v2x communications: a comprehensive survey. Veh Commun 39:100569
24. Almutairi MS (2022) Deep learning-based solutions for 5g network and 5g-enabled internet of vehicles: advances, meta-data analysis, and future direction. Math Probl Eng, 1–27
25. Tang F, Zhou Y, Kato N (2020) Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5g HetNet. IEEE J Sel Areas Commun 38(12):2773–2782
26. Ying J, Wang H, Chen Y, Zheng T-X, Pei Q, Yuan J, Al-Dhahir N (2023) Deep reinforcement learning based joint beam allocation and relay selection in mmWave vehicular networks. IEEE Trans Commun 71(4):1997–2012
27. Ding Y, Huang Y, Tang L, Qin X, Jia Z (2022) Resource allocation in v2x communications based on multi-agent reinforcement learning with attention mechanism. Mathematics 10(19):3415
28. Kose A, Lee H, Foh CH, Shojafar M (2024) Multi-agent context learning strategy for interference-aware beam allocation in mmWave vehicular communications. IEEE Trans Intell Transp Syst
29. Guo Q, Tang F, Kato N (2022) Federated reinforcement learning-based resource allocation in d2d-enabled 6g. IEEE Network
30. Clancy J, Mullins D, Deegan B, Horgan J, Ward E, Eising C, Denny P, Jones E, Glavin M (2024) Wireless access for v2x communications: research, challenges and opportunities. IEEE Commun Surv Tutor

31. Bhattacharya P, Bodkhe U, Zuhair M, Rashid M, Liu X, Verma A, Dewangan RK (2024) Amalgamation of blockchain and sixth-generation-envisioned responsive edge orchestration in future cellular vehicle-to-anything ecosystems: opportunities and challenges. Trans Emerg Telecommun Technol 35(4):e4410

Web-AR Base Support System for Food Tourism

Makoto Hirano and Kayoko Yamamoto

Abstract The purposes of tourism have been becoming diversified in recent years, and a form of travel known as food tourism is becoming increasingly popular. However, little research has been conducted on systems that support food tourism. Against such a backdrop, the present study aims to design, develop, operate and evaluate a food tourism support system that is supported to decide on restaurants for lunch and dinner, tourism spots to visit along the way, and routes to visit these destinations. The system comprises an original tourism plan creation system, web geographic information systems (Web-GIS) and web-augmented realty (Web-AR). In the present study, a location-based Web-AR system is developed. The system was operated for 30 days from December 22, 2023 to January 20, 2024, in Central Yokohama City of Kanagawa Prefecture, Japan. Total number of users was 50 and 20 tourism plans were created during the operation period. Based on the evaluation results, it is clear that the principal functions and the overall system were highly evaluated, regardless of food tourism experience or advance creation of tourism plan. Furthermore, it is evident that there was a high number of visits to the pages for most of the principal functions, and the system was used in a manner consistent with the purpose of the present study.

Keywords Tourism plan creation system · Web-geographic information systems (GIS) · Web-augmented realty (Web-AR) · Location-based augmented realty (Location-based AR) · Food tourism

M. Hirano · K. Yamamoto (✉)
The University of Electro-Communications, Tokyo, Japan
e-mail: kayoko.yamamoto@uec.ac.jp

M. Hirano
e-mail: h2230116@edu.cc.uec.ac.jp

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_2

1 Introduction

The purposes of tourism have been becoming diversified in recent years, and a form of travel known as food tourism is becoming increasingly popular. According to the Japan Tourism Agency [1], food tourism is defined as a form of travel, the purpose of which is to "enjoy food and food culture that are specific to a region," Distinctive gourmet foods exist in every region and are prepared using tourism resources, which gives them the advantage of not being very expensive. However, though research has been actively conducted on regional revitalization through food tourism, little research has been conducted on systems that support food tourism.

There is also a tourist demand to engage in tourism without creating plans in advance. According to the Japan Travel Bureau (JTB) Research Institute [2], an increasing number of people in recent years have been deciding on visiting places by obtaining various information in tourism areas. Therefore, gathering various forms of information for a tourism area, not only for creating tourism plans in advance, but also for when no plans are created, is an effective form of support for tourists. A survey conducted by the Jalan Research Center [3] indicated that only 42% of tourists decided on tourism routes and plans in advance. Thus, there is a need for a tourism system centered on food and drink that provides support for both tourists who create plans in advance as well as those who do not.

Meanwhile, various support methods exist for travelers in tourism areas, one of which is augmented reality (AR). AR superimposes and fuses "digital information" into the "real environment," through the screen of a mobile information device such as a smartphone or tablet device. Location-based AR comprises the use of the global positioning system (GPS) function implemented into a user's mobile information device; thus, it is expected to be used for gathering information in tourism areas or navigating them.

The present study aims to take into consideration the aforementioned context to design, develop, operate and evaluate a system that can accommodate both tourists who create food and drink-centered plans in advance as well as those who do not. The system comprises an original tourism plan creation system, Web-GIS and Web-AR, and implements two unique functions. The first is a function that efficiently creates tourism plans that are centered on restaurants where lunch and dinner are eaten. The second is a function that uses location-based AR to display information on nearby restaurants and tourism spots.

Central Yokohama City of Kanagawa Prefecture, Japan was selected as the operation area for the system. The first reason for this selection is that the city has a large number of diverse restaurants, and organizations such as "YOKOHAMA FOOD LOVERS" are actively attempting to increase the appeal of the food culture in this City. The second reason is that, according to the Yokohama City Tourism Consumption Trends Survey Overview by Yokohama City [4], the most common reason for tourism in this city was experiencing the local food and drink, which accounted for 34% of the total.

2 Related Work

Previous studies related to the present study can be divided into three groups: (1) studies on tourism plan creation support systems, (2) studies on tourism support systems using AR, and (3) studies on food tourism support systems. In the following, a representative review on previous studies of recent years in the three groups is presented, and the originality of the present study is demonstrated.

Regarding (1) studies on tourism plan creation support systems, Ribeiro et al. [5], Chen and Tsai [6] and Yazdeen et al. [7] developed systems to support tourism plan creation using the location information of tourist spots and users. Lim et al. [8], Hida et al. [9], Nitu et al. [10], Tavitiyaman et al. [11], Avval and Harounabadi [12], Li et al. [13], and Vranić et al. [14] developed systems to support tourism plan creation using information and images on social media. The systems in this group have the functions of creating tourism plans that reflect user preferences but do not take into consideration tourism centered on food and drink.

Regarding (2) studies on tourism support systems using AR, Sasaki and Yamamoto [15], Zhou et al. [16], Ferentinos et al. [17], Andrii et al. [18] and Otsu et al. [19] developed systems to provide tourists with sightseeing information using AR. Ikizawa-Naitou and Yamamoto [20], Sasaki and Yamamoto [21], Abe et al. [22] and Sonobe et al. [23] developed tourism support systems for tourist migration behavior using AR. The systems in this group have the functions of providing sightseeing information but were unable to provide navigation for users.

Regarding (3) studies on food tourism support systems, Miyoshi and Okuno [24], Nakano et al. [25], Horibe et al. [26], Kumagai and Okuno [27] and Yoshida et al. [28] developed systems that recommend suitable restaurants that serve food that meets users' tastes. Kumarasiri and Farook [29], Luo and Xu [30], Ichimura [31], Sarasa-Cabezuelo [32] developed systems that provide tourists with information on popular and appropriate restaurants based on the result of online reviews. In this group, since food tourism is a new form of tourism, few systems have been developed to support food tourism.

The originality of the present study involves the development of a unique system that overcomes the problems in the aforementioned previous studies and realizes the following three points; (1) The system facilitates the creation of tourism plans with the principal purpose of food and drink. (2) The system enables the use of all functions on the web without using a dedicated device or installing an application on an information device. (3) The system is comprised with Web-AR, and enables not only tourism plan creation in advance but also supports tourists' activities in the tourism area.

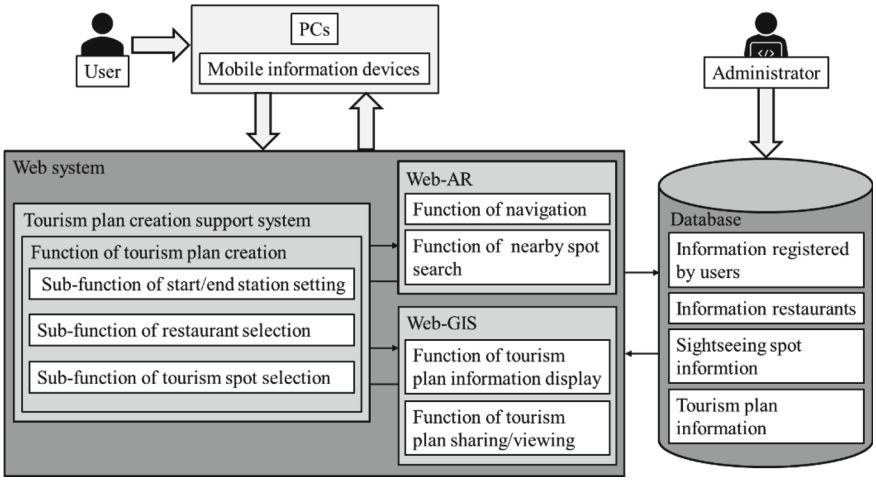


Fig. 1 System features

3 System Design

3.1 System Features

Figure 1 presents the system features. The system comprises an original tourism plan creation system, Web-GIS and Web-AR. When the user uses the system for the first time, they create an account and register user information in the database. The purpose of the system is to support the creation of tourism plans centered on food and drink in advance, as well as tourists' activities in the tourism area. In the system, restaurants can be selected, which is the principal objective, in addition to tourism spots at which tourists can stop along the way, and it can create a tourism route comprising these spots. Moreover, the system can search for restaurants and tourism spots near the user's current location, and display the information on the AR-based mobile information device screen to support the user's gathering of information in the tourism area. The system is operated as a web system with the aforementioned functions.

3.2 Design of Individual Systems

3.2.1 Tourism Plan Creation Support System

The burden on users in creating a tourism plan is reduced by integrating an original tourism plan creation support system for the system. The system implements the function of tourism plan creation, which involves the following three sub-functions:

the sub-function of start/end station setting, the sub-function of restaurant selection, and the sub-function of tourism spot selection.

3.2.2 Web-GIS

The ArcGIS API for JavaScript provided by the ESRI, Inc. is used as Web-GIS. This allows users to use the system on a web browser, perform route searches, and visualize information on a digital map without installing special software. The system implements the function of tourism plan information display as well as the function of tourism plan sharing/viewing.

3.2.3 Web-AR

A location-based Web-AR system is developed using A-frame and AR.js that were explained in Sect. 3.1. The system implements the function of nearby spot search and the function of navigation.

4 System Development

4.1 Frontend

4.1.1 Function of Tourism Plan Creation

(1) *Sub-function of start/end station setting*

After logging into the system, the user moves from the menu bar to the page for the sub-function of start/end station setting, and can select the stations where they will start and end their tour. The user can click or tap the select box at the top of this page or the icon on the digital map to select the stations where they will start and end their tour.

(2) *Sub-function of restaurant selection*

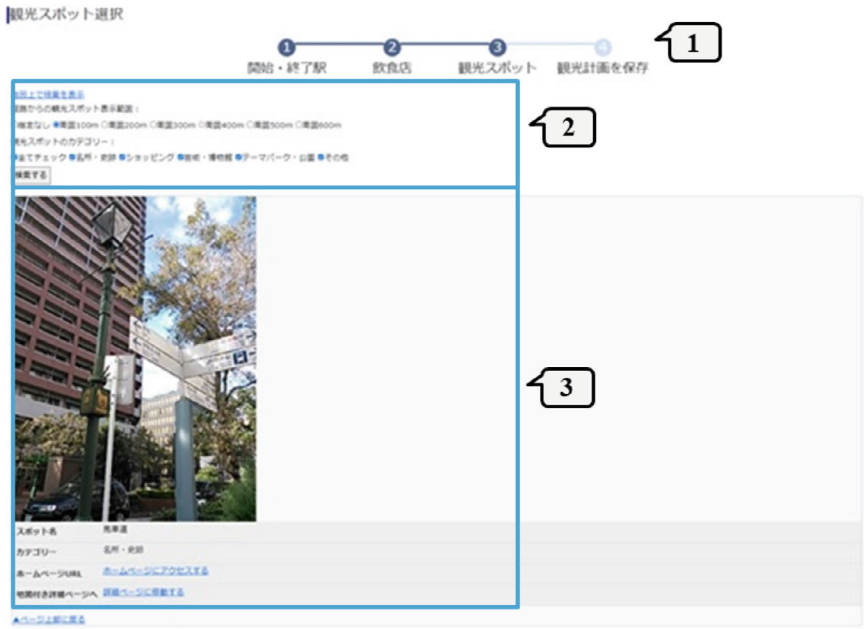
Figure 2 presents the pages for the sub-function of restaurant selection. Users can use this page to specify the conditions for selecting and setting restaurants to visit for lunch or dinner. The selection result can be displayed in a list or on a digital map. Users can move from the list display to a detailed page of the selected restaurants by clicking or tapping “Go to detailed page” or the icon on the digital map. On the detailed page, users can incorporate the selected restaurant into their tourism plan as a restaurant to visit for lunch or dinner, along with the time required for the meal.

(3) *Sub-function of tourism spot selection*

No.	Description
1	Procedure of tourism plan creation
2	Conditions for selecting and setting restaurants to visit for lunch or dinner (WIFI availability, credit card payment, non-smoking area, lunch budget, dinner budget, genre and search keyword)
3	Selected restaurant (Image, name, genre, opening hours, non-smoking area, lunch budget, dinner budget, non-smoking area, menu, website and location)

Fig. 2 Page for the sub-function of restaurant selection

Figure 3 presents the page for the sub-function of tourism spot selection. Users can use this page to specify the conditions for selecting tourism spots and set them as tourism spots to visit "before lunch," "after lunch," or "after dinner." Furthermore, tourism spots are classified into five categories: famous/historical sites, shopping, art/museums, theme parks/parks and others. The specified conditions are the distance from the selected restaurant and the tourism spot category. The selection result can



No.	Description
1	Procedure of tourism plan creation
2	Range from the user‘s current location (0-100m, 0-200m, 0-300m, 0-400m, 0-500m, 0-600m), and categories of tourism spots (Famous/historical sites, shopping, art/museums, theme parks/parks and others)
3	Selected tourism spot (Image, name, genre, category, website and location)

Fig. 3 Page for the sub-function of tourism spot selection

be displayed as a list or on the digital map. Users can move from the list display to a page showing the details of the selected restaurants by clicking or tapping “Go to detailed page” or the icon on the digital map. On the detailed page, users can incorporate the selected tourism spot into their tourism plan at their preferred time of day, along with the time required for the stay.

4.1.2 Function of Tourism Plan Information Display

Users can use the page of tourism plan information display to display and save information on the tourism plan they created. Users can display the tourism route on a digital map and calculate the total walking distance, total walking time, calories burned, and total tour time. Users can save their tourism plan by inputting the tourism plan name, any notes, and a public/private setting. In the system, the total walking time is calculated based on the assumption of movement at 80 m/min, according

to the standards of ArcGIS Pro provided by the ESRI, Inc. The total tour time is calculated by adding the total walking time and the time required at each spot.

4.1.3 Function of Tourism Plan Sharing/Viewing

Figure 4 presents the page for the function of tourism plan sharing/viewing. On this, users can view the tourism plans created by other users and its information. This page also allows them to copy the tourism plans of other users.

No.	Description
1	User-input tourism plan name
2	Name of user who created this tourism plan
3	User's notes of this tourism plan

Fig. 4 Page for the function of tourism plan sharing/viewing

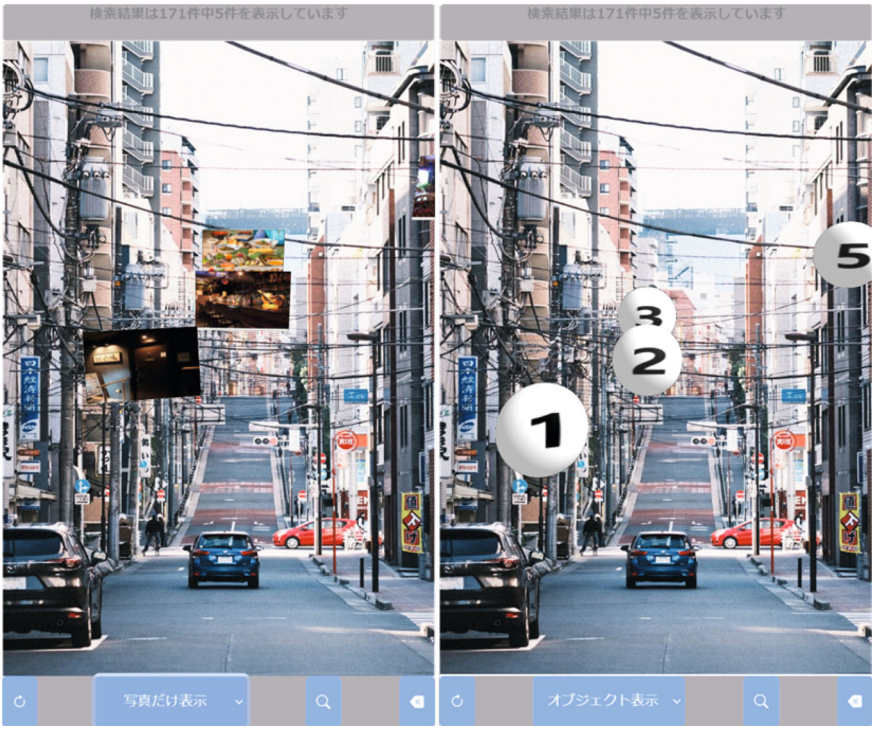


Fig. 5 Pages for the search result display using images and objects on the AR-based mobile information device screen

4.1.4 Function of Nearby Spot Search

The system allows users to search for restaurants and tourism spots near the user's current location. The specified conditions include the search range, sorting conditions, and number of items to display (maximum of 10 items). The search result can be displayed on an AR-based mobile information device screen or on a digital map. The sorting condition is the distance from the user's current location. Furthermore, the AR-based mobile information device screen display has the options of "normal display," which shows items such as the spot name, business hours and budget, "image display," which shows only the image of the spot, and "object display," which shows the visit destinations. Figure 5 presents the search result display using images and objects on the AR-based mobile information device screen.

4.1.5 Function of Navigation

Figure 6 presents the pages for the navigation and the destination information display (image, genre, open hours, lunch budge and dinner budge). As shown in Fig. 6,

Fig. 6 Pages for the navigation and the destination information display

the AR-based mobile information device screen allows users to receive navigation instructions from the current location to the next destination. Users can also use this screen to display objects that show the destination and the distance from the current location. Users can also confirm the destination information and usage methods from the bottom of the screen.

4.2 Backend

4.2.1 Processing Related to the Sub-function of Restaurant Selection

The restaurants saved in the database are each assigned an ID. Users can use the sub-function of restaurant selection to extract information on restaurants that match the specified conditions and include the search words from the database. Backend processing is then performed for the list display or the digital map display of the restaurant information.

4.2.2 Processing Related to the Sub-function of Tourism Spot Selection

As in the case of restaurants, the tourism spots that are stored in the database are each assigned an ID. The sub-function of tourism spot selection derives a route comprising the restaurants and tourism spots according to the created tourism plan. Information on the tourism spots near the route that match the specified conditions is subsequently extracted from the database. Backend processing is then performed for the list display or the digital map display of the tourism spot information.

4.2.3 Processing Related to the Function of Tourism Plan Information Display

The created tourism plan information is used as a basis for displaying the tourism route on the digital map, and the backend processing is then performed for calculating the total walking distance, total walking time, calories burned, and total time required for tourism. The backend processing is also performed for storing the user-input tourism plan name, any notes and public/private settings in the database.

4.2.4 Processing Related to the Function of Tourism Plan Sharing/ Viewing

Backend processing is performed for displaying the tourism plan stored in the database, and copying tourism plans created by other users.

4.2.5 Processing Related to the Function of Nearby Spot Search

Backend processing is performed for obtaining the user's current location from the GPS function developed in the mobile information device, and extracting data on restaurants and tourism spots that match the specified conditions from the database.

4.2.6 Processing Related to the Function of Navigation

The user's current location, which was obtained in a similar manner as aforementioned, is used as a basis for backend processing for calculating the shortest distance to the destination. The shortest distance is calculated using the road search function of ArcGIS pro.

4.3 Database

Tourism spot data are required to be collected in advance and registered in the database to enable the use of the functions of the system immediately after starting its operation. Therefore, data on a total of 113 tourism spots in the operation area of the system (Central Yokohama City) were collected using the tourism web media such as 4travel.jp provided by the Kakaku.com, Inc., Japan Tourism Guide provided by the Recruit Co., Ltd., and TripAdvisor provided by the Tripadvisor LLC. Then, the tourism spots were classified into five categories that were introduced in Sect. 4.1. Additionally, data on 1,187 restaurants in the operation area were also collected from the restaurant data of the Hot Pepper Web Service provided by the Recruit Co., Ltd. These two kinds of data were, respectively, processed in layers using ArcGIS Pro, and registered into the database in advance.

4.4 Interface

The interfaces for the function of tourism plan creation, the function of tourism plan information display, and the function of tourism plan sharing/viewing were developed such that they could be used anywhere on PCs and mobile information devices. There exist design differences depending on when the system is viewed on the PC or the mobile information device, but the interface was developed such that the same functions could be used in either case. Meanwhile, the interfaces for the function of the nearby spot search and the function of navigation, which use Web-AR, were developed with the assumption that they would be used on a mobile information device in a tourism area.

5 System Operation

5.1 Operation Overview

The system was operated for 30 days from December 22, 2023 to January 20, 2024, while targeting people around the operation area. The operation of the system was publicized on website of the authors' lab.

5.2 Operation Result

There were 50 users in total, comprising 32 males, 17 females, and one other person. Users in their 20s accounted for the largest proportion at 60%, followed by users in

their 50s and 60s at 17% each, and users in their 40s and 70s at 3% each. The result demonstrate that the system was primarily utilized by younger people, but also by people across a wide range of ages. Furthermore, 20 tourism plans were created by users during the operation period.

6 System Evaluation

After the end of the operation, the system was evaluated by conducting an online questionnaire survey for users and an access log analysis of their log data during the operation period. All the users responded to the questionnaire survey.

6.1 Evaluation Based on Online Questionnaire Survey

6.1.1 Overview of an Online Questionnaire Survey

An online questionnaire survey for users was conducted to evaluate (1) the principal functions and (2) the overall system in order to address the object of the present study. In the case of the question item on food tourism experience, 67% of respondents said that they had experience, and 33% said that they did not. Furthermore, in the case of the question item on advance creation of tourism plans, 17% responded with "I create solid plans," 64% responded with "I only create rough plans," and 19% responded with "I do not create any plans." The result indicated that 81% of respondents were engaged in tourism with plans created, and 19% of users were engaged in tourism without much of a plan.

Therefore, of the principal functions, the presence or absence of food tourism experience will be focused on the evaluation of the function of tourism plan creation. Meanwhile, the presence or absence of advance creation of tourism plans will be focused on the evaluation of other principal functions and the overall system. Table 1 shows a cross-tabulation of each evaluation item and the presence or absence of food tourism experience and advance creation of tourism plans. Pearson's chi-squared tests were performed on Table 1 to confirm whether there was any bias in the evaluation, because users experienced food tourism and created tourism plans in advance.

6.1.2 Evaluation of the Principal Functions

(1) *Sub-function of restaurant selection*

82% of respondents answered "Agree" or "Somewhat agree," and 18% of respondents answered "Somewhat disagree" or "Disagree." The chi-square test result showed that there was no bias in the evaluation based on the presence or absence of food tourism

Table 1 Evaluation result (person)

Question items	Travel experience	Strongly agree	Agree	Disagree	Strongly disagree	Total
Easy use of the sub-function of restaurant selection	Experience with food tourism	11 (10.9)	17 (17.7)	5 (4.8)	1 (0.6)	34
	No experience with food tourism	5 (5.1)	9 (8.3)	2 (2.2	0 (0.4)	16
	Total	16	26	7	1	50
Easy use of the sub-function of tourism spot selection function	Experience with food tourism	13 (13.6)	16 (15.6)	5 (4.8)	0 (0.0)	34
	No experience with food tourism	7 (6.4)	7 (7.4)	2 (2.2)	0 (0.0)	16
	Total	20	23	7	0	50
Easy use of the function of tourism plan sharing/viewing	Tourism plans created in advance	25 (25.4)	15 (14.8)	1 (0.8)	0 (0.0)	41
	No tourism plans created in advance	6 (5.6)	3 (3.2)	0 (0.2)	0 (0.0)	9
	Total	31	18	1	0	50
Easy use of the function of nearby spot search	Tourism plans created in advance	16 (16.4)	18 (18.0)	7 (6.6)	0 (0.0)	41
	No tourism plans created in advance	4 (3.6)	4 (4.0)	1 (1.4)	0 (0.0)	9
	Total	20	22	8	0	50
Easy use of the function of navigation	Tourism plans created in advance	13 (13.1)	23 (23.0)	5 (4.9)	0 (0.0)	41
	No tourism plans created in advance	3 (2.9)	5 (5.0)	1 (1.1)	0 (0.0)	9
	Total	16	28	6	0	50
Usefulness of the system	Tourism plans created in advance	25 (25.4)	15 (14.8)	1 (0.8)	0 (0.0)	41
	No tourism plans created in advance	6 (5.6)	3 (3.2)	0 (0.2)	0 (0.0)	9
	Total	31	18	1	0	50

(continued)

Table 1 (continued)

Question items	Travel experience	Strongly agree	Agree	Disagree	Strongly disagree	Total
Easy use of the system	Tourism plans created in advance	21 (21.3)	17 (17.2)	2 (1.6)	1 (0.9)	41
	No tourism plans created in advance	5 (4.7)	4 (3.8)	0 (0.4)	0 (0.1)	9
	Total	26	21	2	1	50

Note The numbers in brackets indicate the expected frequency

experience ($\chi^2(3) = 0.992$; $p > 0.05$). Thus, this function was easy to use for most users. However, the free response section included opinions such as "small icon size" and "difficult access to the detailed page after selecting the restaurant."

(2) *Sub-function of tourism spot selection*

84% of respondents answered "Agree" or "Somewhat agree," and 16% of respondents answered "Somewhat disagree" or "Disagree." The chi-square test result showed that there was no bias in the evaluation based on the presence or absence of food tourism experience ($\chi^2(3) = 0.986$; $p > 0.05$). Thus, this sub-function was easy to use for most users. However, the free response section included opinions such as "small icon size" and "unclear operation method."

(3) *Function of tourism plan sharing/viewing*

98.0% of respondents answered "Agree" or "Somewhat agree." The chi-square test result showed that there was no bias in the evaluation based on the presence or absence of advance creation of tourism plans ($\chi^2(3) = 0.960$; $p > 0.05$). From this result, it can be said that this function was extremely easy to use for almost all users.

(4) *Function of nearby spot search*

84% of respondents answered "Agree" or "Somewhat agree," and 16% of respondents answered "Somewhat disagree." The chi-square test result showed that there was no bias in the evaluation based on the presence or absence of advance creation of tourism plans ($\chi^2(3) = 0.979$; $p > 0.05$). Thus, this function was easy to use for most users. However, the free response section included opinions such as "information to be displayed using figures or pictogram."

(5) *Function of navigation*

87% of respondents answered "Agree" or "Somewhat agree," and 13% of respondents answered "Somewhat disagree" or "Disagree." The chi-square test result showed that there was no bias in the evaluation based on the presence or absence of advance creation of tourism plans ($\chi^2(3) = 0.999$; $p > 0.05$). Thus, this function was easy to use for most users. However, the free response section included opinions such as "destination direction to be displayed."

6.1.3 Evaluation of Overall System

(1) *Usefulness of the system*

98% of respondents answered "Agree" or "Somewhat agree." Furthermore, the chi-square test result showed that there was no bias in the evaluation based on the presence or absence of advance creation of tourism plans ($\chi^2(3) = 0.340; p > 0.05$). From this result, it can be concluded that the system was extremely useful even for almost all users who do not create tourism plans in advance.

(2) *Easy use of the system*

94% of respondents answered "Agree" or "Somewhat agree." The chi-square test result showed that there was no bias in the evaluation based on the presence or absence of advance creation of tourism plans n ($\chi^2(3) = 0.911; p > 0.05$). From this result, it can be concluded that the system was extremely easy to use for almost all users, though there is room for improvement in some functions.

6.2 Evaluation Based on Access Log Analysis

Furthermore, an access log analysis using the users' log data during the operation period was conducted using Google Analytics. The proportion of users' means of access to the system was as follows: PC, 82%; smartphone, 17%; and tablet devices, 1%. From this result, it can be said that the system of using the same design regardless of the device used was effective.

Next, Table 2 lists the percentage of visits by page. The total number of visits to the pages of the system is 1892. Table 2 shows that, with the exception of the login page, which was always the first page, the pages with the highest number of visits were, in descending order, the pages for the sub-function of restaurant selection, the sub-function of start/end station setting, the sub-function of tourism spot selection, the function of nearby spot search, and the function of tourism plan sharing/viewing. Therefore, the result showed that users used the system to create tourism plans centered on restaurants, and collect information on nearby spots in the operation area (Central Yokohama City). As a result, the system met the purpose of the present study. However, of the principal functions, the page for the function of navigation had a low access count. This was attributed to the problem of this function not displaying the destination direction, as mentioned in Sect. 6.1.3 (5).

Table 2 Percentage of visits by page (%)

Ranking	Page	Percentage
1	Login	21.4
2	Sub-function of restaurant selection	16.0
3	Digital map display of restaurant selection result	11.4
4	Sub-function of start/end station setting	11.2
5	Home	9.0
6	Sub-function of tourism spot selection	7.7
7	Function of nearby spot search	7.2
8	Sub-function of start/end station setting	5.9
9	Function of tourism plan information display	5.5
10	New user registration	2.7

6.3 System Problems and Improvement Strategies

The system problems and improvement strategies that were derived from the results of the questionnaire survey and the analysis of the access logs can be summarized in the following three suggestions.

(1) *Web page design*

Information on restaurants and tourism spots will be clearly displayed as a figure or pictogram on the digital map such that users can more easily confirm them.

(2) *Function of tourism plan creation*

An easy-to-understand manual will be made with instructions on how to create tourism plans that users can refer to as needed.

(3) *Function of navigation*

The unction of navigation will be equipped with a sub-function that shows the destination direction on the AR-based mobile information device screen. This is expected to reduce the problems that users encounter when searching for objects displayed on the mobile information device screen using AR.

7 Conclusion

In the present study, a system, which combined an original tourism plan creation system, Web-GIS and Web-AR in order to support tourism centered on food and drink, was designed and developed. The system has two unique functions. The first is to efficiently create tourism plans centered on restaurants for lunch and dinner. The second is to use location-based AR to visualize and display nearby restaurant

and tourism spot information on the AR-based mobile information device screen. Furthermore, this is a web system that can be used simply through a connection to the Internet from a web browser.

Central Yokohama City of Kanagawa Prefecture, Japan was selected as the operation area. Data on 1,187 restaurants and 113 tourism spots were collected from tourism web media, and then registered in the database in advance. The system was operated for 30 days from December 22, 2023 to January 20, 2024. During the operation period, there were 50 users and 20 tourism plans were created by them.

The system was evaluated by conducting an online questionnaire survey for users and an access log analysis of their log data during the operation period. The questionnaire survey result showed that the easy use of the principal functions was highly evaluated generally, regardless of food tourism experience or advance creation of tourism plan. Additionally, the usefulness and easy use of the overall system were also highly evaluated. Furthermore, the access log analysis result showed that there was a high number of visits to the pages for most of the principal functions, and the system was used in a manner consistent with the purpose of the present study.

Future research topics include improving the system according to the result of Sect. 6.3, and operating the system in other tourism areas to increase the usage rate and significance of its use.

Acknowledgements We would like to thank all the users of the food tourism support system that uses location-based Web-AR, and the online questionnaires survey. We would like to express our deepest gratitude to them.

References

1. Japan Tourism Agency (2018) Tourism community development case studies—good practices 2018. Tokyo
2. Japan Travel Bureau (JTB) Research Institute (2018) Survey on the spread of new technologies and services and lifestyle/travel. Tokyo
3. Jalan Research Center (2022) 13th Domestic lodging travel needs assessment report. Tokyo
4. Yokohama City (2022) FY2023 Yokohama City tourism consumption survey report. Yokohama
5. Ribeiro FR, Silva A, Barbosa F, Silva AP, Metrôlho JC (2018) Mobile applications for accessible tourism: overview, challenges and a proposed platform. Inf Technol Tour 19(1):29–59
6. Chen CC, Tsai JL (2019) Determinants of behavioral intention to use the personalized location-based mobile tourism application: an empirical study by integrating TAM with ISSM. Futur Gener Comput Syst 96:628–638
7. Yazdeen AA, Zeebaree SR (2022) Comprehensive survey for designing and implementing web-based tourist resorts and places management systems. Acad J Nawroz Univ (AJNU) 11(3):113–132
8. Lim KF, Chan J, Karunasekera S, Leckie C (2018) Tour recommendation and trip planning using location-based social media: a survey. Knowl Inf Syst 60:1247–1275
9. Hida M, Kanaya Y, Kawanaka S, Matsuda Y, Nakamura Y, Suwa H, Fujimoto M, Arakawa Y, Yasumoto K (2020) On-site trip planning support system based on dynamic information on tourism spots. Smart Cities 3(2):212–231
10. Nitu P, Coelho J, Madiraju P (2021) Improvising personalized travel recommendation system with recency effects. Big Data Ming and Analytics 4(3):139–154

11. Tavitiyaman P, Qu H, Tsang WL, Lam CR (2021) The influence of smart tourism applications on perceived destination image and behavioral intention: the moderating role of information search behavior. J Hosp Tour Manag 46:476–487
12. Avval AAN, Harounabadi A (2023) A hybrid recommender system using topic modeling and prefixspan algorithm in social media. Complex Intell Syst 9:4457–4482
13. Li H, Li M, Zou H, Zhang Y, Cao J (2023) Urban sensory map: how do tourists "sense" a destination spatially? Tour Manage 97:104723. https://doi.org/10.1016/j.tourman.2023.104723
14. Vranić V, Lang J, Nores ML, Arias JJP, Solano J, Laseca G (2024) Use case modeling in a research setting of developing an innovative pilgrimage support system. Univ Access Inf Soc 23(4):1543–1560. https://doi.org/10.1007/s10209-023-01047-1
15. Sasaki R, Yamamoto K (2019) A sightseeing support system using augmented reality and pictograms within urban tourist areas in Japan. Int J Geo Inf 8(9):381. https://doi.org/10.3390/ijgi8090381
16. Zhou X, Sun Z, Xue C, Lin Y, Zhang J (2019) Mobile AR tourist attraction guide system design based on image recognition and user behavior. In: Nedjah N, Perez GM, Gupta BB (eds) Advances in intelligent systems and computing. Springer, Heidelberg, pp 858–863
17. Ferentinos KP, Nakos YSH, Pristouris K, Barda MS (2020) Initial design and features of an augmented reality system for urban park touring and management. Int J Comput Theory Eng 12(5):106–112
18. Andrii A, Günther S, Ritzenhofen S, Mühlhäuser M (2020) AR sightseeing: comparing information placements at outdoor historical heritage sites using augmented reality. In: Proceedings of the 6th ACM on human-computer interaction. Honolulu, pp 1–17
19. Otsu K, Ueno T, Izumi T (2023) AR-based visitor support system for enhancing the liveliness of sightseeing spots using CG humanoid models. In: Chen JYC, Fragomeni G (eds) Lecture notes in computer science. Springer, Heidelberg, pp 591–611
20. Ikizawa-Naitou K, Yamamoto K (2020) A support system of sightseeing tour planning using public transportation in Japanese rural areas. J Civ Eng Arch 14(6):316–332
21. Sasaki R, Yamamoto K (2021) Sightseeing navigation system from normal times to disaster outbreak times within urban tourist areas in Japan. Appl Sci 11(10):4609. https://doi.org/10.3390/app11104609
22. Abe S, Sasaki R, Yamamoto K (2021) Sightseeing support system with augmented reality and no language barriers. In: Geertman S, Pettit C, Goodspeed R, Kauppi A (eds) Lecture notes in urban informatics for future cities. Springer, Heidelberg, pp 591–611
23. Sonobe H, Nishino H, Okada Y, Kaneko K (2021) Tourism support system using AR for tourists' migratory behaviors. In: Barolli L, Xhafa F, Javaid N, Spaho E, Kolici V (eds) Advances in internet, data & web technologies. Springer, Heidelberg, pp 669–710
24. Miyoshi Y, Okuno T (2018) Development of a system to support food tourism using knowledge-based recommendation. In: Proceedings of 80th national convention of IPSJ, vol 1. Tokyo, pp 443–444
25. Nakano M, Ohno S, Kamihara Y, Goto D, Suginaka H, Suda M, Sone K (2018) Proposal of restaurant recommendation service "OAISO" based on web data analysis: consideration on the concise restaurant dictionary extracted from varieties of food information in megalopolises. In: Proceedings of the 80th national convention of IPSJ, vol 1. Tokyo, pp 371–372
26. Horibe M, Motoh A, Moriyama I, Inuzuka N (2022) A method for recommending additional orders considering order time using restaurant data. IPSJ SIG Technical Report, Tokyo
27. Kumagai K, Okuno T (2022) Restaurant recommendation considering seasonal food and local ingredients for food tourism using knowledge-based recommendation. In: Proceedings of the 84th national convention of IPSJ, vol 1. Matsuyama, pp 561–562
28. Yoshida N, Motoh A, Shima K, Moriyama K, Matsui T, Ikuzuka N (2023) A method for recommending additional orders considering order time distribution in restaurants. In: Proceedings of the annual conference of JSAI 2023. Tokyo, pp 4F2GS1003
29. Kumarasiri C, Farook C (2018) User centric mobile based decision-making system using natural language processing (NLP) and aspect based opinion mining (ABOM) techniques for restaurant selection. In: Nedjah N, Perez GM, Gupta BB (eds) Advances in intelligent systems and computing. Springer, Heidelberg, pp 43–56

30. Luo Y, Xu X (2019) Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: a case study of Yelp. Sustainability 11(19):5254. https://doi.org/10.3390/su11195254
31. Ichimura S (2020) Searching for delicious restaurants using review. IPSJ J 61(11):1748–1756
32. Sarasa-Cabezuelo A (2023) Development of a restaurant recommendation system. In: Mittal H, Nanda SJ, Meng-Hiot L (eds) Algorithms for intelligent systems. Springer, Heidelberg, pp 443–455

BrainDetective: An Advanced Deep Learning Application for Early Detection, Segmentation and Classification of Brain Tumours Using MRI Images

Nazlı Tokatlı, Mücahit Bayram, Hatice Ogur, Yusuf Kılıç, Vesile Han, Kutay Can Batur, and Halis Altun

Abstract This study aims to create deep learning models for the early identification and classification of brain tumours. Models like U-Net, DAU-Net, DAU-Net 3D, and SGANet have been used to evaluate brain MRI images accurately. Magnetic resonance imaging (MRI) is the most commonly used method in brain tumour diagnosis, but it is a complicated procedure due to the brain's complex structure. This study looked into the ability of deep learning architectures to increase the accuracy of brain tumour diagnosis. We used the BraTS 2020 dataset to segment and classify brain tumours. The U-Net model designed for the project achieved an accuracy rate of 97% with a loss of 47%, DAU-Net reached 90% accuracy with a loss of 33%, DAU-Net 3D achieved 99% accuracy with a loss of 35%, and SGANet achieved 99% accuracy with a loss of 20%, all demonstrating effective outcomes. These findings aim to improve patient care quality by speeding up medical diagnosis processes using computer-aided technology. Doctors can detect 3D tumours from MRI pictures using software developed as part of the research. The work packages correctly handled project management throughout the study's data collection, model creation, and

N. Tokatlı (✉) · M. Bayram · H. Ogur · Y. Kılıç · V. Han · K. C. Batur · H. Altun
Istanbul Health and Technology University, Istanbul, Türkiye
e-mail: n.tokatli@istun.edu.tr

M. Bayram
e-mail: m.bayram@istun.edu.tr

H. Ogur
e-mail: h.ogur@istun.edu.tr

Y. Kılıç
e-mail: y.kilic@istun.edu.tr

V. Han
e-mail: v.han@istun.edu.tr

K. C. Batur
e-mail: k.batur@istun.edu.tr

H. Altun
e-mail: h.altun@istun.edu.tr

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_3

evaluation stages. Regarding brain tumour segmentation, 3D U-Net architecture with multi-head attention mechanisms provides doctors with the best tools for planning surgery and giving each patient the best treatment options. The user-friendly Turkish interface enables simple MRI picture uploads and quick, understandable findings.

Keywords 3D brain tumour diagnosis · Deep learning models · MR imaging · AI applications in Turkish health system

1 Introduction

This study's scientific value comes from applying deep learning and computer-aided systems for the early detection and classification of brain cancers. According to the latest World Health Organization data, brain tumours are among the most common causes of cancer death globally and may develop at any age [1]. They are most fatal to those under the age of 40. Early diagnosis of brain tumours dramatically improves patient survival and treatment success. Rapid and precise picture assessment is critical; however, manual evaluations are time-consuming and prone to errorsAs a result, artificial intelligence applications have become prevalent in techniques such as computed tomography (CT) and magnetic resonance (MR) [2]. The latest developments in deep learning and machine learning have significantly improved pattern identification in biological images [3]. These innovative approaches meet the needs of automated medical decision-making systems because human processes are expensive to accomplish, labour-intensive, and prone to errors [4].

In the present research, a multi-model approach developed with deep learning and machine learning algorithms outperformed existing methods for detecting and classifying brain cancers in terms of accuracy and efficiency. The BraTS 2020 dataset was crucial in training and testing these algorithms. This dataset contains a large amount of data for finding and categorizing cancers in brain MR images. Advanced deep learning models were utilized to speed up and improve brain tumour diagnosis accuracy in these MRI scans. Modern deep learning structures such as U-Net, DAU-Net, DAU-Net 3D, and SGANet were used to analyse brain MRI scans. Each model's performance was evaluated by achieving high accuracy rates and low loss values while segmenting and classifying brain tumours. A user-friendly interface was created to upload MR images and retrieve quick conclusions. The application's effectiveness was increased through case studies and testing procedures. This research is essential for precisely identifying the spatial extent and positioning of brain tumours using 3D brain models generated by MRI scans, which is required for brain telemetry device placement and surgical planning. The literature indicates that machine learning and deep learning algorithms are helpful in the diagnosis of brain tumours [5–11]. However, there are places where these procedures might be increased for more precise and effective results. This study is planned to improve brain tumour detection by filling the gaps in the literature. Furthermore, segmenting brain tumours will improve surgical planning by preserving essential structures surrounding the tumour.

The study's contributions include early diagnosis, ob taining a 3D image of the brain, convenience in brain telemetry and neurosurgical applications, a user-friendly Turkish interface, and developing a clinical decision support system via the application. The study's sections cover related work, research methodology, data preparation, model designs, and findings. The results section presents the findings, and the conclusions section interprets the findings to conclude the study.

2 Related Work

The diagnosis and classification of brain tumours have been a topic of extensive research and experimentation in the field of medical imaging for many years. These studies recommend using various algorithms and deep learning methods to analyse brain MRI images.

DAU-Net is a model that incorporates attention mechanisms along with U-Net. It is particularly effective in capturing detailed features in medical images. Introduced by Zhang et al., DAU-Net provides more detailed and precise segmentation results due to its attention mechanisms. This feature has made DAU-Net one of the models used in this project for distinguishing various types of brain tumours [12].

3D DAU-Net is a model optimized for volumetric data analysis and capable of working with 3D medical images. Developed by Zhao et al., this model offers high accuracy and reliability in 3D analysis of brain MRI images. Its 3D tumour segmentation and classification capability has made it a significant component of the project [13].

SGANet is a model that combines generative adversarial networks (GANs) with U-Net. Proposed by Yu et al., this model has proven effective in improving the quality of medical images and achieving better results in segmentation tasks. The high performance of SGANet plays a crucial role in this project's segmentation and classification of brain tumours [14].

Recent studies have examined deep learning techniques for brain tumour detection, segmentation, and classification using MRI images. Various models have been proposed in this field, such as U-Net [15] for segmentation and CNN [16] for classification. Advanced approaches like YOLOv5 and FastAi have yielded promising results with accuracy rates of 85.95% and 95.78%, respectively [17]. A custom Mask R-CNN model with a DenseNet-41 backbone demonstrated high accuracy in both segmentation (96.3%) and classification (98.34%) tasks [18]. Transfer learning techniques like AlexNet CNN have achieved an impressive accuracy rate of 99.62% [19].Multi-task classification studies using CNN have also been explored for various tumoUr classification tasks [20]. Additionally, 3D-U-Net models have been used for volumetric segmentation, followed by CNN-based classification [21]. These advancements aim to improve the early detection and diagnosis of brain tumours.

3 Proposed Design

The proposed system aims to analyse brain MRI images using deep learning models. This system offers an innovative approach to brain tumour segmentation and detection by leveraging the BraTS 2020 dataset. The system utilizes a comprehensive database consisting of T1, seg, T1ce, T2, and FLAIR modalities to accurately detect brain tumour types in both 3D and 2D. The BraTS 2020 dataset consists of a total of 369 patients, with each patient having five.nii files: flair.nii, t1.nii, t1ce.nii, t2.nii, and seg.nii, which contains the 3D-labelled tumour. Additionally, a .csv file provides information on whether the tumour is benign or malignant, along with another .csv file containing data such as the patient's age, whether the tumour was removed, and survival time. During model training, these data were processed and converted into .npy and .tfRecord formats. The dataset is divided into two main categories: HGG (High-Grade Glioma) and LGG (Low-Grade Glioma). With augmentation techniques, the dataset size was increased to over 100,000 samples. However, the BraTS 2020 dataset has significant limitations, such as the absence of important clinical variables (volume, gender, histopathological data) and tumour subtypes. These limitations may restrict the model's application in real-world scenarios. In the study, these limitations were considered, with a focus on the generalizability of the findings. To enhance applicability across a broader clinical range, the use of various datasets, along with techniques like transfer learning or domain adaptation, can enable the model to adapt to different data sources. The project's software architecture is designed to be modular and scalable, with each model developed as an independent module and integrated into the application. The system architecture is shown in Fig. 1.

3.1 Data Processing

Data processing workflows were based on the Factory and Strategy design patterns. The Factory pattern defines preprocessing stages across numerous modalities, whilst the Strategy style allows for the dynamic deployment of data augmentation approaches [22, 23]. These approaches ensure that data preparation operations are both flexible and efficient. Initially, the system meticulously processes each modality to improve data quality. Modality-based directories are built, and axis corrections and channel placements are performed. Normalization is carried out utilizing Z-score and Min-Max approaches. This allows the model to be trained using consistent, high-quality data. Data augmentation techniques are subsequently applied to improve the model's generalization capability, and the processed data is stored in NumPy arrays and TFRecord format for easy access and processing. This approach ensures that the data is displayed effectively. Figure 2 displays the data processing and storage procedures.

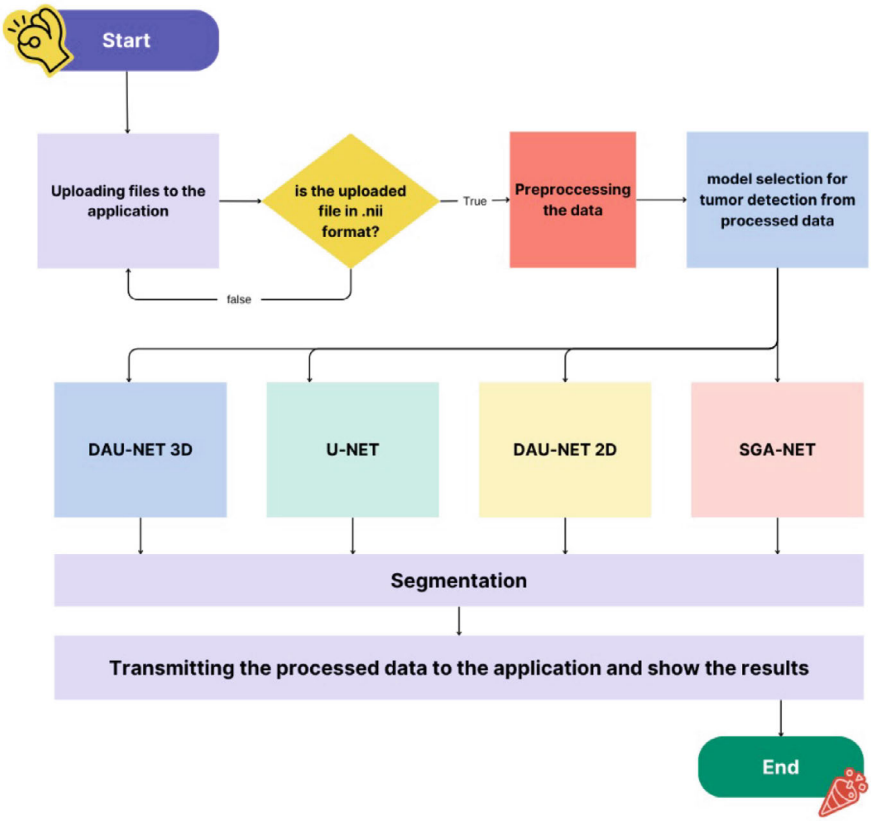


Fig. 1 General design stages of the application

3.2 Model Development

The project focuses on deep learning architectures such as DAUNet, DAUNet 3D, U-Net, and SGANet. Each model is evaluated and improved based on existing approaches in the literature, optimizing their segmentation capabilities. These models aim to achieve high accuracy rates while minimizing the risk of overfitting. Figure 3 explains the step-by-step process of the model development and improvement.

4 Experimental Methodology

In the experimental phase of the project, U-Net, DAU-Net, DAU-Net 3D, and SGANet models were trained with the BraTS 2020 dataset, and the hyperparameter

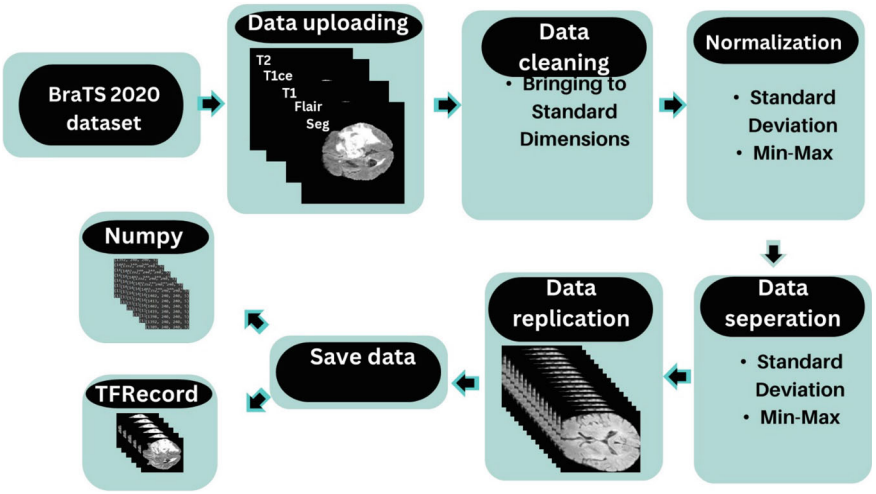


Fig. 2 Data preparation process of the application

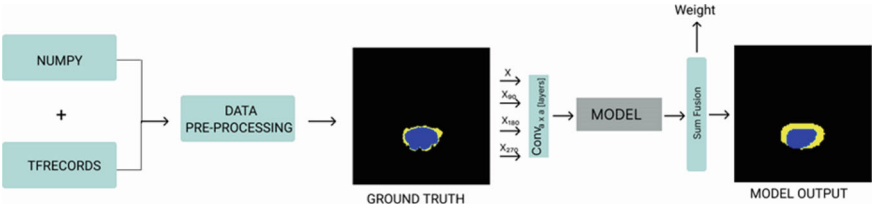


Fig. 3 Processing stages of models

settings of each model were carefully optimized. Throughout the training process, the performance of each model was evaluated based on accuracy rates and loss values.

4.1 Data Preparation and Preprocessing Process

In the experimental phase of the project, U-Net, DAU-Net, DAU-Net 3D, and SGANet models were trained with the BraTS 2020 dataset, and the hyperparameter settings of each model were carefully optimized. Throughout the training process, the performance of each model was evaluated based on accuracy rates and loss values [24, 25].

During the preparation phase, directories were established for each patient in the dataset, and data use and access were streamlined by categorizing directories by modality type. The modes in the files, which contained axis corrections and channel placements, were loaded and processed appropriately. Data was normalized using standardization (Z-score) and min-max normalization methods, with clipping

operations used to limit the influence of outliers [26]. Data augmentation techniques such as rotation, shifting, cropping, and mirroring were used further to develop the model's generalization capabilities [27]. Processed data was saved as NumPy arrays (.npy format) for quick access and then translated to TensorFlow's TFRecord format. The TFRecord format was created for efficient processing of big datasets, and the data was serialized and saved in this format for use in training and testing [28, 29].

4.2 U-Net

The U-Net layout is a highly effective structure for brain tumour segmentation and biomedical imaging. It uses an autoencoder-like structure to process the input image via encoding and decoding routes. While the encoding path collects contextual information, the decoding path preserves details and returns the image to its original size [30, 31]. U-Net has shown outstanding performance, particularly in the precise identification of brain tumours, with high accuracy even with fragile datasets [32–34].

4.3 DAUNet 3D

The DAUNet 3D model uses attention mechanisms and deep learning techniques to achieve remarkable precision in brain tumour segmentation. Attention procedures increase segmentation accuracy by detecting key regions, but deep learning systems may learn complicated structures and fine details. 3D convolutional layers maintain the three-dimensional structure of brain images, allowing for more detailed analysis. Incremental learning techniques incorporate segmentation findings at many levels, improving overall accuracy. DAUNet 3D is a valuable tool for medical imaging applications that require volumetric analysis, and it performs excellently.

4.4 DAUNet

DAUNet and DAUNet 3D models can be helpful for medical image segmentation. DAUNet works with 2D slice data, but DAUNet 3D handles 3D volume data. DAUNet uses 2D convolutional, max pooling, and upsampling layers to improve accuracy using attention methods. Generalized Dice Loss and Categorical Crossentropy are utilized as loss functions. During the training phase, the model's performance was measured using metrics such as the Dice coefficient, sensitivity, and specificity, and the training process was shown. The model produced effective results for brain tumour segmentation.

4.5 SGANet

SGANet is a deep learning model that provides excellent accuracy and precision in medical image segmentation. It improves segmentation problems by integrating Generative Adversarial Networks (GAN) and U-Net architectures [35, 36]. The model employs a U-Net-like generator network to turn images into segmentation masks and a discriminator network to evaluate the masks' realism [37, 38]. The adversarial training technique improves segmentation accuracy through competition [39]. SGANet provides effective results in medical applications, such as the exact segmentation of brain tumours, and has demonstrated effectiveness in this study [40].

In this study, the models were not combined. Instead, the application includes four distinct models capable of detecting tumours in both 3D and 2D, allowing the user to select the desired model. The developed interface currently operates with the 3D DAU-Net model, which has demonstrated the highest performance. The comparison of multiple models was conducted to evaluate the performance of different models on the same dataset and to assess each model's effectiveness in solving specific problems. Although this approach is rare in the literature, it is crucial for selecting the best-performing model or testing whether the integration of models provides new solutions. In future stages, the integration of other models used in the study is planned. The joint evaluation of the models aims to identify the strengths and weaknesses of various model architectures and develop solutions that are better suited to clinical applications. As shown in Table 1, the comparison criteria include performance metrics, which allowed us to objectively assess the models' segmentation success and overall performance. The attention mechanisms and multi-scale feature extraction capabilities of SGANet and DAU-Net enhanced their segmentation and classification performance. These models demonstrated higher accuracy and lower loss rates compared to others (Fig. 4).

Table 1 Performance metrics of models

Model	Dice coefficient	Sensitivity	Specificity
DAU-Net	[0.9798, 0.8940, 0.8950]	[0.9846, 0.8878, 0.8780]	[0.9995, 0.9991, 0.9970]
3D DAU-Net	[0.9798, 0.8940, 0.8950]	[0.9846, 0.8878, 0.8780]	[0.9995, 0.9991, 0.9970]
SGANet	[0.9700, 0.9725, 0.9610]	[0.9846, 0.8878, 0.8780]	[0.9995, 0.9991, 0.9970]
U-Net	[0.9826, 0.9753, 1.0]	[0.9894, 0.9891, 1.0]	[0.9986, 0.9969, 1.0]

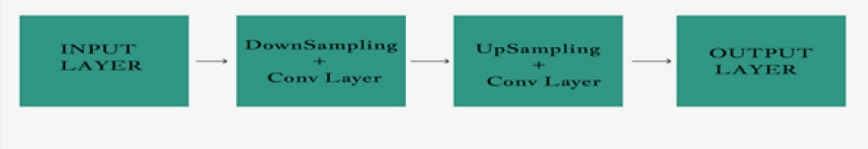


Fig. 4 General models and layers

5 Experiment Results

In conclusion, the developed models have demonstrated high accuracy rates and effective performance. Important metrics such as Dice Coefficient, Sensitivity (Recall), and Specificity were used to evaluate model performance. These metrics are commonly used to measure how close the model's segmentation results are to the ground truth and their accuracy. The Dice Coefficient, Sensitivity, and Specificity calculations for the developed models were performed using the following formulas:

$$\text{Dice Katsayısı} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Duyarlılık (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Özgüllük(Specificity)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

6 Discussion

In recent years, significant progress has been made in using deep learning models for the detection and classification of brain tumours. This study presents a rare approach in the literature by integrating U-Net, DAU-Net, DAU-Net 3D, and SGANet models into a single platform, enabling the comparative evaluation of different models and techniques. In particular, the DAU-Net and SGANet models demonstrated superior performance due to their attention mechanisms and multi-scale feature extraction capabilities. One of the primary reasons for the observed loss rate is the class imbalance in the BraTS 2020 dataset. In this study, 3D MRI images in .nii format were utilized. During the data preprocessing phase, the five .nii files obtained for each patient from the BraTS 2020 dataset were converted into .npy or .tfRecord formats. Subsequently, these data were segmented into 155 slices per patient. After the slicing process, random data augmentation techniques were applied. However, including brain regions without tumours in this process led to the model's insufficient learning

of rare tumour classes. This issue may result in incorrect outcomes during tumour segmentation (Fig. 5).

To mitigate the effects of this issue, improvements can be made during the model training phase, particularly in custom layers and preprocessing steps, taking class imbalance into account. Specifically, implementing oversampling for rare classes or undersampling for the majority class can improve the representation of rare classes, and advanced data augmentation techniques can enhance the overall performance of the model.The DAU-Net model, due to its attention-based architecture, achieved high success in brain tumour segmentation across metrics such as Dice coefficient, sensitivity, and specificity.

On the other hand, the SGANet model achieved a 99% accuracy rate due to its enhanced generalization capacity, which was obtained through the use of ResNet blocks, Guided Attention Blocks, and the GaussianNoise layer. The integration of these four models into a single platform allowed for a comprehensive evaluation of their strengths and made a significant contribution to the literature. The user-friendly interface developed in the study enables healthcare professionals to quickly and accurately analyse MRI data. Additionally, offering the interface in Turkish provides localized benefits, enhancing accessibility within the Turkish healthcare system. The

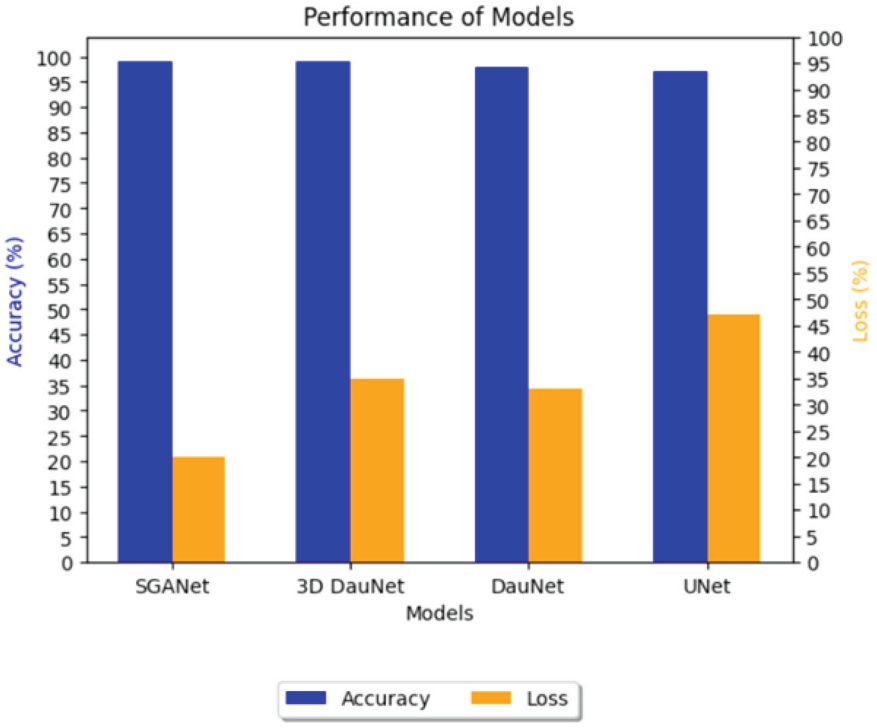


Fig. 5 Performance of models

study has some limitations, particularly regarding the size and diversity of the BraTS 2020 dataset, which may affect the generalizability of the findings. Future studies will use larger datasets, obtain the necessary ethical approvals from public hospitals in Turkey, and conduct large-scale clinical trials to increase the reliability of the results. In conclusion, this study demonstrates the potential of deep learning-based systems in medical image analysis and highlights the importance of these systems in clinical applications. These systems are particularly valuable in reducing the workload of doctors and alleviating the impact of the shortage of neurologists in public hospitals in Turkey. Further research supported by large datasets and clinical trials could enhance the effectiveness and applicability of these models even more.

7 Future Work

The developed Turkish interface and 3D imaging capabilities are expected to provide significant contributions in terms of accessibility and ease of use within the Turkish healthcare system. How the interface will be tested in practice and how user feedback will be collected are critical points for evaluating the software's effectiveness. In future stages, pilot studies are planned to assess the software's applicability in healthcare institutions. These studies will aim to ensure compatibility with international health standards such as Health Level Seven (HL7) and Digital Imaging and Communications in Medicine (DICOM), enabling smooth integration of the software with existing hospital information management systems. Comprehensive tests will be conducted in partnership with healthcare institutions affiliated with our university to evaluate to what extent the software reduces the workload of doctors in clinical settings. Thanks to its automatic 3D imaging and analysis features, the software is expected to significantly reduce doctors' workloads by speeding up diagnostic processes compared to manual methods. The results of these tests will demonstrate how well the software aligns with its goals while providing valuable feedback to optimize clinical integration. Additionally, for economically disadvantaged citizens, the software aims to contribute by reducing the number of tests needed for tumour detection, thus enabling access to healthcare services without financial concerns. Images of the current interface of the project can be found in Figs. 6 and 7.

8 Conclusion

In this research, brain tumours were quickly and accurately diagnosed using deep learning-based models. There is a chance that this system will raise the standard of healthcare. Deep learning techniques help in the early detection of brain tumours, which makes early and efficient treatments possible. Furthermore, the program and user interface that have been developed are user-friendly, allowing for 3D imaging

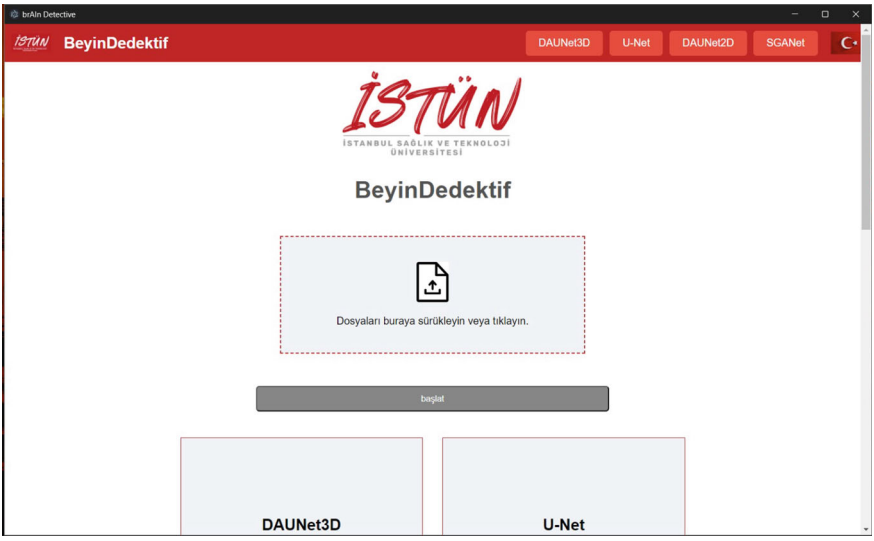


Fig. 6 Application interface

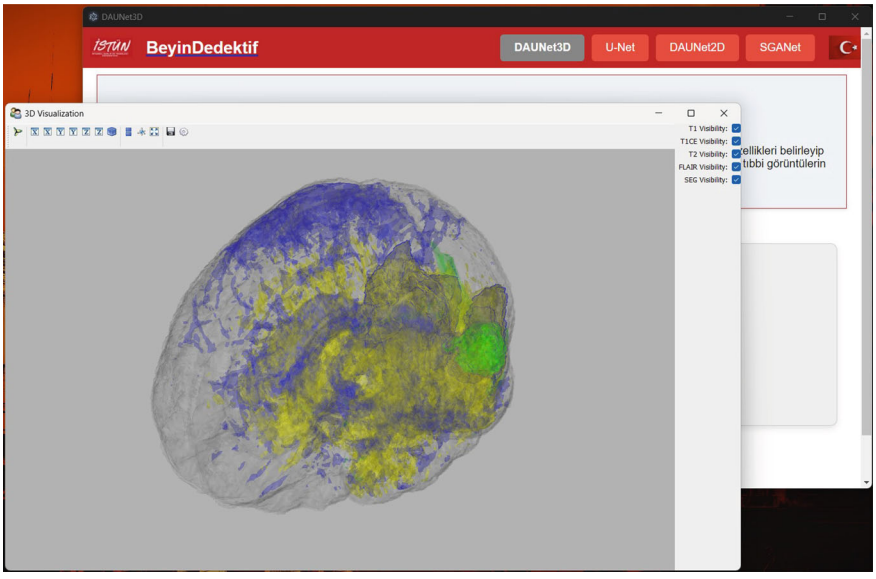


Fig. 7 3D tumour detection in practice

and extensive analysis of the brain and tumours. This information is useful for clinical evaluations and helps with treatment planning. The study has demonstrated the effectiveness of deep learning techniques in medical image analysis and has contributed to the proliferation of artificial intelligence applications in healthcare. This

project highlights the potential of artificial intelligence in the healthcare sector, forming a crucial foundation for future research and clinical applications. The broader application of deep learning-based systems in healthcare will improve patient care and make medical diagnosis processes more effective.

References

1. Abd El Kader I, Xu G, Shuai Z, Saminu S, Javaid I, Salim Ahmad I (2021) Differential deep convolutional neural network model for brain tumor classification. Brain Sci 11:352
2. Baran T (2019) Yapay Zekâ ve Radyoloji
3. Taghadomi-Saberi S, Hemmat M (2015) Görüntü Ã–şleme ve Medikal Görüntü Analizi
4. Karabulut G (2016) Otomatik Medikal Karar Verme Sistemleri
5. Ari A, Alpaslan N, Hanbay D (2015) Computer-aided tumor detection system using brain MR images. In: 2015 Medical technologies national conference (TIPTEKNO). IEEE, Bodrum, Turkey, pp 1–4
6. Bulut F, Kiliç I, Ince IF (2018) Beyin Tümörü Tespitinde Görüntü Bölütleme Yöntemlerine Ait Başarımların Karşılaştırılması ve Analizi. Deu Muhendislik Fakultesi Fen ve Muhendislik 20:173–186
7. Afshar P, Mohammadi A, Plataniotis KN (2018) Brain tumor type classification via capsule networks. In: 25th IEEE International conference on image processing (ICIP). IEEE, Athens, Greece, pp 3129–3133
8. Vani N, Sowmya A, Jayamma N (2017) Brain tumor classification using support vector machine. Int Res J Eng Technol (IRJET) 4(7):792–796
9. Swati ZNK, Zhao Q, Kabir M, Ali F, Ali Z, Ahmed S, Lu J (2019) Brain tumor classification for MR images using transfer learning and fine-tuning. Comput Med Imaging Graph 75:34–46
10. Tas MO, Ergin S (2020) Detection of the brain tumor existance using a traditional deep learning technique and determination of exact tumor locations using K-means segmentation from MR images. Ã–Ã–leri Mühendislik çalışmaları ve Teknolojileri Dergisi 1:91–97
11. Saxena P, Maheshwari A, Tayal S, Maheshwari S (2019) Predictive modeling of brain tumor: a deep learning approach. Adv Intell Syst Comput 1189:275–285
12. Zhang Y, Wang S, Yang Y (2018) DAU-Net: a deep attention u-net for segmenting multi-class medical images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
13. Zhao X, Xie Y, Zhang S, Lu Z (2020) 3D DAU-Net: a deep attention U-Net for 3D biomedical image segmentation. Med Image Anal 60:101606
14. Yu L, Zhang L, Yang M (2021) SGANet: a generative adversarial network for semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
15. Arafat A et al (2023) U-Net based brain tumor segmentation using MRI images. J Med Imaging Health Inform 13:625–634
16. Aamir M et al (2022) Brain tumor classification using convolutional neural networks. Comput Math Methods Med 2022:1–10
17. Dipu NM et al (2021) Brain tumor detection and classification using YOLOv5 and FastAi. IEEE Access 9:16984–16995
18. Masood M et al (2021) Brain tumor segmentation and classification using custom mask R-CNN with DenseNet-41 backbone. J Healthc Eng 2021:1–11
19. Badjie B, ülker E (2022) Brain tumor classification using transfer learning: AlexNet CNN model. Turk J Electr Eng Comput Sci 30:2317–2328
20. Kokila B et al (2021) Multi-task classification using CNN for brain tumor classification. Mater Today: Proc 45:4120–4126

21. Agrawal P et al (2022) Volumetric brain tumor segmentation and classification using 3D-UNet and CNN. J Digit Imaging 35:901–910
22. Gamma E, Helm R, Johnson R, Vlissides J (1994) Design patterns: elements of reusable object-oriented software. Addison-Wesley
23. Larman C (2004) UML and patterns: an introduction to object-oriented analysis and design and iterative development. Prentice Hall
24. Schlageter KE, Molnar P, Lapin GD, Groothuis DR (1999) Microvessel organization and structure in experimental brain tumors: microvessel populations with distinctive structural and functional properties. Microvasc Res 58(3):312–328
25. Menze BH, Jakab A, Bauer S et al (2014) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34(10):1993–2024. https://doi.org/10.1109/TMI.2014.2325008
26. Zhang Y, Yang L, Wang L, Zhang H (2019) A survey on image normalization in deep learning. Pattern Recogn 89:99–112. https://doi.org/10.1016/j.patcog.2019.01.017
27. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60. https://doi.org/10.1186/s40537-019-0197-0
28. Harris CR, Millman KJ, van der Walt SJ, Gommers R et al (2020) Array programming with NumPy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2
29. TensorFlow Documentation (2020) TFRecord and tf.data. https://www.tensorflow.org/api-docs/python/tf/data
30. U-Net is a fully connected CNN used for efficient semantic segmentation of images. Such U-Net deep neural network fits in various analytical tasks of wide ranging application (2020)
31. The U-Net architecture is based on an autoencoder network where the network will copy its inputs to its outputs (2020)
32. This architecture has several applications ranging from consumer videos, earth observations, and medical imaging (2020)
33. The encoder path of U-Net captures the context of the input image, this path is simply a pipeline of convolutional and pooling layers (2020)
34. Similarly to an autoencoder network, U-Net contains two paths, a contraction path (encoder) and a symmetric expanding path (decoder) (2020)
35. Liu X et al (2020) SGANet: a new generative adversarial network for medical image segmentation. Med Image Anal J
36. Wang Y et al (2021) Adversarial learning for medical image segmentation with SGANet. IEEE Trans Med Imaging
37. Zhang H et al (2019) Enhancing medical image segmentation with SGANet: a deep learning approach. J Healthc Eng
38. Kim S, Lee J (2020) Data augmentation techniques for SGANet in medical imaging. Comput Biol Med
39. Sharma R et al (2021) Balancing pixel and adversarial loss in SGANet for accurate segmentation. Pattern Recogn Lett
40. Luo Q et al (2020) Clinical applications of SGANet in brain tumor segmentation. Front Neurosci

SKY CONTROL: A Novel Concept for a Vendor-Agnostic Multi-cloud Framework to Optimize Cost Control and Risk Management for Small and Medium-Sized Enterprises

Christian Baun, Henry-Norbert Cocos, and Martin Kappes

Abstract Multi-cloud setups have become increasingly common in the industry, and adopting this method brings many opportunities for companies like vendor diversification, a selection of Best-of-Breed Services, and an increased resilience of the services used. However, this approach also brings challenges for the users, such as increased complexity of managing the services across vendors and increased vulnerability of the service. Another unsolved issue is the need for more transparency in running costs of using multiple services from many vendors and the compliance of the services with binding regulations. Our proposed framework, SKY CONTROL, will tackle those challenges and develop a comprehensive planning tool for complex, distributed IT infrastructures. With our innovative solution approach, we will conduct static and dynamic resource analyses of the resource inventory. In addition, a cost calculator for hybrid cloud users will be implemented, providing an aggregated cost overview of cloud and on-premise systems. At the same time, critical data requires an equally transparent option for risk management and information governance so that data and processes in hybrid infrastructures are always located on systems with an appropriate level of protection. Our solution marks the first concrete implementation of the innovative Sky computing concept for small-medium enterprises.

Keywords Sky computing · Multi-cloud · Cost control · Risk management · Vendor-agnostic framework

C. Baun (✉) · H.-N. Cocos · M. Kappes
Department of Computer Science and Engineering, Frankfurt University of Applied Sciences, Frankfurt am Main, Germany
e-mail: christianbaun@fb2.fra-uas.de
URL: https://www.frankfurt-university.de

H.-N. Cocos
e-mail: cocos@fb2.fra-uas.de

M. Kappes
e-mail: kappes@fb2.fra-uas.de

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_4

1 Introduction

The global transformation of IT infrastructures from purely organization-internal on-premise systems to complex multi- and hybrid cloud systems is already complete in many companies; in others, it is still in its infancy [12, 13]. A multi-cloud strategy selects platforms, infrastructures, and applications from different cloud service providers (CSPs) for a specialized purpose. It integrates them organizationally and technically with the company's on-premise systems [17]. The advantages of hybrid multi-cloud systems are manifold [11, 13, 26]:

- Reduction of provider dependency.
- Cost optimization.
- Load balancing.
- Business continuity through partial redundancy.
- Selection of the best service offerings for the respective application.
- Increased security through diversification of data storage and processing.

At higher organizational levels, multi-cloud strategies can promote innovation and agility for data analysis. In hybrid on-premise/cloud systems, (hosted) components are outsourced in the platform, and infrastructure-as-a-service models (IaaS) and native cloud applications (software as a service, SaaS) are used. While using PaaS and IaaS only relieves IT of some administrative tasks (i.e., the physical computer platform), the software on these platforms must continue to be maintained and managed traditionally. SaaS components are fundamentally different: deployed and maintained via user-defined configuration interfaces. PaaS and SaaS components can often communicate via defined interfaces and standard protocols (e.g., REST), which facilitates the development of connectors.

As a further development, the various platforms, infrastructures, and applications of different CSPs are not only orchestrated (i.e., managed and scaled) through standardized control interfaces with a high degree of automation but are also made seamlessly interoperable. This enables data processing across different clouds (e.g., advanced functions such as cross-cloud data analytics). This status is still partly visionary [3, 8] and the subject of ongoing research efforts. In an ideal multi-cloud environment, IT staff configure components once during deployment to ensure interoperability and the desired functionality.

Due to these developments, the fundamental change in work organization is to replace recurring administrative and programming tasks without added value with demand-oriented configuration tasks (on deployment or when functions are changed) that enable immediate productivity. The following project proposal is based on this only partially fulfilled promise of a positive evolution of IT infrastructures and makes a specific contribution here. Despite the technological development described above, companies' IT infrastructure costs are constantly rising, both in absolute terms and as a proportion of total operating costs.

The concept of sky computing [30, 37] is relevant in this context, as this new concept aims to abstract the cloud resources of different service providers. Thanks to

Table 1 Spendings on infrastructure in millions dollar

	Expenses (2021)	Growth (2021) (%)	Expenses (2022)	Growth (2022) (%)	Expenses (2023)	Growth (2023) (%)
Data center systems	207.306	6.7	218.643	5.5	230.385	**5.4**
Software	614.494	15.9	674.889	9.8	754.808	**11.8**
Devices	809.452	16.1	824.600	1.9	837.844	1.6
IT services	1185.103	10.6	1265.127	6.8	1372.892	**8.5**
Communication services	1443.419	3.4	1448.396	0.3	1477.798	2.0
Total	4259.773	9.5	4431.646	4.0	4673.728	5.5

sky computing, customers should be able to automatically access the resources from different providers best suited to their requirements. Sky computing is, therefore, a form of orchestration of various service offerings. From the customer's perspective, it has the potential to significantly improve the selection and use of suitable service offerings. So far, sky computing has only existed as a research concept. No solutions are yet available on the market. Table 1 shows that currently, according to Gartner, Inc. Datacenter Insider [27], and in the foreseeable future, the IT-related cost drivers are in data center systems, IT services, and software (highlighted in bold). These are the elements of hybrid infrastructures for which our solution enables effective cost control for the first time.

2 Background and Related Works

The project originates in distributed systems [34] and is based on the methods and technologies of cloud computing [6]. The National Institute of Standards and Technology (NIST) characterizes cloud computing by the five properties [23] on-demand self-service, whereby the provisioning of resources such as computing power or storage runs automatically without the interaction of a provider, the provision of services via the network and interaction with these via standard mechanisms (called Broad Network Access in the NIST definition). Furthermore, the resources are available via a pool that users can access. The resources are transparent and available to the customer in (seemingly) infinite quantities (resource pooling in the NIST definition). Building on this, the resources have an elastic structure. They can adapt to the customer's needs, thus automatically adapting to their requirements (called Rapid Elasticity in the NIST definition). The fifth characteristic is the measurability of the available cloud services, which allows for the exact, demand-based billing of the services.

In recent years, the range of cloud computing offerings has increased, leading to numerous public and private service offerings. Based on the hybrid cloud deployment model, multi-cloud environments have become established in the industry [25]. These have the advantage of avoiding a vendor lock-in, where customers become dependent on a single service provider. By adopting a multi-cloud strategy, companies can pursue a flexible cloud strategy. In multi-cloud environments, services or parts are distributed or operated in parallel by different providers [35]. This has economic and organizational advantages, as the distributed services can be used flexibly. This also leads to increased availability of the service offering and, together with a private service offering, increases the availability of resources. However, this strategy also has a disadvantage, as the range of services offered by CSPs could be more transparent, and it is not possible to make them uniformly usable, which is why much individual effort is required from users [3]. Another issue users of multi-cloud architectures face is the increased complexity and lack of overview of the cloud service zoo, which grows exponentially by using multiple cloud services from multiple CSPs. One additional challenge is the lack of interoperability between the APIs (Application-Program Interface) of cloud services differing between the CSPs. Ensuring a smooth integration between services from different providers is very important in multi-cloud architectures, and maintaining data consistency and synchronization can be very difficult between different vendors [4]. Sky computing addresses some of these issues and tries to solve the lack of clarity and inconsistent usability of the services offered by public CSPs [10, 24, 36, 37]. Here, a further layer of abstraction will be placed over the CSP's offering, enabling users to uniformly provision cloud services regardless of their operating location (more in Sect. 4).

The SkyPilot [35] project at Berkeley University is the first notable application in sky computing. With SkyPilot, the research team led by Stoica et al. has developed the first intercloud broker capable of executing various machine learning workloads across different providers. The broker consists of several components, such as a service catalog that records and stores the prices of the individual services of the various cloud providers and an optimization engine to calculate the optimal price for running an instance with a suitable cloud provider. The work of Stoica et al. can be seen as the first serious attempt to implement a sky computing application. However, this solution is at a very early and experimental stage and only serves to illustrate the paradigm. Initial experiments are limited to the application of machine learning applications such as a Large Language Model or Natural Language Processing applications in IaaS environments. The selection of cloud offerings used here is minimal. In contrast to SkyPilot, SKYCONTROL will look at industry-standard applications and services for small-medium enterprises (SME) and offer a solution for them.

2.1 State of the Art

The state of the art reflects freely or commercially available technologies. We give a relevant example for each. On the technical side, OpenTelemetry [31], is a good

landmark for the state of the art. OpenTelemetry is an open-source suite of instrumentation tools designed to simplify the implementation of tracing, metrics, and logging-in applications.

Two key features are:

- **Tracing**: OpenTelemetry's tracing capabilities enable the capture and visualization of requests and transactions across visualization of requests and transactions across distributed systems, which enables a detailed analysis of the performance and behavior of applications.
- **Metrics**: OpenTelemetry provides mechanisms for collecting, aggregating, and visualization of metrics such as CPU utilization, memory usage, and request rates to provide insights into application performance and behavior.

OpenTelemetry was developed to improve the observability of distributed systems. Many cloud providers offer interfaces to OpenTelemetry so that hybrid infrastructures can be monitored. As you can see, OpenTelemetry focuses entirely on measuring computing power in real time and creating time series data. In other words, it is all about optimizing the performance of distributed systems. Neither cost aspects nor the criticality of data and processes are considered here. Furthermore, OpenTelemetry is aimed at analyzing a given distributed system and does not provide any information on how to relocate processes or data to achieve better efficiency.

Research into the status of the cost-optimized use of (multi-) cloud systems shows, on the one hand, that the boom in research into cloud systems has already peaked a decade ago. There is a strong focus on scheduling algorithms for dynamic load balancing in cloud systems [1, 14], particularly on dynamic load balancing algorithms. Static system planning needs to be studied more. Optimization is mainly considered from the perspective of the cloud provider [20]. Hybrid systems are rarely considered, and cost optimization fed by heterogeneous sources needs to be researched [22, 32].

IBM's Instana Observability product is a benchmark for comprehensive monitoring solutions for (performance) monitoring of highly complex, heterogeneous IT infrastructures [28]. The solution is the most feature-rich on the market. It allows monitoring and performance measurement for various systems, clouds, end devices, and applications and offers uniform visibility for the real-time performance data obtained.

Instana also focuses entirely on optimizing the overall performance of the system. Although the product's website mentions some projects in which Instana helped achieve a more effective infrastructure overall, these are activities for which the software only provides a data basis in terms of performance. Achieving cost efficiency is always a downstream, manual activity.

Control Plane is a product that integrates on-premise and various cloud provider platforms [16]. It is a hybrid platform that enables cloud architects to combine the services, regions, and computing power of Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, and any other public or private cloud to provide developers with a flexible, global environment for developing backend applications

and services. Control Plane's flexibility shines through in its ability to be deployed across any combination of geographic regions and cloud providers, including AWS, Azure, GCP, or other public and private clouds. Kubernetes clusters hosted anywhere can be easily added to Control Plane can be added. Control Plane elastically optimizes resource consumption to run precisely the resources needed and nothing else. However, this solution is limited to the execution of Docker-specific workloads and is designed for the use of microservice architectures. It, therefore, offers no support for the operation of classic workloads.

Google Anthos [38] is a platform for distributing container workloads across multiple clusters. The clusters are Google Kubernetes Engine (GKE), which, as the name suggests, is based on Kubernetes. Anthos offers the possibility of managing different clusters across different cloud providers, such as AWS, Azure, or GCP. Anthos is limited to container workloads and is, therefore, unsuitable for virtual machine-based workloads and only suitable for newer solutions based on microservices architectures. Furthermore, Anthos needs to include an overview or optimization of costs across cloud providers.

The aspect of risk management for critical data and processes is described by many proven methods in literature and applications [19], such as IT security concepts, failure mode and effects analysis, fault tree analysis, event tree analysis. A software tool has been developed around the CRAMM method (CCTA Risk Analysis and Management Method) developed in the UK by the Central Computer and Telecommunications Agency (CCTA), which is used by NATO and facilitates implementing a risk management system but does not automate it.

Further automated and dynamic solutions for parts of risk management are so-called vulnerability management solutions. There are several solutions available on the market from several manufacturers. However, these only look specifically at the known vulnerabilities of IT infrastructure elements and help to eliminate them. The criticality of data and applications needs to be included. Some well-known open-source solutions in the field of vulnerability management can be used in the project. The Greenbone OpenVAS vulnerability scanner should be mentioned here [29].

3 Methodology

We aim to meet the needs of SMEs and the challenges they face when deploying multi-cloud environments. The use of multi-cloud deployments brings a lot of benefits for SMEs and can leverage significant improvements in the effectiveness and efficiency of the used services. A possible vendor lock-in can be prevented by the use of multi-cloud architectures and the use of best-of-breed services from different providers, which has a strong beneficial impact on SMEs. The overview of the services used is a major challenge when using multi-cloud setups. The distribution of workloads to many different providers makes it hard to keep track of the resources used and the costs generated. Another problem is the fulfillment of governance requirements in using multi-cloud setups, especially for SMEs, since they do not have the capabilities

that bigger companies have. Although there are a lot of benefits, which are very promising, some challenges still need to be tackled in order to make its use feasible for SMEs.

3.1 Research Questions

The setup of multi-cloud environments is increasing the complexity of companies using such an approach. Therefore, one important question is the analysis of the challenges that SMEs face when using multi-cloud setups. We need to understand the needs of SMEs in order to present a framework that not only assists but also solves problems that arise when using multi-cloud environments. There is a lot of research in the field of multi-cloud architectures and their challenges [5, 18]. Therefore, this leads to the first research question that we want to answer before designing the framework.

RQ1: *What are the challenges of multi-cloud environments for SMEs?*

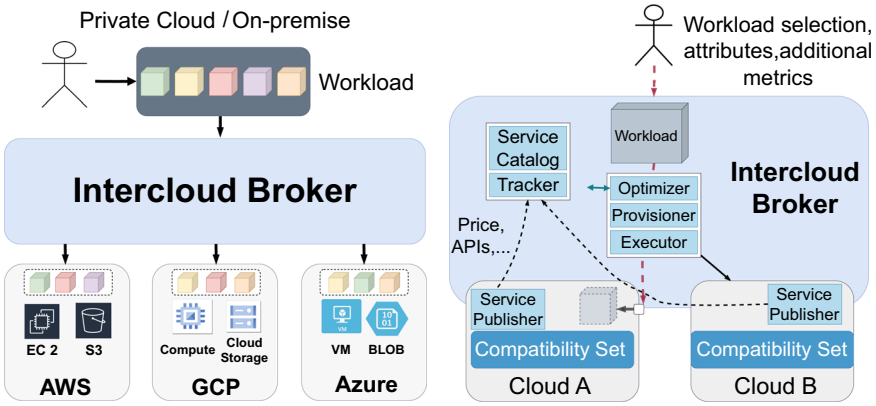
The second question that we want to tackle is the use of Sky computing to implement a framework for multi-cloud environments. Since Sky computing is tackling the challenges that multi-cloud setups face, we first need to understand this paradigm and leverage beneficial knowledge for its implementation. This fact is faced by the second research question.

RQ2: *How can Sky computing be used to improve the efficiency and effectiveness of multi-cloud setups for SMEs?*

Features of Sky computing are the distributed infrastructure and the distribution of individual workloads with the use of resources across various providers. This enables the services to be dynamically scalable, as the resources can be scaled across many cloud providers. This leads to universality in the use of cloud resources, as the placement of workloads is independent of the type of workload and cloud provider. Figure 1a illustrates the overall concept of Sky computing.

The Sky computing concept [30] proposes an additional abstraction layer, which is inserted between the cloud services of the providers (Amazon Web Services, Google Cloud Platform, etc.) and the workloads of the end users (in our case, SMEs). Sky computing aims to complete the abstraction of cloud resources from different providers so that applications and users can access these resources without worrying about where the resources or services are located in the individual clouds. For this reason, the term "*cloud of clouds*" is also used, as the additional abstraction of resources leads to creating a uniform and interoperable cloud and thus includes several individual cloud providers. This abstraction layer is called an intercloud broker [37].

Figure 1b shows the components of an intercloud broker. The Service Catalog captures the instances and services available in each cloud, detailed information about locations that offer them, and the APIs for allocating and accessing them.



(a) Overview of the Sky computing framework (Yang et al., 2023)

(b) Overview of the components of the inter-cloud broker (Yang et al., 2023)

Fig. 1 The Sky computing concept and the intercloud broker

It also stores the long-term pricing for on-demand virtual machines, data storage, egress, and services (typically, these prices stay the same for months). The Service Catalog can provide filtering and search functionality based on information published by the cloud providers, listed by a third party, or collected by the broker.

The tracker tracks spot prices (which may change more frequently, e.g., hourly or daily) and the availability of resources across different cloud providers and locations. This module is of central importance as the prices of cloud providers and the associated services are a decision metric for service placement.

The Optimizer processes the workload requirements and checks the availability of instances and services as well as their prices, which are provided by the Service Catalog and Tracker. This module then calculates the optimal placement of the services. If resource availability and/or price change, the Optimizer can possibly perform a new optimization.

The Provisioner module manages the resources by allocating the resources required for execution. The Optimizer's execution plan allocates the resources accordingly and releases them when each task is completed. The executor manages the application by aggregating the tasks of each workload and executing them based on the resources allocated by the cloud provider and service provider.

Compatibility sets are an essential feature of Sky computing. The focus here is on using existing services and APIs from all cloud providers, intended to offer transparent and standardized options for connecting services without having to reimplement them.

Sky computing sets a reasonable basis for the implementation of our proposed framework since it adds an additional abstraction layer between the users and the cloud service providers. However, it is still important to have a tool and/or component in the Sky computing concept that analyzes the costs that are drawn by the services used. This raises the third research question.

RQ3: *Is it possible to analyze the costs of multi-cloud environments for SMEs?*

We tackle the challenges of cost control and resource supervision through the implementation of the Sky computing paradigm. However, security and governance are very important fields. The information security needs to be analyzed in order to have a strong basis for the analysis of the risks a SME is facing. Governance and regulation are important for SMEs in their interaction with customers and government agencies to fulfill the regulatory requirements that are posed on SMEs, and also, SMEs have a lot of pressure from competitors. Therefore, the analysis of risks and collection of assets, together with the mapping of this collection of data to governance requirements, is important for SMEs. This raises the research question.

RQ4: *How can SMEs keep track of the distributed workloads and make sure security risks are analyzed?*

The research questions raised in this section motivate the conceptual development of our proposed framework called SKY CONTROL. We try to answer the research questions and, along with it, develop a framework that fulfills the needs of SMEs in the use of multi-cloud environments and, at the same time, offers a comprehensive platform for the gathering, analysis, optimization, and control of multi-cloud environments for SMEs. The following section presents the concept of SKY CONTROL and explains the ideas behind the proposed framework.

4 Concept of SKY CONTROL

The distributed nature of multi-cloud environments has many benefits for SMEs but also brings many challenges to the management of workloads on the platforms of the different CSPs. The overview of the costs of the services used by the companies is very challenging and is a major drawback in the choice of such architectures. Another very important point is the overview and analysis of potential security risks and governance of the company's assets. SKY CONTROL is designed to cope with the challenges SMEs face when using multi-cloud deployments. The following sections present the architecture of SKY CONTROL, along with the desired functionalities that are needed to fulfill the requirements of SMEs.

4.1 Architecture of SKY CONTROL

SKY CONTROL is divided into two distinct modules, the **Cost Control** and the **Risk Management** module, that solve different challenges of SMEs. The **Cost Control** module is responsible for the analysis, calculation, and visualization of on-premise and cloud resources. The Cost Control module also takes care of the static analysis of the resources, where the metadata of the resources is analyzed. These static attributes

are information such as the ID of the resource, the characteristics of the (virtual) resource, and hardware capabilities (CPU information, main memory size, etc.). The dynamic analysis focuses on the attributes of the resources, such as the consumption of CPU, main memory, and bandwidth (for Ingres and Egress). Based on this data, the price for the operation is analyzed, and trends for predictions can be calculated by the monitoring component. This, together with the control and planning tool, makes it possible for users to get detailed information on the resources running on-premise as well as in the cloud. Cloud resource monitoring also takes place for different cloud providers distributed over multiple CSPs. The data gathered by the sub-components are visualized and by rendering the data into a human-friendly format making the vast amount of data comprehendible.

The Second module of SKY CONTROL is the **Risk Management** module, which is responsible for the analysis and management of assets of the customer. This module consists of components for the management of assets and the analysis of risks for the assets. The management of the assets starts by collecting information on the assets and gathering detailed insights that can be stored for analysis. Based on the data, the Risk analysis component can give insight into the expected threats and the classification of risks for the individual assets and the whole company. The analysis of the risks, together with the collection of assets, is a crucial basis for governance in complex multi-cloud environments. This is a strong database for audits that are important for the fulfillment of governance requirements for SMEs, and the assistance in gathering this information makes SMEs more compatible with bigger companies. The visualization of the risks and assets is another benefit for information security officers (CIO) and reduces efforts for the preparation of audits and the mitigation of possible risks for the company. Figure 2 presents the architecture of the proposed framework.

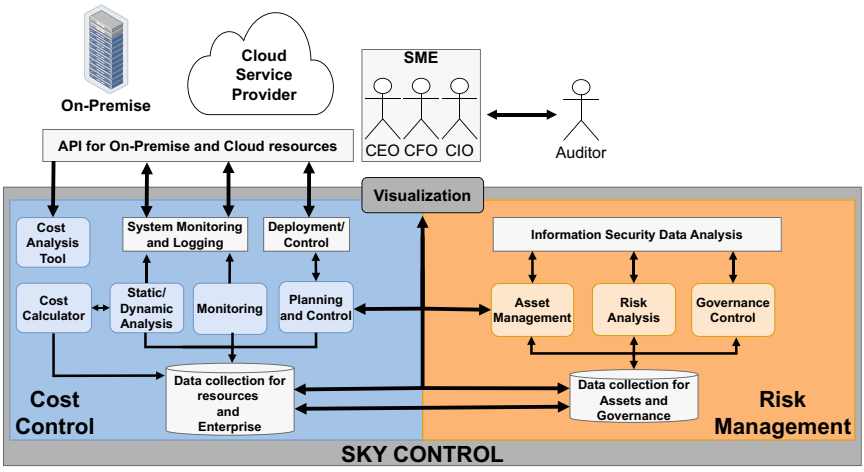


Fig. 2 SKY CONTROL architecture

4.2 Features Planned

We are, therefore, faced with a twofold problem for users of hybrid multi-cloud infrastructures:

1. **Effective cost control and management in hybrid infrastructures**:
   - Recording fixed and ongoing costs and monitoring (preferably in real time) all costs and resource consumption (especially energy) across all systems of a complex hybrid infrastructure.
   - Transparent visualization of costs and dynamic optimization, partially automated.
2. **Security of information assets, information security management**:
   - A systematic and up-to-date assessment of the security/protection level of subsystems and the criticality of the information assets on them.

We want to tackle and solve these core problems in our project. Table 2 shows that these problem areas have a lot in common and are, in principle, amenable to a common solution approach. The core idea is, first and foremost, to create transparency regarding the relevant properties of the entire infrastructure and its components.

The planned system's solution approach is described separately below for the two problem areas of **cost control** and **risk management**. Of course, technical interactions and architectural synergies exist between the two areas. Risk management and cost control should not hinder each other operationally (through conflicting objectives) as far as possible but should cooperate in a common system architecture. We begin with a description of the solution approach for cost control, which also describes the general architecture of the planned system.

Table 2 Comparison of the two problem areas of the project

Module	Cost control	Risk management
Problem area	Cost/efficiency of the IT infrastructure	Data protection and security of the corporate information
Risk	Underutilization of systems; bad investments; cost inefficiency in operations	Data and processes at risk processes on inadequately protected systems
Target criterion	Optimal costs for data storage and processing	Adequate level of protection of systems for critical data and processes
Strategic condition	Transparency of costs and effective cost management	Information security risk management
Primary measure	Mobility of data and processes to the most favorable systems	Control of processes and the mobility of data

4.3 Module: Cost Control

Today's IT infrastructures are faced with the challenge of efficiently planning and implementing mixed systems of on-premise and cloud resources while considering various cost factors. The prevalent nature of hybrid IT infrastructures requires a solution that enables static system planning, keeps an eye on, and minimizes dynamic costs. Our research project aims to develop a comprehensive planning tool for hybrid cloud systems that analyzes static and dynamic resource inventory. A central feature of this solution is optimizing CPU, memory, and traffic resources to minimize the costs of combined on-premise and cloud systems. In addition, a cost calculator for hybrid cloud users will be implemented to provide an aggregated cost overview of cloud and on-premise systems.

Advanced features of our solution include the implementation of dynamic resource control, which makes it possible to adjust resource requirements and costs in real time. By combining static and dynamic analysis and comprehensive cost aggregation, we aim to develop an effective and efficient solution for planning and implementing cost-efficient hybrid cloud systems. The system concept for the cost efficiency subproblem shown in Fig. 2 highlights the new components to be developed as part of this project.

- In the first step, the primary data sources for calculating operational system costs must be collected, whether from cloud hosting or SaaS providers.
- This must be compared with the company's infrastructure costs, including energy costs.
- Particular attention is also paid to the cost optimization of software licenses in the rental models commonly used today.

The data sources are stored and processed in an analysis tool in an interoperable data format. Human expertise is required to define, select, and weigh cost factors. In functional terms, the cost analysis tool has the character of common economic controlling tools but must be redeveloped for the planned area and, in particular, expanded to include the specific development of data sources. According to the state of the art, IT infrastructures are constantly monitored by system monitoring and logging functions (in the case of hybrid systems, these are several different functions). We want to establish the core functions of the system on the basis of these established methods:

- The static/dynamic system analysis provides the relevant data for cost monitoring and relevant data for cost monitoring and makes it available for further processing.
- The central component of the system is the cost monitor and efficiency calculator. This is where the system information is combined with the analyzer's cost information to calculate the ongoing operating costs.
- The third pillar of the system visualizes the costs in granular form for the human planner and provides information for cost optimization.
- A planning and control tool can make cost-optimizing decisions and initiate implementation. This tool enables the execution of various control functions for

hybrid cloud systems (resource scaling) and hosted/on-premises systems (e.g., deployment tools such as Terraform).

In the implementation of our technical/architectural concept for SKY CONTROL, the automation of infrastructure and its properties plays a decisive role. Essential technologies have been available in recent years and can be used effectively in our system. Here tools and concepts from the SASE (Secure Access Service Edge) area described in the next section. Another technological basis for the development of SKY CONTROL is the Sky computing concept [30].

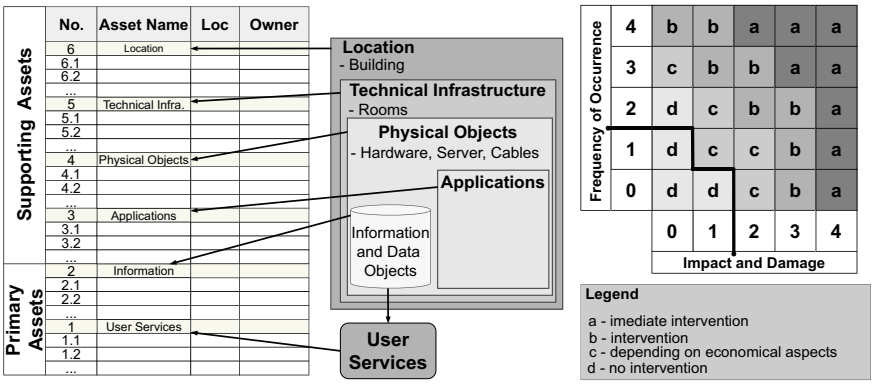
4.4 Module: Risk Management

The last section describes our approach to solving the cost/resource efficiency subproblem. The solution to the second subproblem is structurally similar and will essentially be based on the same system platform but will incorporate other data sources and additional Free and Open Source Software (FOSS) tools from the risk management area (some of which are described in the related works Sect. 2). Information security risk management is an advanced discipline in the field of management systems (see the comprehensive textbook [19]). Within the framework of our planned overall system for effective cost control, the risk control subsystem is intended to make a selective and delimited contribution here, whereby the established methods of risk management are to be applied consistently. The main task of risk control in our context is to ensure that relevant data and its processing are always protected in such a way that risks are appropriately minimized and the operation of cost control—which manages and transports data and processes in different operating environments under certain circumstances—is not hindered. This complex task is divided into three subtasks:

1. Recording the data and processes affected by the SKY CONTROL system as (critical) assets
2. Evaluating the protection level of the environment in which the assets are located, i.e., assessing the current risk for/by an asset
3. Ensuring the appropriate (best possible) level of protection for each asset

As part of the overall concept for SKY CONTROL, the first task falls into the area of information gathering/source development and extends the "cost analysis tool" module accordingly to a "cost/risk analysis tool". Standard methods such as the asset register (see Fig. 3a), for which a suitable interface is to be set up, are suitable for recording assets and their criticality.

Once the assets have been identified, their criticality is usually determined by classifying them in a so-called risk matrix (see Fig. 3b). This usually has to be done manually and is a necessary preliminary step when adapting SKY CONTROL to an operational environment. A suitable interface should be created for this, or existing



(a) Classification of assets in an asset register (Konigs, 2017, p. 256)

(b) Risk matrix with acceptance criteria (Konigs, 2017, p. 254)

Fig. 3 Traditional visualization of the assets and risks [19, pp. 254]

FOSS solutions be used. Consequently, all relevant data for SKY CONTROL is conceptually accessible.

We deploy (cloud, infrastructure, on-premise) systems. Cost factors, assets, the criticality of the assets, and the system can, in principle, begin to make and implement optimization decisions. This is where the last two subtasks of risk management come into play: the current security level of the systems used must be constantly determined and kept up to date. This is a typical vulnerability management task for which we intend to use standard solutions such as OpenVAS [2, 29] as part of the static/dynamic system analysis component. At this point, only gradual progress would be achieved compared to conventional vulnerability management systems—the extension to heterogeneous infrastructures and the simultaneous visibility of costs and risks. In SKY CONTROL, however, we are now tackling the more ambitious third sub-task, namely—as far as possible - the automated assurance of suitable security levels for all assets. A few years ago, this would have been an almost impossible task in heterogeneous environments. However, the concept of Secure Access Service Edge (SASE)—was introduced in 2019 and has since evolved into a solid technology that offers new possibilities for our purposes [15, 21]. SASE is a synergetic combination of techniques for controlling security—i.e., enforcing security policies—in complex, heterogeneous infrastructures.

SASE generally includes the following functionalities:

- Software-Defined Wide Area Network (SD-WAN) is an innovative technology fundamentally changing how companies manage and optimize their wide-area networks. SD-WAN is defined as a virtual wide-area network that enables organizations to use any combination of transport combination of transport services—including MPLS, LTE and 5G as well as broadband—to connect users to network locations securely [15]

- Secure Web Gateway (SWG) is a critical component of modern network security architectures that aims to protect organizations and their users from web-based threats. It acts as an intermediary between users and the Internet by filtering and monitoring web traffic to ensure compliance with corporate and regulatory guidelines.
- Cloud Access Security Broker (CASB) acts as a security checkpoint between cloud service users and cloud service providers and enables internal security policies and compliance enforces internal security policies and compliance regulations.
- Firewall-as-a-Service (FWaaS) is a cloud-based solution that offers firewall functionalities in a flexible and scalable form and fulfills the same basic tasks as a conventional firewall.
- Zero Trust Network Access (ZTNA) is based on the principle of *never trust, always verify*, which means that neither internal nor external networks are automatically trusted. In the SASE context, ZTNA provides protection for users' sessions, regardless of whether they are inside or outside the corporate network.

Figure 4 presents the core components of SASE. SASE enables centralized policy management with distributed enforcement points that are logically close to the entity and enable local decision-making when needed [33]. An example of this is the local enforcement of policies in a branch office using a Customer Premises Equipment (CPE) device or by local agents on managed devices. This architecture significantly improves the flexibility and responsiveness of security measures. Elements of SASE, in appropriate combination, are crucial to addressing the challenges of SKY CONTROL for active, asset, and risk-based security management. Their precise use in the planning and control tool of SKY CONTROL will be the subject of research and development.

SKY CONTROL is a complex project that is adaptable and incorporates many existing and new technologies. Some of them are inspirations for the technology to

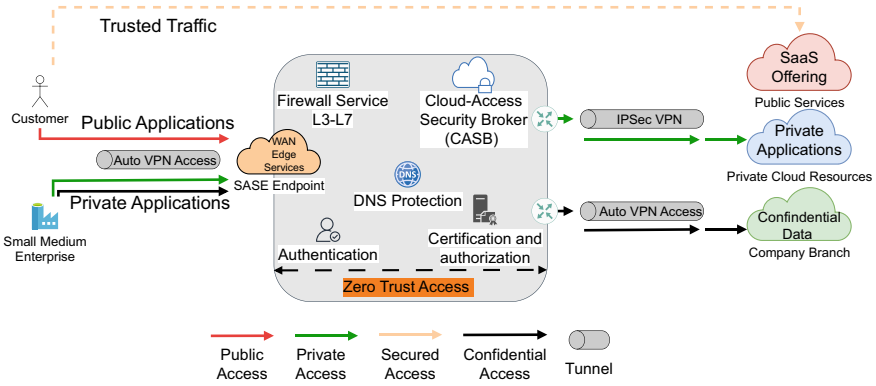


Fig. 4 SASE and its underlying components

be developed in SKY CONTROL, while others (to a lesser extent) can be integrated into the SKY CONTROL architecture as components.

5 Impact of the Project

In summary, the SKY CONTROL solution stands out with its unique features, setting it apart from the published state-of-the-art and existing solutions (see Sect. 2 for the basic concepts and technologies that may be used).

SKY CONTROL takes a holistic view of the value of digital assets and the costs of their storage and processing. This goes far beyond the classic performance analysis, bringing the costs and protection of critical data and its processing into the focus of monitoring, which until now could not be fully monitored by technical solutions.

This applies to hybrid infrastructures, i.e., mixtures of on-premise and multi-cloud infrastructures, which still need to be inspected for an overall overview. For the first time, SKY CONTROL thus enables uniform visibility of previously hidden data and processes to more suitable systems, which has not yet been attempted by any existing solution. The further ambitious goal of optimizing the factors of costs and risks and to implement it in real time wherever possible—by transparently inspecting the parameters under consideration.

In the medium to long term, a solution such as SKY CONTROL could also have a macroeconomic impact. If this were to create a market, cloud offerings could be traded in the manner of stocks at a stock exchange. Even companies themselves could monetize idle resources without evoking security risks. This would allow customers to distribute their workloads at prices dynamically and offers currently traded on the market.

SKY CONTROL reduces potential vendor lock-ins by providing an active and transparent view of the public cloud providers' offerings. Cloud services' dynamic query and cataloging can give SMEs a decisive advantage when rolling out a multi-cloud environment. An overview of services and costs makes cloud services much less risky for SMEs. Optimization of the distribution of workloads makes the operation of cloud instances more efficient, and automated provisioning of workloads reduces the effort required for operation.

SKY CONTROL can reduce the barrier to entry for SMEs into modern IT infrastructure, as it gives them a clear framework for operating their infrastructures and an overview of medium to long-term costs—in other words, greatly improved planning capability. Furthermore, the overview of the distributed workloads helps maintain an overview of one's infrastructure and, thus, for example, to fulfill compliance and other regulatory requirements. The integration of the application into existing frameworks for enterprise architecture (e.g., Enterprise Architect [9], HOPEX [7], etc.) seems possible. It would have the potential to become a holistic cloud tool for SMEs. For research, SKY CONTROL offers an opportunity to find answers to many current questions in cloud computing, especially regarding the setup and efficient administration of multi-cloud environments. One of the goals is to optimize the process

of placing resources across different cloud providers. Methods for the efficient and needs-based allocation of resources are to be investigated and evaluated. A taxonomy for cost factors is also to be developed.

A central problem of Sky computing for which SKY CONTROL intended to provide solutions is the distributed processing of services across different cloud providers. The allocation and distribution of tasks across providers is a significant challenge, and the orchestration and composition of the individual services is also a major challenge. These are to be investigated as part of the research project, and solutions for the operation of the services will be provided.

6 Conclusion and Outlook

In conclusion, the SKY CONTROL framework offers a promising solution for small and medium-sized enterprises (SMEs) facing the complexity of managing multi-cloud environments. By combining cost control with risk management, it addresses two critical challenges: optimizing resource allocation and safeguarding sensitive data. SKY CONTROL's dynamic analysis, real-time cost tracking, and risk assessment tools provide SMEs with greater transparency and control over their IT infrastructure. Additionally, its application of sky computing enhances the flexibility and efficiency of multi-cloud setups, offering SMEs a competitive edge in today's cloud-driven landscape.

In terms of future development, the SKY CONTROL framework opens up several avenues for further research and practical application. As cloud computing continues to evolve, enhancing SKY CONTROL's capabilities through integration with emerging technologies such as AI-driven automation and advanced analytics could further optimize resource allocation and risk management. Additionally, exploring the scalability of SKY CONTROL for larger enterprises or adapting it to specialized industries could expand its utility.

Acknowledgements We thank our partners from Systrade GmbH for their support. We especially thank Andreas Schmidt, Thorsten Luft, Abo El Hage.

References

1. Afzal S, Kavitha G (2019). Load balancing in cloud computing—a hierarchical taxonomical classification. J Cloud Comput Adv Syst Appl 8(1)
2. Aksu MU, Altuncu E, Bicakci K (2019) A first look at the usability of openvas vulnerability scanner. In: Workshop on usable security (USEC)
3. Ardagna D (2015) Cloud and multi-cloud computing: current challenges and future applications. In: 2015 IEEE/ACM 7th International workshop on principles of engineering service-oriented and cloud systems, pp 1–2
4. Barker A, Varghese B, Thai L (2015) Cloud services brokerage: a survey and research roadmap. In 2015 IEEE 8th International conference on cloud computing, pp 1029–1032

5. Baryannis G, Garefalakis P, Kritikos K, Magoutis K, Papaioannou A, Plexousakis D, Zeginis C (2013) Lifecycle management of service-based applications on multi-clouds: a research roadmap. In Proceedings of the 2013 International workshop on multi-cloud applications and federated clouds, MultiCloud '13. Association for Computing Machinery, New York, NY, USA, pp 13–20
6. Baun C, Kunze M, Nimis J, Tai S (2011) Cloud computing. Informatik im Fokus. Springer, Berlin Heidelberg
7. Bouille R (2016) Enterprise architecture: the process mapping with the 2.0 standard using mega hopex
8. Brogi A, Carrasco J, Cubo J, D'Andria F, Ibrahim A, Pimentel E, Soldani J (2014) Seaclouds: seamless adaptive multi-cloud management of service-based applications. In: Castro J, Ayala CP, Giachetti G, Lucena M, Cares C, Franch X, Barcellos MP, Lencastre M, Marín B, Gacitúa R (eds) Proceedings of the XVII Iberoamerican conference on software engineering, CIbSE 2014. Curran Associates, Pucon, Chile, April 23–25, 2014, pp 95–108
9. Doomen P (2016) Introduction to sparxsystems enterprise architect
10. Fortes JAB (2010) Sky computing: when multiple clouds become one. In: Proceedings of the 2010 10th IEEE/ACM International conference on cluster, cloud and grid computing, CCGRID '10. IEEE, USA, p 4
11. Georgios C, Evangelia F, Christos M, Maria N (2021) Exploring cost-efficient bundling in a multi-cloud environment. Simul Model Pract Theory 111:102338
12. Gundu SR, Panem CA, Thimmapuram A (2020) Hybrid it and multi cloud an emerging trend and improved performance in cloud computing. SN Comput Sci 1(5)
13. Hong J, Dreibholz T, Schenkel JA, Hu JA (2019) An overview of multi-cloud computing. In: Barolli L, Takizawa M, Xhafa F, Enokido T (eds) Web, artificial intelligence and network applications. Springer International Publishing, Cham, pp 1055–1068
14. Ijeoma CC, Inyiama CH, Samuel A, Okechukwu OM, Chinedu AD (2022) Review of hybrid load balancing algorithms in cloud computing environment
15. Islam MN, Colomo-Palacios R, Chockalingam S (2021) Secure access service edge: a multi-vocal literature review. In: 2021 21st International conference on computational science and its applications (ICCSA), pp 188–194
16. Iványkó NR, Colvin J (2024) The blueprint for kubernetes compliance (whitepaper)
17. Jamshidi P, Pahl C, Mendonça NC (2016) Pattern-based multi-cloud architecture migration. Softw Pract Exp 47(9):1159–1184
18. Kavitha MG, Radha D (2022) Quality, security issues, and challenges in multi-cloud environment: a comprehensive review. Springer International Publishing, Cham, pp 269–285
19. Konigs H-P (2017) It-Risikomanagement Mit system, 5th edn. Springer Vieweg
20. Li S, Zhou Y, Jiao L, Yan X, Wang X, Lyu MR-T (2015) Towards operational cost minimization in hybrid clouds for dynamic resource provisioning with delay-aware optimization. IEEE Trans Serv Comput 8(3):398–409
21. MacDonald N, Orans L, Skorupa J (2019) The future of network security is in the cloud. Gartner
22. Malawski M, Figiela K, Nabrzyski J (2013) Cost minimization for computational applications on hybrid cloud infrastructures. Futur Gener Comput Syst 29(7):1786–1794. Including Special sections: Cyber-enabled Distributed Computing for Ubiquitous Cloud and Network Services & Cloud Computing and Scientific Applications—Big Data, Scalable Analytics, and Beyond
23. Mell P, Grance T, NI Standards, TUCS Division (2011) The NIST definition of cloud computing. NIST special publication. U.S, Department of Commerce, National Institute of Standards and Technology
24. Monteiro A, Teixeira C, Pinto JS (2014) Sky computing: exploring the aggregated cloud resources—Part II. In: 2014 9th Iberian conference on information systems and technologies (CISTI), pp 1–6
25. Mulder J (2020) Multi-cloud architecture and governance: leverage Azure, AWS, GCP, and VMware vSphere to build effective multi-cloud solutions. Packt Publishing
26. Petcu D (2013) Multi-cloud: expectations and current approaches. In: Proceedings of the 2013 International workshop on multi-cloud applications and federated clouds, ICPE'13. ACM

27. Rimoli M (2022) Gartner forecasts worldwide it spending to reach 4.4 trillion in 2022
28. Seo C, Yoo D, Lee Y (2024) Empowering sustainable industrial and service systems through AI-enhanced cloud resource optimization. Sustainability 16(12)
29. Sharma M, Desai D, Arun ARLP, Rajagopalan N (2024) Openvas vs the rest: unveiling the competitive edge in vulnerability scanners. In: 2024 3rd International conference for innovation in technology (INOCON), pp 1–6
30. Stoica I, Shenker S (2021) From cloud computing to sky computing. In: Proceedings of the workshop on hot topics in operating systems, HotOS '21. Association for Computing Machinery, New York, NY, USA, pp 26–32
31. Thakur A, Chandak M (2022) A review on opentelemetry and http implementation. Int J Health Sci 6:15013–15023
32. Unuvar M, Steinder M, Tantawi AN (2014) Hybrid cloud placement algorithm. In: 2014 IEEE 22nd International symposium on modelling, analysis & simulation of computer and telecommunication systems, pp 197–206
33. van der Walt S, Venter H (2022) Research gaps and opportunities for secure access service edge. In: International conference on cyber warfare and security, vol 17, pp 609–619
34. van Steen M, Tanenbaum A (2017) Distributed systems, 3rd edn. Self-published, Open publication
35. Wei X, Mohaimenur Rahman ABM, Wang Y (2021) Data placement strategies for data-intensive computing over edge clouds. In: 2021 IEEE International performance, computing, and communications conference (IPCCC), pp 1–8
36. Wu Z, Chiang W-L, Mao Z, Yang Z, Friedman E, Shenker S, Stoica I (2024) Can't be late: optimizing spot instance savings under deadlines. In: 21st USENIX symposium on networked systems design and implementation (NSDI 24). USENIX Association, Santa Clara, CA, pp 185–203
37. Yang Z, Wu Z, Luo M, Chiang W-L, Bhardwaj R, Kwon W, Zhuang S, Luan FS, Mittal G, Shenker S, Stoica I (2023) SkyPilot: an intercloud broker for sky computing. In: 20th USENIX symposium on networked systems design and implementation (NSDI 23). USENIX Association, Boston, MA, pp 437–455
38. Yeboah-Ofori A, Jafar A, Toluwaloju T, Hilton I, Oseni W, Musa A (2024) Data security and governance in multi-cloud computing environment

Computing Political Power: The Case of the Spanish Parliament

Aitor Godoy, Ismael Rodríguez, and Fernando Rubio

Abstract In this paper, we present a series of algorithms to calculate the power of each political party in a parliamentary system. For this purpose, it is necessary to calculate the proportion of parliamentary majorities in which their participation is necessary. The usefulness of the proposed methods is illustrated with a real case study: the Spanish electoral system. For this system, we analyze all the elections that have taken place since the establishment of democracy in the country. For each electoral process, we compare the power that each party would have if the allocation of deputies were proportional to the number of votes, and the real power it has with the current electoral system. The results obtained contradict intuitions that the Spanish population usually has about its own electoral system.

Keywords Electoral systems · Power measures · Counting problems · Approximation methods

1 Introduction

There are a multitude of different electoral systems (see [4] for a detailed classification). Some systems are purely presidential, meaning that only the president is elected in them (see e.g. [15]), either in a single round or in a two-round system where only the two most voted candidates go to the second round. In our case, we will focus on parliamentary electoral systems (see e.g. [9, 18]), where voters do not

A. Godoy (✉) · I. Rodríguez · F. Rubio
Facultad Informática, Universidad Complutense, Madrid, Spain
e-mail: aitorgod@ucm.es

I. Rodríguez
e-mail: isrodrig@ucm.es

F. Rubio
e-mail: rubiod@ucm.es

I. Rodríguez · F. Rubio
Instituto de Tecnología del Conocimiento, Universidad Complutense, Madrid, Spain

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_5

just elect a president but elect a parliament consisting of a certain number of representatives, being each representative assigned to a political party. Subsequently, this parliament must pass laws (see e.g. [7]) or even elect the prime minister. In order to do so, majorities must be formed by means of pacts between the political parties.

Within parliamentary systems, there are different ways of electing representatives. In some cases, each constituency elects only one representative, where the elected representative will be the one who obtains the most votes in the constituency. In other cases, each constituency is assigned a certain number of representatives, which are distributed among the different political parties depending on the number of votes obtained by each party, following some mathematical law (such as the D'Hont law [12] or the Webster method [1]). In other cases, the allocations of seats by constituencies are complemented by additional allocations at the global level, which may seek to favor the most voted party or to achieve a proportional distribution of representatives with respect to the percentage of global votes obtained by each party.

In any case, once the parliament is configured, each political party has a certain amount of power. This power depends fundamentally on how many possible majorities the party is part of—or more specifically, on how many majorities the party is actually relevant in. For example, if we have four parties (P1, P2, P3, P4) with a number of representatives of 15, 10, 7 and 5, respectively, then 19 representatives are needed to obtain a parliamentary majority. This can be achieved by a coalition of P1 with any other party, but also by a coalition of P2, P3 and P4. There are other possible coalitions (e.g., P1-P2-P3) but in such cases some party is irrelevant to form the coalition, since the majority would be obtained equally without that party (e.g., with P1-P2 or with P1-P3 but not with P2-P3). Thus, in this example, party P1 forms a relevant part of 6 majorities, while the rest of the parties form a relevant part of only 2 possible absolute majorities. Thus, although the party P4 has half as many seats as P2, its real power to form majorities is exactly the same.

However, if the weight of each party were not given by its number of parliamentarians, but by the number of votes it actually obtained, the possible majorities could change significantly. In fact, a central issue in the design of any electoral system is to determine how proportional the system is (see e.g. [3]). This means deciding whether the number of representatives obtained is a linear function with respect to the number of votes or not, and if not (which is the most common case), in which direction it deviates from linearity and how much. Given this scenario, the obvious question is: Which type of party benefits from each electoral system? That is, which type of party increases more its power to form majorities if such electoral system is considered instead of simply counting its number of votes?

In order to choose the most suitable algorithm to (exactly or approximately) solve a given problem, first of all its computational complexity should be identified (see e.g. [6, 16, 20]), including its approximability if it is possible (see e.g. [10, 13, 17]), as even the approximation hardness of a problem has been observed to affect the suitability of solving it by means of a genetic algorithm [14]. Although it has been shown that calculating the electoral power of each party to form majorities is a #P-hard problem (see [8]), in practice it is possible to solve the problem when the number of parties and seats is relatively small (as is usual in most parliaments). However,

when we want to calculate the power to form majorities from the number of votes, the size of the problem can make it convenient to use approximation algorithms to obtain the solution to the problem, due to the #P-hard nature of the problem. In this paper, we show several algorithms, including an approximation algorithm for larger problems, and apply them to analyze a real case study. In particular, we will analyze the case of the Spanish electoral system, studying quantitatively which type of political parties benefit the most from the system.

The rest of the paper is structured as follows. In the next section, we introduce the main concepts on how to measure the power of each political party. Then, in Sect. 3 we show the basic scheme of our algorithms. Afterward, in Sect. 4 we analyze the case study of the Spanish electoral system. Finally, in Sect. 5 we present our conclusions.

2 Power Measures

In a weighted voting game, the weights of each agent (e.g., its number of representatives in a parliament) are not always the best way to measure their power. For example, in a majority weighted voting game with three agents where the weight of two of them is 4 and the weight of the remaining agent is 1, the only way any agent can form a majority is by making a coalition with another agent. Hence, the agent with weight 1 can be part of the same number of coalitions as the other agents with weight 4. In this sense, each agent has exactly the same power as the others, despite them having different weights.

Power indexes exist to give a mathematical formulation to what really is the influence of a player in a weighted voting game or in a coalition game. They have been greatly studied in the literature before (see e.g. [2, 5, 19]). Let $(A = \{a_1, \ldots, a_n\}, v)$ be a simple coalition game where $\{a_1, \ldots, a_n\}$ are the weights of the agents and v is a function that indicates if the coalition succeeds. The power indexes used in this paper are the following.

Definition 1 (*Shapley-Shubik index*) The Shapley-Shubik index for i is the number of different orders of arrival in which player i can join a coalition, where we say that player i joins a coalition if its arrival transforms a losing coalition into a winning coalition. Then, the Shapley-Shubik index for i, φ_i is defined by:

$$\varphi_i = \sum_{S \subseteq A \setminus \{i\}} (|S|!(|A| - |S| - 1)!(v(S \cup \{i\}) - v(S))) \tag{1}$$

Definition 2 (*Raw Banzhaf index, Banzhaf-Coleman index, Absolute Banzhaf index*) The raw Banzhaf index for i is the number of coalitions where i is pivotal (we will say that i is pivotal in a coalition if the coalition is successful when i is included but it is not successful otherwise). The Banzhaf-Coleman index and the Absolute Banzhaf index are simply the raw Banzhaf index divided by the total amount of agents that are pivotal and the raw Banzhaf index divided by the total amount of coalitions (2^{n-1}),

respectively. The raw Banzhaf index (β_i'), the Banzhaf-Coleman index, (β_i), and the Absolute Banzhaf index (β_i''), are defined as follows:

$$\beta_i' = |\{S \subseteq A \setminus \{i\} | v(S) = 0 \wedge v(S \cup \{i\}) = 1\}| \tag{2}$$

$$\beta_i = \frac{\beta_i'}{\sum_{j=1}^{n} \beta_i'} \tag{3}$$

$$\beta_i'' = \frac{\beta_i'}{2^{n-1}} \tag{4}$$

3 Implementation

Although the calculation of the raw Banzhaf index is a #P-hard problem, it is relatively straightforward to provide practical algorithms in case the number of political parties to be considered is relatively low. This is the case when considering only parties that have obtained representation in the parliament. For example, in the case of the Spanish parliament that we will analyze in the following section, the number of political parties that obtain representation usually varies between 10 and 20. In this case, it is sufficient to analyze the 2^{20} possible pacts that can be formed between the parties with parliamentary representation. However, in order to compare the power obtained by these parties in the parliament with the power they would obtain by considering their votes, it is necessary to analyze all the parties, not only those that obtain parliamentary representation. In this case, the number of parties to be considered would be around 50, which makes it unfeasible to analyze the 2^{50} cases.

In order to deal with all political parties running for election, we propose to use a pseudo-polynomial algorithm similar to the one used to solve the knapsack problem (see e.g. [11]). As in the knapsack problem, we use a dynamic programming technique, although a single-dimension array is used in this case. In each position j of the array, we store the number of different ways in which the votes of the parties can combine to sum to exactly j. This information will be updated by iteratively counting the values of cells $j - x$ for all x values denoting the numbers of votes of some party. The concrete pseudocode of the algorithm is the following, where w_i denotes the weight (i.e., number of votes, or number of deputies) of each party, and S denotes the set of weights of all the parties:

1. target = $\lceil \sum S/2 \rceil - w_i$
2. let count, count' be arrays of size $\sum S - w_i + 1$
3. set count[0] = 1 and count[j] = 0 for all other j
4. for every x in $S \backslash \{w_i\}$

count' = count
for every j in $\{1, \ldots, \sum S - w_i\}$ with $j \geq x$

count'[j] = count[j] + count[j - x]

count = count'

5. sol = $\sum_{j=\text{target}}^{\sum S - w_i}$ count[j]

Note that the computational complexity of this solution is $\mathcal{O}(n \sum S)$. Thus, it is a reasonable solution as long as the weights (i.e., votes) of each party are not too large. In particular, this method is feasible for a few million votes, but it would not be so good in electoral systems such as the Indian one with a billion voters. For such larger cases, we have also created a heuristic algorithm that provides a good approximation using a sampling technique. More specifically, our algorithm counts how many of the k randomly generated agreements reach absolute majority, and then normalizes with respect to the maximum number of possible agreements to extrapolate the actual number of them over the entire agreement space:

1. valid = 0
2. loop k times:[1]
   (a) target = $\sum S/2$
   (b) make a Boolean array of size n where:
      – pick[i] = True
      – $\forall j \neq i$ pick[j] = True with 1/2 prob else False
   (c) total = $\sum_i \{s_j \mid pick[j]\}$
   (d) if total $>$ target then valid = valid + 1
3. sol = $2^{|S|-1} * \frac{\text{valid}}{k}$

4 Case Study

In this section, we consider a real case study in which we will apply our algorithms to calculate how the electoral system affects the power of each political party. More specifically, we focus on the case of national elections in Spain. For each of the electoral processes that have taken place in Spain since the implementation of Democracy during the 1970s, we will analyze the electoral power of each party in three ways: (1) considering the number of votes they obtained in the elections; (2) considering the number of seats they obtained in the parliament; (3) calculating the ratio between the two previous values. In the first two cases, we will calculate the

[1] k is an arbitrary number, and the bigger it is, the more accurate results we will get.

Banzhaf-Coleman index for each political party, using the algorithms described in the previous section and the formula described using the raw Banzhaf index. For the third step it will be sufficient to make the division between both values (the ratio by seats and the ratio by votes). Thus, a ratio greater than 1 will indicate that the party has been favored by the electoral system (as its relative power is greater considering deputies than considering votes), while a ratio lower than 1 will indicate that it has been disadvantaged.

In the Spanish electoral system 350 deputies are elected. There is a constituency for each of the 50 provinces of the country. In addition to these 50 constituencies, there are another 2 corresponding to the autonomous cities of Ceuta and Melilla. The number of deputies elected in each constituency depends directly on the population of the constituency. However, even the smallest provinces are guaranteed to have at least two representatives (except for the autonomous cities of Ceuta and Melilla, with only one representative each). Thus, very small provinces may be over-represented in parliament. On the other hand, within each constituency, the D'Hont law is used to distribute the deputies taking into account the number of votes obtained by each political party. As is well known, this system of distribution slightly rewards the majority parties within the constituency, with the impact of the prize being smaller as the constituencies become larger. In fact, in small constituencies it is almost impossible for minority parties to obtain representation.

The majority belief among Spanish voters is that the current electoral system greatly favors those nationalist parties that only run in a few constituencies. For example, parties such as ERC or CiU (or more recently Junts) only run in Catalan constituencies, while EAJ-PNV or Bildu only run in Basque constituencies. In fact, more recently, new regionalist political parties have emerged that run in a single constituency, in order to try to gain high influence for their territories. Examples of this style are PRC or TeruelExiste.

The most relevant data on the electoral results and the power of each party in each of the electoral processes are summarized in the tables shown at the end of the paper. For each party, it shows (in this order) the number of votes it obtained, the relative power that such votes conferred to form voting majorities, the number of deputies obtained, the relative power that such number of deputies conferred to form parliamentary majorities, and the ratio between both powers. That is, the last column will be greater than 1 if the electoral system has favored it, or less than 1 if it has harmed it. Moreover, those political parties that only run in a few constituencies are marked in blue, while those with a national scope remain in black.

The study of the power of each party according to the number of votes obtained has been carried out taking into account all the political parties that ran in each electoral process, regardless of whether they obtained parliamentary representation or not. However, given that for parties without parliamentary representation the ratio between both powers will always be 0, we have preferred to show only the data of the parties that obtained parliamentary representation. The only exception is that sometimes we also include data from some parties (such as CDS or very specially PACMA) that despite not having obtained representation, they did obtain a much higher number of votes than other parties that did obtain deputies. The purpose of

showing these data is simply to illustrate that, indeed, this situation usually exists with the Spanish electoral system.

Finally, we would like to note that when calculating the possible pacts, we assume that any political party can pact with any other, regardless of its ideology. We have made this decision because ideological issues are independent of the electoral model itself, which is what we are really evaluating. Besides, the decision on who could pact with whom would not be objective. In fact, there have been investiture pacts between parties that, in principle, were very distant ideologically.

4.1 Analysis of Results

Before analyzing the general rules that can be drawn, it is worth commenting separately on the 1982, 1986, 2000, and 2011 elections. In those elections, the winning party obtained an absolute majority. Thus, all possible government coalitions went through that winning party. That is, its relative index of power in terms of deputies was 1 and that of the rest of the parties was 0. In the 1982 elections, the winning party was also very close to obtaining an absolute majority of the votes, so its relative index of power per votes was also close to 1. In other words, the winning ratio remained near 1. On the other hand, in the electoral processes of 1986, 2000, and 2011 the winning party did not obtain an absolute majority of votes. Thus, in those elections the winning party obtained a significant gain of power by using the electoral system, while the rest of the parties obtained a 0 ratio, since their power to form majorities after the elections was nil. The situation was very similar in 1989, where the winning party did not obtain an absolute majority but came within one deputy.

From such cases it is easy to infer that the electoral system favors the majority party in those cases in which the amount of votes obtained is close to the absolute majority without reaching it (obtaining winning ratios of 1.52 in 1986, 1.72 in 2000, or 1.44 in 2011) and disadvantages the rest, which are left with a 0 improvement ratio.

If we turn to the more general case, it is striking that, in absolutely all electoral processes, the majority party obtains a winning ratio greater than 1. However, it undergoes very significant variations, from as high as 1.02 in 1982 to 2.59 in 1979. That is, the winning party is always favored, but the ratio is not usually very high, as it has only been greater than 2 in two electoral processes.

The fact that the majority party is favored by the electoral system is to be expected. In fact, the popular belief is that the system favors the largest parties. However, this belief is in clear contradiction with the second general conclusion that can be drawn: the second most voted party always worsens its power ratio in all electoral processes. That is, the system does not reward the *largest* parties, since the second largest party is always punished with ratios always strictly below 1 and, in nearly half of the electoral processes, below 0.5. In fact, if we calculate the average ratio obtained during the 16 electoral processes by the second most voted party, it is 0.48, while that of the first

party is 1.51. That is, the system rewards (on average) with more than 50% of extra power to the winner, while the power of the second most voted party is halved.

When we move on to analyze the case of parties that only run in a few constituencies, the situation is more interesting. The two most paradigmatic cases are EAJ-PNV (in the Basque Country) and CiU (in Catalonia), which have obtained representation in all electoral processes, although under different names.[2] These parties are perceived by society as parties that are rewarded by the electoral system above the rest. However, if we analyze their electoral results, we can see that EAJ-PNV has only obtained ratios higher than 1 in 8 occasions, while CiU has only obtained positive rewards in 6 out of 16 electoral processes. That is to say, the electoral system has harmed them on more occasions than it has benefited them. However, these results also admit another alternative view, because the variance of their ratios is much higher than that of other political parties. In fact, in several electoral processes their relative power has increased to a very high degree, reaching its peak in 1996, where PNV's improvement ratio was 7.17 and that of CiU was 66.25. In other words, we can conclude that, in general, the electoral system does not benefit them (in fact, there are more occasions in which it harms them), but it is true that the electoral system favors them in certain situations. In fact, if we calculate the average winning ratio, we obtain 1.28 for EAJ-PNV and 4.87 for CiU, because when they obtain profit, they obtain very large profits.

The situation of other parties that only run in a few provinces is worse than in the case of EAJ-PNV and CiU. For example, ERC has only obtained ratios above 1 in 6 out of 16 occasions, obtaining an average ratio of 0.71, while the left-independence voting spectrum (HB, Amaiur, Bildu) in the Basque Country has obtained ratios above 1 in only 3 occasions, with an average ratio of 0.82. Something similar happens with BNG in Galicia, which has only obtained positive ratios 3 times. Slightly better have been CC's results in the Canary Islands, with 5 positive ratios in the 11 elections it has contested, with an average ratio of 1.33.

Another widespread belief is that national minority parties are always disadvantaged by the Spanish electoral system. This statement is almost true, since in most situations the ratios obtained by this type of parties (PCE, IU, CDS, UPyD, Cs, Podemos, UP, Vox, Sumar, or PACMA) are not only less than 1, but usually even less than 0.3. However, in a few occasions, both the third and fourth national parties can obtain ratios strictly higher than 1. In fact, in 2015 and in 2023 both the third and the fourth national parties obtained at the same time ratios strictly greater than 1 (Cs and Podemos in 2015, Vox and Sumar in 2023), while in the first electoral process of 2019 the fourth party (UP) obtained 1.14, and in the second electoral process of 2019 the third party (Vox) obtained 1.12.

[2] CiU's ideological space has run for the different elections under different names such as PDPC, CiU, DiL, CDC, or Junts.

5 Conclusions

Perceptions of the strengths and weaknesses of electoral systems are often overly influenced by personal biases. In order to objectively analyze the influence of an electoral system, it is necessary to be able to accurately, objectively and unambiguously compute the results of the system. Unfortunately, such analyses are not frequent in the literature, partly due to the computational difficulty of the problem.

In this paper, after developing specific algorithms to calculate the power of parties to form majorities, we have been able to analyze a real case study: the Spanish electoral system. Our computational study has allowed us to confirm some common beliefs (such as that the most voted party usually benefits from the electoral system), but it has also allowed us to discard other widely held beliefs. In particular, we have shown that the large national parties do not benefit from the system, because although the first party always benefits, the second party always loses. On the other hand, we have also ruled out the belief that the system favors nationalist parties that only run in a few constituencies. Our computational study has found that, while it is true that on certain occasions the system gives them an enormous advantage, in most cases they are disadvantaged. That is, it is true that sometimes it favors them a lot, but it is also true that most of the time it does not favor them.

Note that in our computational study, when calculating the relative power of each party, we have only taken into account the votes and the deputies obtained by them. However, we have not introduced restrictions related to *impossible* pacts due to the fact that the ideology of the corresponding parties may be completely incompatible. We have preferred not to include ideological aspects for two reasons. First, there is no objective way to determine which parties are compatible with each other (in fact, in the electoral processes studied there are several cases of coalitions that include opposing parties on the left-right ideological axis). Second, and more importantly, such ideological aspects are outside the electoral system itself. That is to say, whether or not a party can ideologically agree with another party does not depend on how the seats are distributed, but on its political affinity.

1977 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
UCD	6310391	0.326079	165	0.725624	2.225297
PSOE	5371866	0.183181	118	0.048375	0.264082
PCE	1709890	0.139601	20	0.048375	0.346522
FPAP	1504771	0.113382	16	0.048375	0.426654
PDPC	514647	0.041827	11	0.048375	1.156561
EAJ-PNV	296193	0.019386	8	0.035525	1.832493
PSP-US	816582	0.056544	6	0.012850	0.227251
UCDCC	172791	0.011761	2	0.011338	0.963988
EE	61417	0.004215	1	0.005291	1.255149
CE EC	143954	0.009843	1	0.005291	0.537550
CIC	29834	0.002049	1	0.005291	2.582248
CAIC	37183	0.002553	1	0.005291	2.072102

1979 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
UCD	6268593	0.317087	168	0.821259	2.590013
PSOE	5469813	0.201875	121	0.026599	0.131760
PCE	1938487	0.195861	23	0.026599	0.135805
CD	1060330	0.063008	9	0.026599	0.422151
CiU	483353	0.034845	8	0.026599	0.763363
EAJ-PNV	296597	0.020429	7	0.026392	1.291905
PA	325842	0.022507	5	0.020596	0.915093
HB	172110	0.011897	3	0.006106	0.513290
UPC	58953	0.004067	1	0.003208	0.788909
UPN	28248	0.001948	1	0.003208	1.646653
PUN	378964	0.026202	1	0.003208	0.122449
PAR	38042	0.002624	1	0.003208	1.222682
ERC-FN	123452	0.008522	1	0.003208	0.376484
EE	85677	0.005912	1	0.003208	0.542727

1982 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	10127392	0.975734	202	1.000000	1.024869
AP-PDP	5548107	0.001793	107	0.000000	0.000000
CiU	772726	0.001793	12	0.000000	0.000000
UCD	1425093	0.001793	11	0.000000	0.000000
PNV-EAJ	395656	0.001793	8	0.000000	0.000000
PCE	846515	0.001793	4	0.000000	0.000000
CDS	604309	0.001793	2	0.000000	0.000000
HB	210601	0.001768	2	0.000000	0.000000
ERC	138118	0.001501	1	0.000000	0.000000
EE-IPS	100326	0.001146	1	0.000000	0.000000

1986 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	8901718	0.656149	184	1.000000	1.524043
AP	5247677	0.053833	105	0.000000	0.000000
CDS	1861912	0.053833	19	0.000000	0.000000
CiU	1014258	0.053832	18	0.000000	0.000000
IU	935504	0.053814	7	0.000000	0.000000
PNV	309610	0.020370	6	0.000000	0.000000
HB	231722	0.014278	5	0.000000	0.000000
EE	107053	0.006330	2	0.000000	0.000000
CG	79972	0.004712	1	0.000000	0.000000
UV	64403	0.003788	1	0.000000	0.000000
CAIC	65664	0.003863	1	0.000000	0.000000
PAR	73004	0.004298	1	0.000000	0.000000

1989 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	8115568	0.532103	175	0.997078	1.873844
PP	5285972	0.099711	107	0.000243	0.002442
CiU	1032243	0.058495	18	0.000243	0.004163
IU	1858588	0.099706	17	0.000243	0.002442
CDS	1617716	0.099283	14	0.000243	0.002452
EAJ-PNV	254681	0.011684	5	0.000243	0.020839
HB	217278	0.009898	4	0.000243	0.024600
PA	212687	0.009682	2	0.000243	0.025148
UV	144924	0.006554	2	0.000243	0.037148
EA	136955	0.006191	2	0.000243	0.039326
EE	105238	0.004752	2	0.000243	0.051242
PAR	71733	0.003236	1	0.000243	0.075238
AIC	64767	0.002922	1	0.000243	0.083339

1993 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	9150083	0.325946	159	0.500000	1.533997
PP	8201463	0.211562	141	0.166667	0.787792
IU	2253722	0.211408	18	0.166667	0.788366
CiU	1165783	0.057346	17	0.166667	2.906329
EAJ-PNV	291448	0.021571	5	0.000000	0.000000
CC	207077	0.015014	4	0.000000	0.000000
HB	206876	0.014999	2	0.000000	0.000000
ERC	189632	0.013702	1	0.000000	0.000000
PAR	144544	0.010372	1	0.000000	0.000000
EA-EE	129293	0.009260	1	0.000000	0.000000
UV	112341	0.008031	1	0.000000	0.000000
CDS	414740	0.033905	0	0.000000	0.000000

1996 elections					
Party	Votes	V.Power	Seats	Se.Power	S/V power
PP	9716006	0.329087	156	0.453879	1.379206
PSOE	9425678	0.324827	141	0.171655	0.528452
IU	2639774	0.324827	21	0.171655	0.528452
CiU	1151633	0.002130	16	0.141112	66.245234
EAJ-PNV	318951	0.002130	5	0.015272	7.169398
CC	220418	0.002077	4	0.015272	7.352791
BNG	220147	0.002076	2	0.009163	4.413040
HB	181304	0.001919	2	0.009163	4.775024
ERC	167641	0.001827	1	0.004276	2.340796
EA	115861	0.001298	1	0.004276	3.295044
UV	91575	0.001060	1	0.004276	4.032584

2000 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PP	10321178	0.580909	183	1.000000	1.721439
PSOE	7918752	0.064889	125	0.000000	0.000000
CiU	970421	0.064889	15	0.000000	0.000000
IU	1263043	0.064889	8	0.000000	0.000000
EAJ-PNV	353953	0.036507	7	0.000000	0.000000
CC	248261	0.023962	4	0.000000	0.000000
BNG	306268	0.030327	3	0.000000	0.000000
PA	206255	0.019579	1	0.000000	0.000000
ERC	194715	0.018426	1	0.000000	0.000000
ICV	119290	0.011146	1	0.000000	0.000000
EA	100742	0.009375	1	0.000000	0.000000
CHA	75356	0.006990	1	0.000000	0.000000

2004 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	11026163	0.402820	164	0.515777	1.280414
PP	9763144	0.139408	148	0.105583	0.757365
CiU	835471	0.084544	10	0.099515	1.177073
ERC	652196	0.055164	8	0.080097	1.451980
EAJ-PNV	420980	0.045152	7	0.066748	1.478290
IU	1284081	0.135137	5	0.042476	0.314316
CC	235221	0.021073	3	0.035194	1.670071
BNG	208688	0.018940	2	0.021845	1.153381
CHA	94252	0.008934	1	0.010922	1.222525
EA	80905	0.007530	1	0.010922	1.450425
NABAI	61045	0.005649	1	0.010922	1.933535

2008 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	11289335	0.412045	169	0.560256	1.359698
PP	10278010	0.136917	154	0.096154	0.702276
CiU	779425	0.110312	10	0.096154	0.871653
EAJ-PNV	306128	0.028651	6	0.093590	3.266506
ERC	298139	0.028065	3	0.044872	1.598872
IU	969946	0.130017	2	0.026923	0.207073
BNG	212543	0.020337	2	0.026923	1.323836
CC-PNC	174629	0.017534	2	0.026923	1.535495
UPyD	306079	0.028648	1	0.014103	0.492273
NABAI	62398	0.006577	1	0.014103	2.144291

2011 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PP	10866566	0.692979	186	1.000000	1.443044
PSOE	7003511	0.045191	110	0.000000	0.000000
CiU	1015691	0.045172	16	0.000000	0.000000
IU-Ver	1686040	0.045191	11	0.000000	0.000000
Amaiur	334498	0.019368	7	0.000000	0.000000
UPyD	1143225	0.045191	5	0.000000	0.000000
EAJ-PNV	324317	0.018621	5	0.000000	0.000000
ERC	256985	0.014246	3	0.000000	0.000000
BNG	184037	0.009963	2	0.000000	0.000000
CC-NC	143881	0.007688	2	0.000000	0.000000
IV-Eq-Com	125306	0.006685	1	0.000000	0.000000
PACMA	102144	0.005432	0	0.000000	0.000000

2015 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PP	7236965	0.344767	123	0.428364	1.242473
PSOE	5545315	0.180118	90	0.167418	0.929494
Podemos	3198584	0.119143	42	0.151709	1.273341
Cs	3514528	0.143175	40	0.143273	1.000680
ECP	929880	0.036087	12	0.021091	0.584452
Compr-Pod	673549	0.025918	9	0.017600	0.679067
ERC	601782	0.023145	9	0.017600	0.760420
DyL	567253	0.021823	8	0.015855	0.726502
Marea	410698	0.015526	6	0.012655	0.815076
EAJ-PNV	302316	0.011558	6	0.012655	1.094856
IU	926783	0.035973	2	0.004800	0.133435
Bildu	219125	0.008346	2	0.004800	0.575109
CC-PNC	81917	0.003153	1	0.002182	0.692055
PACMA	220369	0.008394	0	0.000000	0.000000

2016 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PP	7941236	0.445271	137	0.482317	1.083199
PSOE	5443846	0.166494	85	0.142073	0.853321
UP	3227123	0.154739	45	0.142073	0.918144
Cs	3141570	0.150957	32	0.133537	0.884602
ECP	853102	0.013740	12	0.020732	1.508876
Compr-Pod	659771	0.013006	9	0.017683	1.359619
ERC	632234	0.012696	9	0.017683	1.392843
CDC	483488	0.010684	8	0.016463	1.540870
Marea	347542	0.007261	5	0.010976	1.511560
EAJ-PNV	287014	0.006179	5	0.010976	1.776210
Bildu	184713	0.003896	2	0.003659	0.939069
CC-PNC	78253	0.001804	1	0.001829	1.013824
PACMA	286702	0.006174	0	0.000000	0.000000

2019A elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	7513142	0.338596	123	0.437233	1.291313
PP	4373653	0.161041	66	0.145744	0.905016
Cs	4155665	0.153609	57	0.145744	0.948801
UP	2897419	0.082688	33	0.094577	1.143780
VOX	2688092	0.071412	24	0.051167	0.716505
ERC	1020392	0.043468	15	0.046470	1.069068
ECP	615665	0.028236	7	0.017151	0.607410
Junts	500787	0.022402	7	0.017151	0.765581
EAJ-PNV	395884	0.017574	6	0.014731	0.838232
Bildu	259647	0.011447	4	0.011457	1.000885
NA+	107619	0.004722	2	0.004626	0.979597
EC-UP	238061	0.010476	2	0.004626	0.441557
CC-PNC	137664	0.006044	2	0.004626	0.765340
Compromís	173821	0.007643	1	0.002348	0.307260
PAC	52266	0.002294	1	0.002348	1.023612
PACMA	328299	0.014505	0	0.000000	0.000000

2019N elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PSOE	6792199	0.327092	120	0.360294	1.101507
PP	5047040	0.196526	89	0.194782	0.991123
VOX	3656979	0.174015	52	0.194545	1.117979
UP	2381960	0.084539	26	0.079813	0.944094
ERC	874859	0.028048	13	0.037129	1.323758
Cs	1650318	0.060439	10	0.026826	0.443849
Junts	530225	0.018787	8	0.021321	1.134847
ECP	549173	0.019463	7	0.018631	0.957258
EAJ-PNV	379002	0.013485	6	0.016061	1.191022
Bildu	277621	0.009882	5	0.013372	1.353202
CC-NC	124289	0.004426	2	0.005313	1.200322
CUP	246971	0.008792	2	0.005313	0.604238
EC-UP	188231	0.006706	2	0.005313	0.792250
+PaÃƒÂs-Eq	330345	0.011760	2	0.005313	0.451760
NA+	99078	0.003529	2	0.005313	1.505300
TeruelE	19761	0.000704	1	0.002666	3.786117
MÃƒÂL°S COM	176287	0.006280	1	0.002666	0.424486
BNG	120456	0.004290	1	0.002666	0.621435
PRC	68830	0.002452	1	0.002666	1.087123
PACMA	228856	0.008147	0	0.000000	0.000000

2023 elections					
Party	Votes	V.Power	Seats	S.Power	S/V power
PP	8160837	0.282760	137	0.414691	1.466584
PSOE	7821718	0.206616	121	0.177559	0.859369
VOX	3057000	0.123601	33	0.157895	1.277451
Sumar	3044996	0.121086	31	0.138230	1.141584
ERC	466020	0.055362	7	0.028340	0.511902
Junts	395429	0.044981	7	0.028340	0.630050
Bildu	335129	0.037584	6	0.023713	0.630941
EAJ-PNV	277289	0.031191	5	0.015616	0.500658
BNG	153995	0.015990	1	0.005205	0.325546
CC	116363	0.012172	1	0.005205	0.427664
UPN	52188	0.005471	1	0.005205	0.951400
PACMA	169237	0.017352	0	0.000000	0.000000

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgements This work has been partially supported by Spanish projects PID2019-108528RB-C22 and PID2023-149943OB-I00.

References

1. Balinski ML, Young HP (1980) The Webster method of apportionment. Proc Natil Acad Sci 77(1):1–4
2. Banzhaf J (1965) Weighted voting doesn't work: a mathematical analysis. Rutgers Law Rev 19(2):317–343
3. Benoit K (2000) Which electoral formula is the most proportional? a new look with new evidence. Polit Anal 8(4):381–388
4. Bormann N-C, Golder M (2013) Democratic electoral systems around the world, 1946–2011. Electoral Stud 32(2):360–369
5. de Keijzer B (2008) A survey on the computation of power indices and related topics
6. Galiana J, Rodríguez I, Rubio F (2023) How to stop undesired propagations by using bi-level genetic algorithms. Appli Soft Comput 136:110094
7. Godoy A, Rodríguez I, Rubio F (2022) On the hardness of finding good pacts. In: 2022 IEEE CEC. IEEE, pp 1–8
8. Godoy A, Rodríguez I, Rubio F (2023) Majority problems: formal study and practical resolution. In: 2023 IEEE SMC. IEEE, pp 452–459
9. Kam C, Bertelli AM, Held A (2020) The electoral system, the party system and accountability in parliamentary government. Am Polit Sci Rev 114(3):744–760
10. Kochetov YA, Panin AA, Plyasunov AV (2017) Genetic local search and hardness of approximation for the server load balancing problem. Autom Remote Control 78(3):425–434
11. Martello S, Toth P (1987) Algorithms for knapsack problems. In: Surveys incombinatorial optimization, vol 132. North-Holland, pp 213–257
12. Medzihorsky J (2019) Rethinking the D'Hondt method. Politi Res Exchange 1(1):1–15
13. Mondal D, Parthiban N, Kavitha V, Rajasingh I (2021) APX-hardness and approximation for the k-burning number problem. In: WALCOM'21: algorithms and computation. Springer, pp 272–283
14. Muñoz A, Rubio F (2021) Evaluating genetic algorithms through the approximability hierarchy. J Computat Sci 53:101388
15. Passarelli G (2020) The presidential party: a theoretical framework for comparative analysis. Polit Stud Rev 18(1):87–107
16. Rodríguez I, Rosa-Velardo F, Rubio F (2020) Introducing complexity to formal testing. J Log Algebraic Methods Program 111:100502
17. Rodríguez I, Rubio D, Rubio F (2023) Complexity of adaptive testing in scenarios defined extensionally. Front Comput Sci 17(3):173206
18. Sartori G (1994) Parliamentary systems. In: Comparative constitutional engineering: an inquiry into structures, incentives and outcomes, pp 101–119
19. Shapley LS, Shubik M (1954) A method for evaluating the distribution of power in a committee system. Am Polit Sci Rev 48(3):787–792
20. Sharma VS, Srivastava P (2020) The pspace-hardness of understanding neural circuits. arXiv:2006.08266

Unraveling Social Network Factors in Predicting Depression with a Machine Learning Approach

Eunjae Kim, Kyu-man Han, and Eun Kyong Shin

Abstract This study identifies the key factor contributing to major depressive disorder using a machine learning approach. Depression is a global public health concern, particularly significant in South Korea due to its strong association with high suicide rates. While demographic, socioeconomic, medical history, and social network-focused factors are associated with depression, the consensus on the most critical one is challenging due to methodological limitations. To address this, we applied Partial Least Squares Discriminant Analysis (PLS-DA) and evaluated selectivity ratios. 172 participants were included, 70 depressed and 102 non-depressed, assessed by the Hamilton Depression Rating Scale. To gauge the social embeddings of participants, we used UCLA Loneliness Scale (UCLA-3). We included demographic, socioeconomic, and medical history features for the all-inclusive model. We found that the social network related factors were more critical than others. Seven items from the UCLA, including "No one really knows me well," had a selectivity ratio greater than 2. No features from other factors were found significant. This study underscores that poor-quality social relationships are strongly associated with depression. These findings can enhance early screening for depression and enable the development of tailored interventions for effective treatment and management.

Keywords Major depressive disorder · Social networks · Loneliness · Depression prediction · Machine learning · PLS-DA

E. Kim · E. K. Shin (✉)
Department of Sociology, Korea University, Seoul, Republic of Korea
e-mail: eunshin@korea.ac.kr

K. Han
Department of Psychiatry, Korea University College of Medicine, Seoul, Republic of Korea

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_6

1 Introduction

Major depressive disorder (depression, hereafter) is a significant and debilitating global public health issue [1]. It is characterized by a depressed mood, diminished interest in daily activities, and impaired cognition due to functional changes in brain circuits [2]. Depression is associated with weakened social functioning, including reduced interactions and communication [3]. In South Korea, depression is particularly concerning due to its rapidly increased prevalence and its high suicide rates, necessitating an epidemiological approach to prevention and treatment [2, 4].

Depression is associated with various factors. In addition to psychological and physiological factors [5–7], social factors, such as gender and socioeconomic status, play a crucial role in explaining depression [8–11]. Recent depression studies emphasize the importance of the quality of social networks [12–15]. Loneliness, a distressing psychological state due to unsatisfactory social relationships, is known to be strongly associated with depression [16–18].

However, the lack of consensus on the most critical factors associated with depression is due to methodological challenges [19]. This study identifies the key factor contributing to depression using a machine learning approach. We employed Partial Least Squares Discriminant Analysis (PLS-DA), which is well-suited for high-dimensional data and small sample sizes, to examine the importance of demographic, socioeconomic, medical history, and social network features among 172 study participants (70 depressed individuals and 102 non-depressed). We also evaluated selectivity ratios to assess the most significant features that distinguish depressed individuals from non-depressed individuals. Identifying the crucial factor is essential not only for the early screening of depression but also for developing tailored interventions for effective depression treatment and management.

2 Method

2.1 Data and Operationalization

Study participants ($N = 172$) were recruited from May 2019 to February 2021 at an outpatient psychiatric clinic in a university hospital in Seoul, South Korea. To minimize external influences on symptom, individuals were excluded based on the following criteria: comorbidity with other major psychiatric disorders, depression with psychotic features, acute suicidal or homicidal ideation requiring hospitalization, history of a serious or unstable medical illness, primary neurological conditions (e.g., Parkinson's disease, cerebrovascular disease, or epilepsy), and recent abnormalities detected in physical examination or laboratory tests. The Hamilton Depression Rating Scale was used to evaluate clinical depressive symptoms [20]. Participants scoring 8 or higher (up to 54) were classified as clinically depressed [21], while those scoring below 8 were classified as non-depressed.

To gauge the social embedding of the participants, we used the UCLA Loneliness Scale (UCLA–3) questionnaire items [22], which is widely used self-report tool evaluates loneliness in clinical settings [23]. The scale consists of 20 items that assess the subjective quality of social connectedness and psychological loneliness [22, 24]. The total score ranges from 20 to 80, with higher scores indicating greater levels of loneliness. Severity categories include low (20–34), moderate (35–49), moderately high (50–64), and high (65–80) [25]. For demographic and socioeconomic factor, we included age, gender, education, employment status, and marital status. We additionally add prescription history (medication usage for psychiatric conditions), the number of past depressive episodes, and family history of depression [26–28]. Finally, we incorporated stressful life events associated with depression [26].

2.2 Analysis

We employed Partial Least Squares Discriminant Analysis (PLS-DA) to evaluate the distinctive contribution of each explanatory factor for depression. PLS-DA is a supervised algorithm that effectively handles high-dimensional and imbalanced clinical data, performing well for disease classification including depression [29–31]. We used selectivity ratio (SR), a variable selection technique, to further evaluate the most critical explanatory features [32]. SR is a widely adopted for ranking features based on their contribution to outcome variance, and it is particularly effective with highly correlated features [30, 33]. We selected items with SR greater than 2, as this threshold is associated high probability of accurate classification [34].

We built two models: an all-inclusive model, which includes all available variables in the study and a social network-focused model, which includes only relational factors. The all-inclusive model identifies the most critical features from each factor, whereas the social network-focused model assesses the robustness of relational features exclusively. We used component 1 due to its highest explanatory power (all-inclusive = 50.37%, network-focused = 57.37%). The remaining minor components contributed less than 6% of explanatory power. All analyses were performed using R version 4.2.1 (R Foundation for Statistical Computing, Vienna, Austria).

2.3 Ethnical Approval

This study protocol was approved by the Institutional Review Board of the Korea University Anam Hospital (IRB no. 2019AN0174). Informed consents were obtained from all participants after a thorough explanation of the study.

3 Results

Among our study participants, 70 (40.7%) individuals exhibited depressive symptoms, while 102 (59.3%) were not depressed. The mean age was 37.4 years (SD = 13.90), ranging from 19 to 64 years. 41% ($N = 71$) were female, and 59% ($N = 101$) were male. 58% ($N = 100$) held a college degree or higher, 52% were employed ($N = 89$), and only 37% ($N = 63$) were married.

Table 1 summarizes the characteristics of the participants categorized by the presence of clinical depressive symptoms. Depressed individuals reported higher level of loneliness, lower socioeconomic status, and a personal or family history of depressive episodes. While non-depressed individuals reported low levels of loneliness, with a mean score of 32.8, depressed individuals were more likely to experience moderately high levels, with a mean score of 52.6 ($p < 0.000$). Depressed individuals are more likely to have lower educational attainment ($p < 0.000$), be unemployed ($p < 0.05$), have a prior history of depression ($p < 0.000$), be under pharmaceutical treatment ($p < 0.000$), and have a family member with a history of depression ($p < 0.05$). No significant differences were found in age, gender, or marital status.

Table 2 shows the classification performance of both the all-inclusive and network-focused models. The all-inclusive model demonstrated strong performance, with an accuracy of 0.831, specificity of 0.829, and sensitivity of 0.743 during calibration. Cross-validation results were slightly lower but remained reliable. Both Leave-One-Out (LOO) validation and venetian blinds cross-validation with four segments yielded an accuracy of 0.826, specificity of 0.882, and sensitivity of 0.743. The network-focused model also performed well, with an accuracy of 0.820, specificity of 0.882, and sensitivity of 0.729 across calibration and all validations, showing consistent predictive power. Despite excluding the demographic, socioeconomic and medical history variables, network-focused model performs exceptionally well compared to the all-inclusive model, which shows the critical contribution of relational factors in predicting depression.

To detect the most important features we conducted SR test, and features with an SR greater than 2 in both models are shown in Fig. 1. The relational factors were the most important features for both models. In the all-inclusive model, "No one really knows me well" and "I feel isolated from others" had the highest SR of 2.95, followed by "I feel alone" (SR = 2.63), "I feel left out" (SR = 2.47), "I am no longer close to anyone" (2.21), and "There is no one I can turn to" (SR = 2.12). The network-focused model showed similar results. "No one really knows me well" ranked highest with an SR of 3.08, followed by "I feel isolated from others" (SR = 2.98), "I feel alone" (SR = 2.47), "I feel left out" (SR = 2.44), "I am no longer close to anyone" (SR = 2.22), and "There is no one I can turn to" (SR = 2.12). Additionally, "There are no people I feel close to" (SR = 2.09) appeared in the network-focused model but not in the all-inclusive model. The significance of demographic, socioeconomic, or medical history features were below 1 and less important compared to the network features.

Table 1 Characteristics of participants

	Depressed ($N = 70$)	Non-depressed ($N = 102$)	p-value
Loneliness	52.6 ± 11.7	32.8 ± 10.0	***
Age	36.2 ± 14.1	38.1 ± 13.8	
Gender			
Female	32 (45.7%)	39 (38.2%)	
Male	38 (54.3%)	63 (61.8%)	
Education			***
High school or below	42 (60.0%)	30 (29.4%)	
College or higher	28 (40.0%)	72 (70.6%)	
Marital status			
Married	50 (71.4%)	59 (57.8%)	
Not married	20 (28.6%)	43 (42.2%)	
Employment status			*
Employed	28 (40.0%)	61 (59.8%)	
Not employed	42 (60.0%)	41 (40.2%)	
Recurrence of depression			
Yes	46 (65.7%)	1 (1.0%)	
No	24 (34.3%)	101 (99.0%)	
N of past depressive episodes	2.0 ± 2.7	0.0 ± 0.2	***
Medication			***
Medicated	51 (72.9%)	6 (5.9%)	
Not mediated	19 (27.1%)	96 (94.1%)	
Family history of depression			*
Yes	10 (14.3%)	3 (2.9%)	
No	60 (85.7%)	99 (97.1%)	

*** $p < 0.000$, ** $p < 0.01$, * $p < 0.05$

Table 2 Model performance of PLS-DA models on depressed individuals during calibration

	Accuracy	Specificity	Sensitivity
All-inclusive model	0.831	0.892	0.743
Network-focused model	0.820	0.882	0.729

For the top seven features, the mean differences between depressed and non-depressed individuals were all significant (Table 3). Although "No one really knows me well" had the highest SR in both the all-inclusive and network-focused models, the largest mean differences were observed in "I feel left out" (difference = 1.46, $p < 0.000$) and "I feel alone" (difference = 1.45, $p < 0.000$). Other significant features included: "No one really knows me well" (difference = 1.35, $p < 0.000$), "There is

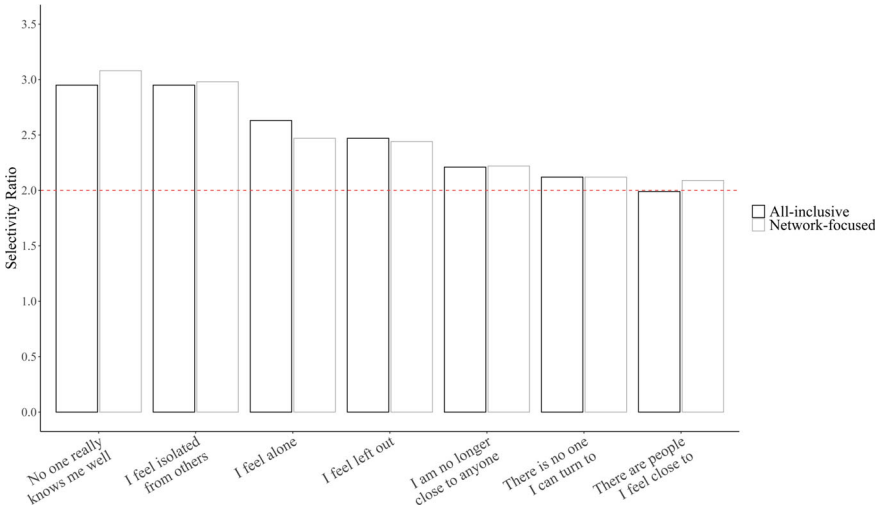


Fig. 1 Features with selectivity ratios greater than 2

no one I can turn to" (difference = 1.25, $p < 0.000$), "I feel isolated from others" (difference = 1.24, $p < 0.000$), "I am no longer close to anyone" (difference = 1.12, $p < 0.000$), and "There are no people I feel close to" (difference = 0.91, $p < 0.000$).

Additionally, self-reported open-ended answers for stressful life events further support the strong link between poor-quality social relationships and depressive symptoms. Among the depressed, 33 individuals (62% out of 53 respondents) indicated that significant social relationships were a major source of psychological distress. As shown in Fig. 2, familial relationships (42%)—including those with a spouse, parents-in-law, parents, and children—were the most common stressors, followed by romantic (9%) and work-related relationships (9%). These findings underscore the strong association between social networks and depression.

Table 3 Independent t-test results of features with selectivity ratio over 2

	Depressed ($N = 70$)	Non-depressed ($N = 102$)	Differences
No one really knows me well	2.80	1.45	1.35
I feel isolated from others	2.54	1.30	1.24
I feel alone	3.16	1.71	1.45
I feel left out	3.09	1.63	1.46
There is no one I can turn to	2.57	1.32	1.25
I am no longer close to anyone	2.49	1.37	1.12
There are no people I feel close to	2.57	1.66	0.91

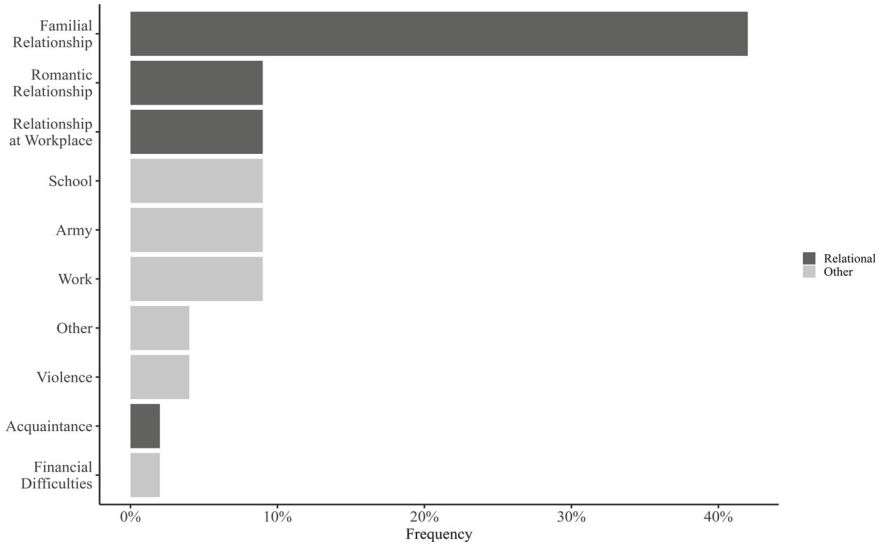


Fig. 2 Stressful life events

4 Conclusion

This study examined the distinctive predictive contribution of different factors associated with depression. We found that the relational factor was the most significant compared to demographic, socioeconomic, and medical history factors. Depressed individuals were more likely to lack meaningful social relationships. "No one knows me well" ranked the highest and all other important features are social network features. This suggests that the feeling of being understood is a key differentiator between depressed and non-depressed individuals. Furthermore, poor-quality social relationships, often resulting from interpersonal conflicts, can cause considerable distress leading to the development of depression. Our results align with existing research that emphasize the importance of social networks. Previous studies suggests that the quality of social relationships, especially loneliness, is a key predictor of mental health symptoms [16, 35]. We further demonstrated that poor-quality social relationships are the most critical factor associated with depression.

The Korean sociocultural context may amplify the impact of social relationships on depression. Traditionally, Korean society is tightly knitted and densely connected [36, 37]. High expectations for social relationships, which seek deep emotional acceptance and inclusion, can easily lead to perceived social isolation when expectations are unmet [38]. Rapid individualization and modernization may have increased prevalence of loneliness in South Korea [39]. Such demanding social relationships can become significant source of stress, which can lead to depression [40].

This study makes several contributions. First, we demonstrated that the social network-focused factor is more important than demographic, socioeconomic, and

medical history factors in predicting clinical depressive symptoms. This suggests that maintaining healthy relationships can help protect individuals against depression. Second, we addressed methodological challenges by applying a machine learning approach. Using PLS-DA models, we identified the most critical factor associated with depression in a small, high-dimensional clinical sample. Third, we revealed specific features from the UCLA-3 that differentiate depressed individuals from non-depressed individuals. Screening for high scores on these seven features could help detect individuals at risk of developing depression.

This study has some limitations. First, because it uses the cross-sectional data, we cannot establish a causal relationship between explanatory features and the outcome. Second, the findings have limited generalizability due to the small sample size, which included only patients from an outpatient psychiatric clinic. People with undiagnosed or untreated depression were not included. Third, the results may have been influenced by social distancing policies during the COVID-19 pandemic. Forced physical separation and lockdowns may have exacerbated psychological distress, particularly among individuals confined to limited spaces [41].

Despite the limitations, our findings strongly suggest that the quality of social relationships plays a critical role in depression. Depressed individuals reported a greater sense of a lack of meaningful relationships compared to non-depressed individuals. No features from demographic, socioeconomic, and medical history factors were found to be significant. Designing support group programs to address these psychosocial challenges in social relationships, such as the need to feel understood as reflected in the feature "No one really knows me well," could aid in the recovery of individuals with major depressive disorder.

Acknowledgements This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (No. 2022R1A4A1033856). This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2022R1A2C4001313). This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A3A130391).

References

1. Santomauro DF, Mantilla Herrera AM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, Abbafati C, Adolph C, Amlag JO, Aravkin AY, Bang-Jensen BL, Bertolacci GJ, Bloom SS, Castellano R, Castro E, Chakrabarti S, Chattopadhyay J, Cogen RM, Collins JK, Dai X, Dangel WJ, Dapper C, Deen A, Erickson M, Ewald SB, Flaxman AD, Frostad JJ, Fullman N, Giles JR, Giref AZ, Guo G, He J, Helak M, Hulland EN, Idrisov B, Lindstrom A, Linebarger E, Lotufo PA, Lozano R, Magistro B, Malta DC, Månsson JC, Marinho F, Mokdad AH, Monasta L, Naik P, Nomura S, O'Halloran JK, Ostroff SM, Pasovic M, Penberthy L, Reiner Jr RC, Reinke G, Ribeiro ALP, Sholokhov A, Sorensen RJD, Varavikova E, Vo AT, Walcott R, Watson S, Wiysonge CS, Zigler B, Hay SI, Vos T, Murray CJL, Whiteford HA, Ferrari AJ (2021) Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 398(10312):1700–1712

2. Otte C, Gold SM, Penninx BW, Pariante CM, Etkin A, Fava M, Mohr DC, Schatzberg AF (2016) Major depressive disorder. Nat Rev Dis Primers 2(1):1–20
3. Kupferberg A, Bicks L, Hasler G (2016) Social functioning in major depressive disorder. Neurosci Biobehav Rev 69(313–33
4. Kim GE, Jo M-W, Shin Y-W (2020) Increased prevalence of depression in South Korea from 2002 to 2013. Sci Rep 10(1):16979
5. Hasler G (2010) Pathophysiology of depression: do we have any solid evidence of interest to clinicians? World Psychiatry 9(3):155–161
6. Liu CH, Zhang E, Wong GTF, Hyun S (2020) Factors associated with depression, anxiety, and PTSD symptomatology during the COVID-19 pandemic: Clinical implications for US young adult mental health. Psychiatry Res 290(113172)
7. Smith JM, Bradley DP, James MF, Huang CLH (2006) Physiological studies of cortical spreading depression. Biol Rev 81(4):457–481
8. Lorant V, Deliège D, Eaton W, Robert A, Philippot P, Ansseau M (2003) Socioeconomic inequalities in depression: a meta-analysis. Am J Epidemiol 157(2):98–112
9. Miech RA, Shanahan MJ (2000) Socioeconomic status and depression over the life course. J Health Soc Behav 41(162–176)
10. Ross CE, Mirowsky J (2006) Sex differences in the effect of education on depression: resource multiplication or resource substitution. Soc Sci Med 63(5):1400–1413
11. Kessler RC, Bromet EJ (2013) The epidemiology of depression across cultures. Annu Rev Public Health 34(2013):119–138
12. Leigh-Hunt N, Bagguley D, Bash K, Turner V, Turnbull S, Valtorta N, Caan W (2017) An overview of systematic reviews on the public health consequences of social isolation and loneliness. Public Health 152(157–171)
13. Steen OD, Ori APS, Wardenaar KJ, van Loo HM (2022) Loneliness associates strongly with anxiety and depression during the COVID pandemic, especially in men and younger adults. Sci Rep 12(1):9517
14. Schwarzbach M, Luppa M, Forstmeier S, König HH, Riedel-Heller SG (2014) Social relations and depression in late life—A systematic review. Int J Geriatr Psychiatry 29(1):1–21
15. Barger SD, Messerli-Bürgy N, Barth J (2014) Social relationship correlates of major depressive disorder and depressive symptoms in Switzerland: nationally representative cross sectional study. BMC Public Health 14(1):1–10
16. Holvast F, Burger H, de Waal MM, van Marwijk HW, Comijs HC, Verhaak PF (2015) Loneliness is associated with poor prognosis in late-life depression: Longitudinal analysis of the Netherlands study of depression in older persons. J Affective Disorders 185(1–7)
17. Cacioppo JT, Hughes ME, Waite LJ, Hawkley LC, Thisted RA (2006) Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. Psychol Aging 21(1):140
18. Weiss R (1973) Loneliness: The experience of emotional and social isolation. MIT Press, Cambridege, MA
19. Shin EK, Mahajan R, Akbilgic O, Shaban-Nejad A (2018) Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. NPJ Digit Med 1(1):1–5
20. Hamilton M (1960) A rating scale for depression. J Neurol Neurosurg Psychiatry 23(1):56
21. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K (2013) Severity classification on the Hamilton depression rating scale. J Affect Disord 150(2):384–388
22. Russell DW (1996) UCLA loneliness scale (version 3): reliability, validity, and factor structure. J Pers Assess 66(1):20–40
23. Ausín B, Muñoz M, Martín T, Pérez-Santos E, Castellanos MÁ (2019) Confirmatory factor analysis of the Revised UCLA Loneliness Scale (UCLA LS-R) in individuals over 65. Aging Ment Health 23(3):345–351
24. Hawkley LC, Browne MW, Cacioppo JT (2005) How can I connect with thee? Let me count the ways. Psychol Sci 16(10):798–804

25. Deckx L, van den Akker M, Buntinx F (2014) Risk factors for loneliness in patients with cancer: A systematic literature review and meta-analysis. Eur J Oncol Nurs 18(5):466–477
26. Roca M, Gili M, Garcia-Campayo J, Armengol S, Bauza N, García-Toro M (2013) Stressful life events severity in patients with first and recurrent depressive episodes. Soc Psychiatry Psychiatric Epidemiol: Int J Res Soc Genetic Epidemiol Mental Health Services 48(12):1963–1969
27. Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J (2010) Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 303(1):47–53
28. Monroe SM, Slavich GM, Gotlib IH (2014) Life stress and family history for depression: the moderating role of past depressive episodes. J Psychiatr Res 49(90–95)
29. Lee LC, Liong C-Y, Jemain AA (2018) Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst 143(15):3526–3539
30. Mehmood T, Sæbø S, Liland KH (2020) Comparison of variable selection methods in partial least squares regression. J Chemom 34(6):e3226
31. Zhang F, Wu C, Jia C, Gao K, Wang J, Zhao H, Wang W, Chen J (2019) Artificial intelligence based discovery of the association between depression and chronic fatigue syndrome. J Affective Disord 250(380–390)
32. Andersen CM, Bro R (2010) Variable selection in regression—A tutorial. J Chemom 24(11–12):728–737
33. Farrés M, Platikanov S, Tsakovski S, Tauler R (2015) Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. J Chemom 29(10):528–536
34. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M, Kvalheim OM (2009) Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. Anal Chem 81(7):2581–2590
35. Domènech-Abella J, Lara E, Rubio-Valera M, Olaya B, Moneta MV, Rico-Uribe LA, Ayuso-Mateos JL, Mundó J, Haro JM (2017) Loneliness and depression in the elderly: the role of social network. Soc Psychiatry Psychiatric Epidemiol 52(381–390)
36. Youm Y, Laumann EO, Ferraro KF, Waite LJ, Kim HC, Park Y-R, Chu SH, Joo W-T, Lee JA (2014) Social network properties and self-rated health in later life: comparisons from the Korean social life, health, and aging project and the national social life, health and aging project. BMC Geriatr 14(1):102
37. Kim E, Shin EK (2023) Double-edged network effects on disclosing traumatic experiences among Korean "Comfort Women." J Interpers Violence 38(11–12):7728–7753
38. An S-J, Seo Y-S (2024) Exploring loneliness among Korean adults: a concept mapping approach. Behav Sci 14(6):492
39. Lee J, Man Chang S, Hahm BJ, Park JE, Seong SJ, Hong JP, Jeon HJ, An H, Kim BS (2023) Prevalence of loneliness and its association with suicidality in the general population: results from a nationwide survey in Korea. J Korean Med Sci 38(36):e287
40. Heu LC, van Zomeren M, Hansen N (2021) Does Loneliness thrive in relational freedom or restriction? the culture-loneliness framework. Rev Gen Psychol 25(1):60–72
41. Ernst M, Niederer D, Werner AM, Czaja SJ, Mikton C, Ong AD, Rosen T, Brähler E, Beutel ME (2022) Loneliness before and during the COVID-19 pandemic: A systematic review with meta-analysis. Am Psychol 77(5):660

Enhancing Reliability in Heavy Duty Autonomous Mobile Machines Through Fault Tolerant Edge Computing

Kalle Hakonen, Jussi Aaltonen, and Kari Koskinen

Abstract This paper presents a novel fault-tolerant edge computing architecture for heavy-duty autonomous mobile machines in industrial environments. The proposed system integrates two virtual machine hosts with a circular topology of four Ethernet switches, ensuring network resilience and operational continuity. A key feature is the implementation of automatic takeover protocols, enabling seamless transition between hosts during hardware failures. The network infrastructure leverages Rapid Spanning Tree Protocol (RSTP) and additional loop protection mechanisms to maintain stability and achieve rapid fault recovery. Virtual machines running Ubuntu Linux or FreeBSD are strategically deployed to handle specific tasks with critical services replicated for enhanced reliability. The system incorporates advanced data management through a PostgreSQL database with master-slave replication. Comprehensive fault tolerance mechanisms, including redundant connections and graceful degradation capabilities, ensure robust performance in challenging industrial settings. This architecture significantly enhances the reliability and autonomy of heavy-duty mobile machines, addressing the critical need for uninterrupted operation in industrial automation applications.

Keywords Edge computing · Graceful degradation · Autonomous mobile machine · Fault tolerance · High availability

K. Hakonen (✉) · J. Aaltonen · K. Koskinen
Tampere University, Tampere, Finland
e-mail: kalle.hakonen@tuni.fi
URL: https://research.tuni.fi/mrg/

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_7

1 Introduction

1.1 Background

The growth of complex, autonomous, heavy-duty mobile machines in industrial environments has necessitated robust, reliable computing systems capable of managing mission-critical operations [7, 8]. These machines, such as mining equipment like excavators and loaders, operate in challenging environments characterized by extreme conditions [5]. Operating conditions include, for example:

- Underground mines with complex tunnel networks
- GNSS denied environments and limited external network connectivity
- Harsh conditions, including extreme temperatures, dust, and mist
- Continuous operation cycles of 16–24 h per day.

Traditionally, these machines have been operated by onboard human drivers, utilizing hydraulic actuators and control systems built around CAN-bus technology. They incorporate a limited number of sensors for essential surveillance and operational monitoring. However, the increasing demand for efficiency and safety has driven the development of more autonomous capabilities [2]. The system proposed in this paper does not eliminate the possibility of human operation but enhances it, allowing for various levels of autonomy as operational needs dictate. This flexibility is crucial in adapting to different scenarios within industrial settings. A promising solution for these applications has emerged from edge computing, moving computational resources closer to the origin of data [1, 11]. In edge computing, data is processed near its origin, thus conserving bandwidth, reducing latency, and enhancing real-time decision-making capabilities [10]. Nevertheless, the distributed essence of edge computing raises challenges in ensuring system reliability and fault tolerance, particularly in the context of heavy-duty mobile machines operating in remote and harsh environments.

1.2 Problem Statement

Reliability is paramount for heavy-duty autonomous mobile machines operating in challenging, remote environments. The central problem lies in the vulnerability of edge computing systems to unexpected failures, potentially compromising continuous operation [5, 12]. Challenges include:

- Hardware failures in remote locations
- Network instability and intermittent connectivity
- Limited computational resources at the edge
- Diverse failure modes in mobile industrial settings
- Need for graceful degradation and rapid recovery without manual intervention.

1.3 Objective

This research aims to develop and implement a robust, fault-tolerant edge computing architecture for heavy-duty autonomous mobile machines while advancing the theoretical understanding of fault tolerance in mobile edge computing environments. The primary objectives are:

1. To develop a scalable and adaptable fault-tolerant edge computing architecture that enhances reliability and ensures continuous operation of autonomous mobile machines in challenging industrial environments.
2. To implement comprehensive fault management mechanisms, including automatic fault detection, isolation, recovery, and graceful degradation strategies.
3. To study IT infrastructure technologies not commonly used inside mobile machines but which might be beneficial for improving reliability.
4. To study the trade-offs between system complexity, fault tolerance, and operational efficiency in the context of autonomous mobile machines.

By achieving these objectives, we aim to make mobile industrial automation more efficient, safer, and reliable in harsh environments, eventually improving productivity and safety across various industries.

2 System Architecture

2.1 Overview

The fault-tolerant edge computing architecture for heavy-duty autonomous mobile machines integrates redundancy, distributed processing, and intelligent fault management. At its core, a dual-host virtual machine (VM) configuration ensures high availability and resilience against hardware failures.

This configuration comprises two computing nodes capable of hosting multiple VMs, working in tandem to share computational load and provide failover capabilities. The virtualization layer enables efficient resource utilization and process isolation, enhancing performance and security.

Four Ethernet switches in a circular topology connect the VM hosts, maintaining network connectivity despite potential link failures and further connecting the hosts to sensors and actuators crucial for the machine's operations. These layer 2 switches provide some layer 3 features to assist advanced management and control functions. Industrial features like RSTP allow quick network reconfiguration, minimizing downtime. Connections are visualized in Fig. 1.

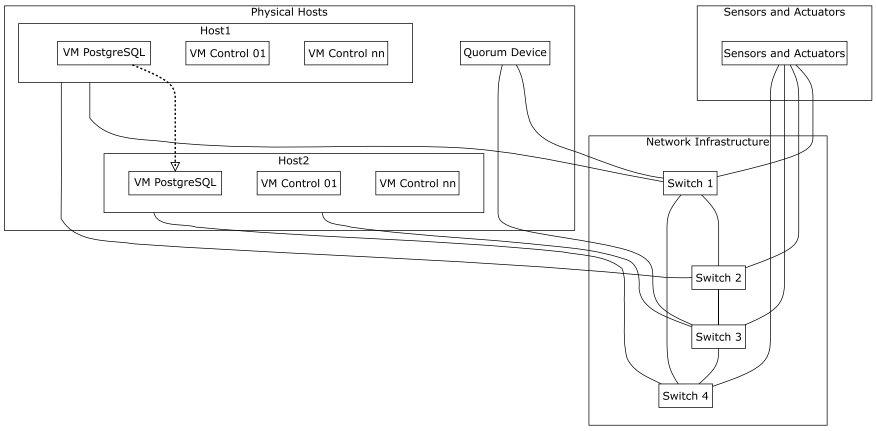


Fig. 1 Network architecture of the robust control system

Automatic takeover protocols enable a seamless transition of operations to the surviving host in case of failure. The architecture incorporates distributed data management techniques to ensure consistency across the system, employing consensus algorithms and distributed storage solutions to maintain data integrity during intermittent network connectivity.

The system includes mechanisms for graceful degradation, prioritizing critical functions when resources are constrained or components fail. This ensures uninterrupted essential operations while less critical tasks are suspended or run at reduced capacity.

The architecture designed for scalability and adaptability, can accommodate additional nodes as needed, and evolves to meet increasing computational demands or provide higher levels of redundancy for critical applications.

In conclusion, this fault-tolerant edge computing architecture represents a comprehensive approach to ensuring reliability and performance in heavy-duty autonomous mobile machines. Combining redundant hardware, intelligent software protocols, and adaptive resource management provides a robust platform capable of withstanding various failure scenarios while maintaining critical functions in challenging mobile industrial environments.

2.2 Virtual Machine Hosts

The fault-tolerant edge computing architecture utilizes dual high-performance VMs hosts as its computational core. These industrial-grade servers are engineered to operate reliably in harsh industrial environments, withstanding vibrations, temperature extremes, and other environmental stressors.

Each host employs a specialized hypervisor based on the Proxmox Virtual Environment, which is well-fitted for edge computing. This configuration allows the deployment of over 30 VMs across both hosts, each dedicated to specific autonomous machine operations. The VMs operate Ubuntu Linux or FreeBSD, selected for their stability, security features, and suitability for the specific task.

The dual-host setup is integral to the system's fault tolerance strategy. Under normal conditions, the computational load is balanced between hosts. In the event of host failure, critical VMs are quickly migrated or restarted on the surviving host, a process managed by health monitoring algorithms.

A PostgreSQL database system, configured in a master-slave replication setup, ensures data consistency and seamless failover. This centralized data store manages operational data, sensor readings, machine states, and configuration information.

The architecture is designed for scalability, allowing integration of additional hosts for increased complexity or redundancy, thus accommodating evolving computational demands in mobile industrial automation applications.

2.3 Virtual Machines

The fault-tolerant edge computing architecture strategically deploys over thirty VMs across two hosts. Dynamic resource allocation optimizes performance by configuring VMs with varying resource allocations based on their roles, ensuring optimal performance for each function. For instance, VMs responsible for real-time sensor data processing receive higher CPU and memory allocations to ensure low-latency performance, while less time-sensitive tasks are allocated fewer resources.

Each VM operates in an isolated environment, enhancing overall system security and preventing resource contention between services. The isolation is vital for the integrity of critical processes and for containing potential issues within individual VMs. Critical services run on VMs configured for real-time replication between hosts to ensure high availability, enabling rapid failover in case of host failure and minimizing downtime. VM checkpointing for critical VMs further enhances fault tolerance, allowing quick rollback to known good states during software failures or data corruption.

The system includes a sophisticated resource allocation algorithm that supports graceful degradation, dynamically adjusting VM priorities and resources in response to hardware failures or extreme processing loads. This approach ensures the autonomous machine can operate safely and effectively, even under constrained resources. Integrating a centralized PostgreSQL database keeps data consistent throughout the system and helps with fault tolerance and failover strategies. It also allows for real-time data sharing between the different parts of the autonomous system.

This VM-based approach, emphasizing fault tolerance and graceful degradation, is the key to achieving the high reliability and performance required for autonomous operations in challenging industrial environments. The VM architecture's flexibility

makes it possible to adapt to changing conditions and maintain operational continuity even in the case of hardware malfunction or resource constraints, making it well-suited for the demands of heavy-duty autonomous mobile machines.

2.4 Virtual Machine Versus Container

The comparison between complete VMs and containers is a topic of great interest in information technology. Containers, often managed by systems like Docker, offer lightweight solutions for rapid deployment and scalability. They encapsulate applications and dependencies into mostly isolated packages that share the kernel of the host operating system, resulting in lower utilization of resources but inducing potential security vulnerabilities due to this shared architecture.

In contrast, VMs provide robust isolation by encapsulating entire guest operating systems at the hardware level through hypervisors. This approach creates firm security boundaries, crucial for applications requiring stringent compliance or legacy systems unsuitable for containerization. VMs offer greater flexibility in operating system choice, benefiting applications tied to specific OS versions or environments requiring replication.

VMs demonstrate superior fault tolerance and resilience, allowing seamless migration between hosts during failures. They leverage established management tools and security controls, providing administrators with familiar frameworks for virtualized environments. While containers may initiate faster, VMs offer more consistent and predictable performance profiles, especially for resource-intensive applications. Additionally, VMs provide mature, feature-rich, persistent storage options, which are advantageous for long-term data retention.

In the context of fault-tolerant edge computing for heavy-duty autonomous mobile machines, VMs are preferable due to their strong isolation, diverse OS support, resource allocation flexibility, and performance predictability. The mature fault tolerance mechanisms of VMs, such as live migration and checkpointing, are essential for maintaining continuous operation in mobile industrial settings. Furthermore, VMs' persistent storage capabilities better suit robust data management and long-term retention needs critical for autonomous machine learning and system diagnostics.

While containers excel in agility and efficiency, VMs remain indispensable when strong isolation, diverse OS support, resource-intensive applications, and established management practices are required. The choice between containers and VMs ultimately depends on specific application needs and organizational requirements. However, VMs' advantages in security, flexibility, fault tolerance, performance consistency, and persistent storage make them crucial tools for heavy-duty autonomous mobile machine applications.

2.5 Network Switches

The proposed fault-tolerant edge computing system for heavy-duty autonomous mobile machines utilizes a network infrastructure comprising four Teltonika TSW202 industrial-grade Ethernet switches. The key innovation is how these technologies are applied and integrated to enhance reliability in mobile industrial environments.

A circular topology, atypical in mobile machinery, is implemented to maintain network integrity in harsh, high-vibration environments. Each switch maintains dedicated connections to its adjacent peers, creating redundant pathways essential for continuous operation during environmental stresses.

An advanced configuration of the RSTP (IEEE 802.1AC) manages this loop topology effectively [4]. Switch-to-switch uplink ports are configured with point-to-point link type settings and automatic cost calculations based on link speed, allowing quick adaptation to dynamic conditions. The implementation achieves instant failure detection in case of link failure and less than 3 seconds to adapt to the failure when using RSTP hello-time of 1 s.

This approach addresses the critical need for uninterrupted operation in mobile industrial automation applications, enhancing the reliability and autonomy of heavy-duty mobile machines.

2.6 Sensors and Actuators

Integrating sensors and actuators in heavy-duty autonomous mobile machines presents challenges distinct from stationary industrial systems. Our approach focuses on improving the reliability and fault tolerance of these critical components, which operate under extreme conditions and constant motion. Unlike stationary industrial systems executing single tasks, these machines operate in multi-modal configurations, requiring adaptive and robust sensor-actuator networks.

We implement a multi-layered sensor redundancy strategy, employing diverse sensing modalities to measure identical parameters. This redundancy is crucial in mobile environments prone to sensor failure due to vibration, shock, or environmental factors. For instance, our system combines LiDAR, radar, and camera data for obstacle detection, ensuring reliable environmental perception even if one sensor type fails.

Our actuator control system utilizes a distributed architecture where control units can operate independently if isolated from the central system. This design ensures that critical functions remain operational during partial system failures.

We have developed a novel predictive maintenance approach for sensors and actuators in mobile environments. The system detects signs of potential failures by

monitoring performance characteristics and environmental conditions. This proactive approach is useful for mobile machinery; unexpected faults between regular maintenance breaks often disrupt continuous operation.

Our fault-tolerant design considers the practical limitations of implementing a complete backup system in mobile machinery while focusing on improving availability and implementing software-based fault detection and compensation mechanisms. This balanced approach improves system resilience without increasing the costs and compromising economic feasibility.

3 Fault Tolerance Mechanisms

3.1 Automatic Takeover

The fault-tolerant edge computing system for heavy-duty autonomous mobile machines incorporates an automatic takeover mechanism utilizing a Proxmox High Availability (HA) cluster. The Proxmox HA cluster has been adapted for mobile settings through a three-node configuration: two active nodes on the machine and a third quorum device in a vehicle-mounted unit. This configuration maintains quorum and prevents split-brain scenarios during malfunction.

The implementation leverages Proxmox's live migration capabilities that proactively initiate VM migrations based on predictive analysis of machine movement and anticipated network conditions. This approach significantly reduces the risk of service interruptions during critical operations.

3.2 Redundant Connections

The circular network topology forms the basis for a robust redundancy strategy explicitly designed for mobile industrial environments. VM hosts are multi-homed, connected to multiple switches, maintaining network resilience despite the frequent vibrations and jolts that heavy machinery experiences. Link aggregation techniques have been adapted to the mobile context. A single logical link is formed by combining similar physical links to provide seamless failover and increase bandwidth in harsh operational conditions. This redundancy approach extends to sensors and actuators, often overlooked in less mobile-focused systems. Connecting these devices to multiple switches ensures data transmission to processing units through alternative paths in case of connection failures due to machine movement or environmental factors.

3.3 Protocols

Various protocols have been adapted to enhance fault tolerance in the challenging conditions faced by heavy-duty autonomous mobile machines. The RSTP implementation has been optimized for swift network reconfiguration during frequent link failures common in mobile environments. For virtual machine hosts running Proxmox, the Corosync Cluster Engine has been customized to maintain cluster integrity in high-vibration and electromagnetically noisy environments. Modifications allow frequent heartbeat messages and tolerant quorum calculations, which are crucial for maintaining an accurate cluster state in mobile settings. At the application level, protocols such as Fast DDS with ROS2 and MQTT have been enhanced with secondary distribution services and brokers, providing essential redundancies for maintaining messaging operations despite network instabilities typical in mobile industrial settings.

4 Reliability Analysis

This section presents an overview of the reliability calculations and methodologies for analyzing the fault-tolerant edge computing system designed for heavy-duty autonomous mobile machines.

The system's reliability is modeled using a hybrid series-parallel approach. This method accounts for the complex interdependencies between components and subsystems. The overall system reliability is calculated for different operational modes, reflecting the system's ability to adapt to various tasks and conditions.

1. For components in parallel:

$$R_{\text{parallel}} = 1 - \prod_{i=1}^{n} (1 - R_i) \tag{1}$$

   where R_i is the reliability of the i-th component.
2. For components in series:

$$R_{\text{series}} = \prod_{i=1}^{n} R_i \tag{2}$$

3. Overall system reliability:

$$R_{\text{system}} = R_{\text{series}} \times R_{\text{parallel}} \tag{3}$$

Table 1 shows individual component reliability over 10 years. These values are used as inputs for the above calculations. The table shows the components of the

Table 1 Reliability of traditional system components

Subsystem	Reliability
PLC	0.90
Communication bus	0.90
Software	0.90
24 V power	0.90
Hydraulics	1.00
Traditional sensors	1.00
Whole system	**0.66**

The bold values indicate the reliability of overall system.

Table 2 Reliability of new system components

Component	Reliability	Amount	P/S	Total reliability
Computer	0.90	2	2P	0.99
Communication bus	0.90	2	2P	0.99
Software with autorecovery	0.90	2	2P	0.99
Hydraulics	1.00			1.00
Traditional sensor	1.00			1.00
24 V power	0.90	2	2P	0.99
Supportive sensor	0.90	8	8S	0.43
Sensor for work and drive	0.90	12	12S	0.28
Sensor for safety	0.90	8	2P4S	0.96
Whole system				**0.11**

The bold values indicate the reliability of overall system.

traditional system for reference, which follows a pure series configuration. For comparability, every subsystem is claimed to have a reliability of 0.9 over ten years. For the same reason, Hydraulics and sensors, which will stay the same, are claimed to have a reliability of 1.0.

The traditional system might look superior when comparing it to the new one. As seen in Table 2, overall reliability decreases quickly when the number of subsystems increases. This happens even though some subsystems are doubled, allowing graceful degradation.

In practice, the new system's reliability varies across different operational modes due to the engagement of different components and subsystems. Several frameworks exist to classify autonomy levels in the development of autonomous mobile machines. The Society of Automotive Engineers (SAE) International's J3016 standard, initially developed for on-road vehicles, has been widely adopted and adapted for off-road and mining applications. This framework defines six levels of driving automation, from 0 (no automation) to 5 (full automation), as shown in Table 3.

Table 3 SAE levels of driving automation

Level	Description
0	No automation
1	Driver assistance
2	Partial automation
3	Conditional automation
4	High automation
5	Full automation

The SAE levels provide a standardized way to describe the capabilities of autonomous systems, facilitating communication between developers, regulators, and end-users [9]. In the context of mining mobile mining equipment, these levels help define the degree of human intervention required and the extent of the machine's decision-making capabilities. In our work, we primarily reference the SAE levels when discussing the autonomy capabilities of mobile machines. The autonomous mobile machine system described here operates in five modes, each tailored to specific operational requirements and operating under varying SAE levels.

Salvage mode bypasses standard safety protocols and autonomy features, allowing emergency operations or recovery in extreme situations. It relies solely on traditional sensors and a minimal control system, without utilizing any new or advanced sensor technologies, to ensure basic functionality in critical scenarios. This is clearly SAE level 0.

Human operator mode is the traditional system with enhanced safety features. This mode relies on direct human control while maintaining advanced safety measures. It incorporates new sensor technologies specifically designed to improve safety, providing operators with enhanced awareness of the machine's environment and potential hazards. The safety features are comparable to emergency braking and blind spot warnings in the car; thus, it operates at SAE level 0.

Augmented human operator mode builds upon the human operator mode. This configuration incorporates augmented reality technologies to enhance the operator's situational awareness and decision-making capabilities. Additionally, it utilizes supportive sensors for work-related tasks. While these supportive sensors may introduce a degree of complexity that could potentially decrease overall system reliability, it is important to note that they are not critical for basic operation. The system can continue to function safely even if these supportive sensors fail. These functions are comparable to lane centering in the car. These functions raise the SAE level to 1.

Work mode focuses on on-spot tasks, activating components and systems specific to stationary work operations and optimizing efficiency for localized tasks. It represents a partially autonomous configuration that builds upon the augmented human operator mode. In addition to the existing sensor array, this mode utilizes four critical extra sensors to enhance its autonomous capabilities for specific work-related operations. These additional sensors are essential for the autonomous functions in

Table 4 Reliability across different operational modes

Operational mode	SAE	Reliability	Description
Salvage	0	0.96	Bypass safety and autonomy
Operator	0	0.82	Traditional system, safety
Augmented operator	1	0.54	Traditional system, safety, augmented reality
Work	2	0.43	Stationary work systems
Driving and SLAM	4	0.35	Autonomous driving

this mode, distinguishing it from the previous modes in terms of both capability and operational requirements. SAE level is about 2.

Driving and SLAM mode represent the most autonomous operating mode of the system. It heavily utilizes sensor arrays and advanced computational systems for autonomous navigation and environment mapping, employing Simultaneous Localization and Mapping (SLAM) techniques. This mode incorporates 8 critical sensors, significantly enhancing its ability to perceive and interpret complex environments. The high number of critical sensors underscores the sophisticated level of autonomy achieved in this mode, enabling the machine to navigate and operate with minimal human intervention in dynamic industrial settings.

Table 4 shows the reliability and the level of autonomy for each operational mode. Reliability value is calculated using Eqs. 1–3 with values from Table 2. It highlights the evident fact that as the autonomy level of a system increases, the system's complexity naturally grows, leading to a decrease in overall reliability. To mitigate this, subsystems needed for safe human operations are connected in parallel, unlike serial connections, which have lower reliability since one component malfunctioning would lead to the loss of the entire system.

Critical components in the system, such as computer systems and sensors, are essential and shared across most operational modes. To ensure the reliability of these components, the system incorporates redundancy. This is reflected in the calculations where the reliability of these components is significantly improved. For instance, the reliability of two computers is calculated as:

$$R_{\text{one\_computer}} = 0.9 \tag{4}$$

$$R_{\text{two\_computers}} = 1 - (1 - R_{\text{one\_computer}})^2 = 1 - (1 - 0.9)^2 = 0.99 \tag{5}$$

This demonstrates how incorporating redundancy, in this case using two computers, dramatically improves reliability from 90 to 99%. With redundancy, the system

can continue operating even if one computer fails, which is crucial for maintaining performance across different modes.

The system's ability to operate in different modes with varying levels of reliability indicates a design philosophy of graceful degradation. This is particularly notable in the "salvage" mode, where the system can operate with minimal functionality, bypassing safety systems if necessary.

While not explicitly shown in the table, the reliability calculations implicitly consider the results of an FMEA. This is evident in how different components contribute to the reliability of various operational modes. This analysis has at least the following limitations and assumptions.

- Perfect reliability is assumed for hydraulics and traditional sensors, which may not reflect real-world conditions.
- The analysis assumes independence between failure modes of different components, which may not always be the case in a tightly integrated system.
- Environmental factors and their impact on component reliability are not explicitly accounted for in this model.

The reliability analysis demonstrates the complex interplay between components' reliability and system-level performance across different operational modes. While the new system shows improved reliability in some areas, the trade-offs in others highlight the challenges in designing multi-modal autonomous systems. This analysis provides a foundation for future reliability engineering efforts and system optimizations.

5 Upgrading and Reconfiguring

In today's interconnected industrial world, the ability to update, upgrade, and reconfigure autonomous mobile machines is not merely advantageous but essential. The ubiquity of network connectivity, whether local or global, necessitates a proactive approach to system maintenance and enhancement. Our fault-tolerant edge computing architecture, leveraging high-level operating systems, provides a robust foundation for addressing these requirements.

The proposed architecture offers several key advantages over traditional systems. Primarily, it enables Over-the-Air (OTA) updates, significantly reducing downtime and maintenance costs. The modular software architecture facilitates targeted upgrades and re-configurations without system-wide disruptions. Virtualization allows for isolated testing of updates before full deployment, enhancing system stability. Moreover, high-level operating systems provide robust security features and regular patches against emerging threats.

Traditional systems, often reliant on Programmable Logic Controllers (PLCs) and rigid Communication Area Network (CAN) protocols, present significant challenges in upgrading and reconfiguring. These systems typically require on-site interventions, leading to extended downtime and increased operational costs. In contrast, our

architecture's flexibility allows for rapid adaptation to changing demands and swift responses to security vulnerabilities.

Quickly deploying security patches is crucial in an era of evolving cyber threats. Our system's network-centric design, coupled with high-level operating systems, enables prompt responses to identified vulnerabilities. This agility in security management is vital for autonomous mobile machines operating in diverse and potentially hostile environments [5].

Mobile industrial environments are dynamic, with evolving operational requirements. The proposed architecture's reconfigurability allows for integrating new functionalities and optimizing existing processes without necessitating complete system overhauls. This adaptability ensures that autonomous mobile machines remain at the forefront of technological capabilities, maintaining their operational relevance and efficiency over time [10].

In conclusion, the upgrading and reconfiguring capabilities of our fault-tolerant edge computing architecture represent a significant advancement over traditional systems. By embracing modern, connected technologies, we provide a platform that is more responsive to current needs and well-positioned to adapt to future challenges in mobile industrial automation systems.

6 Results and Discussion

The proposed fault-tolerant edge computing architecture for heavy-duty autonomous mobile machines demonstrates significant improvements in reliability and operational flexibility compared to traditional systems. Our analysis reveals several key findings:

Table 4 summarizes the reliability calculations for different operational modes of the system. The results indicate a trade-off between system complexity and reliability as the level of autonomy increases. The result aligns with findings from similar studies in autonomous systems [7].

The system's ability to operate in different modes with varying levels of reliability demonstrates the effective implementation of graceful degradation principles. That is particularly evident in the "salvage" mode, which maintains a high reliability of 0.96 by bypassing non-critical systems.

The redundancy in critical components, such as computer systems and communication buses, significantly enhances system reliability. For instance, the reliability of the computer system improves from 0.90 for a single unit to 0.99 with redundancy.

The circular topology of the network switches, coupled with the implementation of RSTP, provides robust network resilience. In some cases, the proposed system achieves instant failure detection and network reconfiguration within 3 seconds when using an RSTP hello-time of 1 second. This rapid recovery is crucial for maintaining continuous operation in mobile industrial environments, outperforming traditional linear network topology [4].

While the overall system reliability appears lower than a traditional system (0.66 vs. 0.35 for the most complex mode), it is important to note that:

1. The new system offers significantly higher functionality and autonomy levels.
2. The reliability calculation for the new system includes a more comprehensive set of components, reflecting its increased complexity.
3. The system's ability to operate in multiple modes allows for continued operation at reduced functionality levels, which is not captured in a single reliability figure.

The architecture's high-level operating systems and virtualization technologies enable OTA updates and flexible reconfiguration. Those improvements represent a significant advancement over traditional PLC-based systems, aligning with industry trends toward more adaptable and updateable mobile industrial automation systems [5].

The current analysis provides valuable insights, but it has limitations. The assumption of perfect reliability for hydraulics and traditional sensors may not reflect real-world conditions. Additionally, the study assumes independence between failure modes of different components, which may only sometimes hold in tightly integrated systems.

7 Conclusion

Our system's performance can be attributed to several key features:

The circular network topology provides robust redundancy against link failures. The dual-host VM configuration with automatic takeover protocols ensures high availability. Implementing graceful degradation mechanisms allows the system to maintain critical functions even under partial failures. Adapting fault tolerance mechanisms specific for mobile high-vibration environments decreases the risk of physical failure.

While cloud-based systems offer high-computational redundancy, they fall short in scenarios with unreliable network connectivity, a common challenge in mobile industrial environments. Traditional systems, while robust in certain aspects, lack the flexibility and advanced fault tolerance features required for autonomous operations.

In conclusion, our proposed fault-tolerant edge computing architecture demonstrates significant advantages over existing systems, particularly in its ability to maintain operational continuity in the challenging conditions faced by heavy-duty autonomous mobile machines.

7.1 Summary

This paper presents a novel fault-tolerant edge computing architecture for heavy-duty autonomous mobile machines operating in challenging industrial environments.

The proposed system addresses critical reliability issues in industrial automation applications, particularly in scenarios with limited network connectivity and harsh operating conditions. Key features of the architecture include:

- A dual-host virtual machine configuration with over 30 VMs, enhancing computational redundancy and fault tolerance.
- A circular network topology utilizing four Ethernet switches, ensuring network resilience.
- Automatic takeover protocols for seamless transitions during hardware failures.
- Implementation of the RSTP for rapid fault recovery.
- A PostgreSQL database with master-slave replication for robust data management.
- Comprehensive fault tolerance mechanisms, including redundant connections and graceful degradation capabilities.

The system's reliability is analyzed across different operational modes, corresponding to various SAE levels of autonomy. The architecture demonstrates improved fault tolerance compared to traditional systems, particularly in maintaining critical functions during partial failures. The paper also discusses the system's upgradeability and reconfigurability, highlighting advantages over traditional PLC-based systems in terms of OTA updates and security patch deployment. This research enhances the reliability and autonomy of heavy-duty mobile machines in industrial settings, addressing the growing demand for robust, autonomous operations in challenging environments.

7.2 Future Work

The complexity of the fault-tolerant edge computing system for heavy-duty autonomous mobile machines necessitates advanced reliability analysis techniques. While the current study provides valuable insights, future work should focus on more nuanced approaches to examine the system's complicated behavior under various operational conditions. One promising future research direction is using Multi-State System (MSS) Reliability Analysis methods. Unlike traditional binary (functioning/failed) reliability models, MSS methods can account for systems with multiple performance levels [6]. This approach is particularly relevant for our system, which exhibits graceful degradation and can operate in various modes with different levels of functionality and reliability. The application of MSS methods could provide several benefits:

1. More accurate representation of system behavior, accounting for partial failures and degraded performance states.
2. Better modeling of the system's ability to transition between operational modes in response to component failures or environmental conditions.
3. Enhanced understanding of the system's resilience and capacity to maintain critical functions under diverse failure scenarios.

4. Improved decision-making for maintenance scheduling and resource allocation based on a more granular understanding of system states.

This approach would allow for a more refined analysis of the system's reliability across its various operational modes and provide deeper insights into its fault-tolerant capabilities. Such analysis could inform future design iterations and optimization strategies, ultimately enhancing the system's overall performance and reliability in mobile industrial environments.

Cybersecurity is critical for future research in distributed edge computing systems, particularly for autonomous mobile machines that operate near people and other machines. As these systems become more interconnected and rely heavily on data exchange, they become increasingly vulnerable to cyber threats and face heightened challenges in maintaining operational reliability and physical safety [3].

Acknowledgements This research was co-funded by European Union by grant agreement 101091895.

References

1. Gasmi K, Dilek S, Tosun S, Ozdemir S (2022) A survey on computation offloading and service placement in fog computing-based IoT. J Supercomput 78(2):1983–2014. https://doi.org/10.1007/s11227-021-03941-y
2. Gustafson A, Schunnesson H, Galar D, Kumar U (2013) Production and maintenance performance analysis: manual versus semi-automatic LHDS. J Qual Mainten Eng 19(1):74–88. https://doi.org/10.1108/13552511311304492. Mar
3. Halgamuge MN, Niyato D (2025) Adaptive edge security framework for dynamic IoT security policies in diverse environments. Comput Secur 148:104128. https://doi.org/10.1016/j.cose.2024.104128
4. IEEE (2004) IEEE standard for local and metropolitan area networks: media access control (MAC) bridges. IEEE Std. 802.1D-2004 (revision of IEEE Std. 802.1D-1998), pp 1–281. https://doi.org/10.1109/IEEESTD.2004.94569
5. Leitão P, Colombo AW, Karnouskos S (2016) Industrial automation based on cyber-physical systems technologies: prototype implementations and challenges. Comput Ind 81:11–25. https://doi.org/10.1016/j.compind.2015.08.004. Sep
6. Lisnianski A, Frenkel I, Ding Y (2010) Multi-state system reliability analysis and optimization for engineers and industrial managers. Springer, London. https://doi.org/10.1007/978-1-84996-320-6
7. Machado T, Fassbender D, Taheri A, Eriksson D, Gupta H, Molaei A, Forte P, Rai PK, Ghabcheloo R, Mäkinen S, Lilienthal AJ, Andreasson H, Geimer M (2021) Autonomous heavy-duty mobile machinery: A multidisciplinary collaborative challenge. In: 2021 IEEE international conference on technology and entrepreneurship (ICTE). pp 1–8. https://doi.org/10.1109/ICTE51655.2021.9584498
8. Pavloudakis F, Roumpos C, Agioutantis Z (2024) Using overall equipment effectiveness as a driver for improving the productivity of continuous mining systems. Mining Metal Explor. https://doi.org/10.1007/s42461-024-01096-x
9. SAE International (2021) J3016: taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE International. https://doi.org/10.4271/J3016_202104

10. Shi W, Cao J, Zhang Q, Li Y, Xu L (2016) Edge computing: vision and challenges. IEEE Internet of Things J 3(5):637–646. https://doi.org/10.1109/JIOT.2016.2579198. Oct
11. Xie R, Tang Q, Qiao S, Zhu H, Yu FR, Huang T (2021) When serverless computing meets edge computing: architecture, challenges, and open issues. IEEE Wirel Commun 28(5):126–133. https://doi.org/10.1109/MWC.001.2000466. Oct
12. Zhang W, Zeadally S, Zhou H, Zhang H, Wang N, Leung VCM (2023) Joint service quality control and resource allocation for service reliability maximization in edge computing. IEEE Trans Commun 71(2):935–948. https://doi.org/10.1109/TCOMM.2022.3227968. Feb

Business Information System Consultant Competences

Małgorzata Pańkowska

Abstract Business information system (BIS) consultants are working on solving problems of client companies, providing them with high-quality services, helping them quickly respond to changes in their ecosystems, and to the changes initiated by new technologies. Client is usually the most important actor in the consulting process. Therefore, the consultants are to be well educated to ensure the best satisfying solutions. This study focuses on business information system analysts' competences development to enable them participation in the consulting projects. In this study, the thematic review of literature was applied; the author's framework of consultants' competencies for business information system strategic analysis has been provided, and finally, the author formulate a recommendation on business analysis course for students of computer science at university. The findings indicate that both the students' motivation, knowledge, experience, as well as a strong theoretical background and a methodological support from cooperative business units influence innovativeness and creativity of BIS consultants.

Keywords Information system · Consultant · Strategic analysis · Requirement engineering · Prototyping

1 Introduction

Business information system (BIS) consulting organizations are usually placed in a business of selling services. Information communication technologies (ICTs) products, frameworks or platforms are separate objects of transactions, but occasionally they play a role in the sale of consulting services [1]. In the consulting process, the client's satisfaction may lead to their loyalty; hence, the ICT companies are strongly oriented toward employees, who have well developed consulting competencies. The BIS consulting company team covers various professionals, i.e., analysts, strategy

M. Pańkowska (✉)
University of Economics in Katowice, Katowice, Poland
e-mail: pank@ue.katowice.pl

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_8

planners, deployers, testers, risk and quality managers. Their number, morale, commitment, and competencies are vital to the consulting company's success. Currently, on the market, you can encounter various size consulting companies, e.g., big management consulting firms, system integrators, original equipment manufacturers (OEMs), software application developers, business process re-engineers, or boutique consulting firms specializing in a particular technology, strategy or industry groups. The big consulting firms, i.e., Accenture, KPMG, PwC, Deloitte, EY, or Capgemini provide a full range of the ICT and managerial services.

Usually, business organizations treat consulting differently than buying ICT products, or any other tangible goods. Consulting projects require an understanding of the client processes, people, internal regulations, and other resources. Consultants are required to investigate the client organization's decisions, management style, business requirements, values, and constraints. Consultants are asked to provide the right solution at the right time to the client organization's top-managers and decision-makers. Therefore, the client company demands people with specific competencies, i.e., agility, adaptability, ease of communication in foreign languages, business awareness, learnability, creative and holistic thinking, and customer focus. Consultant should easily adapt to new and unusual circumstances, understand the impact of decisions on the business strategy and operations. These challenges formulate research questions about what competencies can be developed at university and how they can be developed during the consulting project course. Answer that research questions was searching through the thematic literature survey as well as in a case study on student education at university. In this study, instead of systematic review, the thematic literature survey is to reveal what other authors say about the consulting projects. The rest of the paper is as follows. The second section covers definitions and a theoretical background of consulting services development. The third section covers the BIS modeling framework for pre-implementation analysis following the software engineering and deployment. Further, the author provides the BIS consultant roles' and competencies' schema. The last section includes discussion on the consulting project course development as well as observations after that course realization.

2 Theoretical Background

Generally, consulting has many interpretations, because it is applied in various disciplines and industry branches. According to Mullany [2], consulting is a professional communication research, where academia people and business practitioners jointly conduct a project by a set of individuals, who have expertise and experience enough to provide satisfactory results. Nissen [3] argues that consulting is characterized by the uncertainty associated with the actions under research. Hence, certain informal social relations play an important role in reducing the uncertainty. Managerial consulting is conducted in the form of a project, which is a temporary effort undertaken to create a unique product, service or result. However, although that project focuses on realization of tasks in a well established time, the consulting does not focus only on

the tasks' identification and realization, but rather on justification of those tasks, on explanation tools, methods, and other resources needed for those operations.

The consulting services need results from an asymmetry of knowledge between the professional consultant and a client. The prerequisite for consulting is the exchange of knowledge and information between the consulting service provider and the client company, as well as a trustworthy way of interactions among them. The valuable social resources needed in that project comprise interpersonal trust, knowledge exchange, and relationship proneness [4]. In systemic consulting, the theoretical concepts cover the architecture of changed ecosystems or processes, the design workshops and consultancy meetings, the use of tools in particular work settings as well as the attitude and knowledge of consultants [5]. Mauerer [6] emphasizes that consulting is a professional service provided to an external and financially independent client company. Kargulowa [7] defines consulting as an intellectual effort, including synergetic communication and reflective construction of self-identity of the involved partners. Bodenstein and Herget [8] argue that the central issue is a set of rules on the consulting goal formation, decision-making, and steering processes.

Although the consulting project is a temporary involvement of stakeholders, its results or consequences remain for years with the client company, which has implemented the project products. In the consulting processes, the cooperation of client and consultants require traceability of procedures and results, mutual trust, creation of opportunities for involvement evaluation and acceptance of results [8]. Both sides, i.e., stakeholders and consultants are interested in compliance with agreed contract and regulations, as well as in confidentiality. However, they are also expecting objectivity and independence in formulation of opinions and recommendations. Although the consulting contract is fundamental determinant of cooperation, the consulting project may fail because of several risks, i.e., lack of qualifications of consultants being engaged, lack of support from the client company top management, confused intentions, as well as unclear mutual expectations, imprecise consulting project goals, unsettled project course organization, undefined roles of the deployed consultants, lack of coordination during the project tasks realization, lack of explanations from the stakeholders and poor their involvement, lack of engagement of key opinion leaders and experts [8].

The ISO standard 20700:2017 [9] determines guidelines for managing the consultancy services. The aim of this norm is to improve transparency and increase understanding of the roles of clients and service providers in order to achieve better results in the consulting project, provide values for clients, and reduce risks in the consulting management process. That standard emphasizes professional behavior, social responsibility, avoiding the conflict of interest, and integrity as fundamental determinants of consulting project success.

3 Business Information System Modeling Framework

The BIS modeling for further designing and implementation is based on the requirement identification. According to Meyer [10] a requirement is a statement that defines product, process, design characteristics, or constraints. Requirements are identified with a software engineering product, along with other artifacts, such as code, designs, and tests. The specification of requirements is a challenge to business stakeholders; therefore, the information system modeling is expected to simplify the process. This particular process includes the requirements' elicitation and covers a use of systematic techniques, such as prototyping and structured surveys to recognize and describe the end-user needs. Each business information system should be developed in a particular ecosystem. For small and medium enterprises (SMEs), the description of the ecosystem may start from a user story, which is a usage scenario for actors interacting with the information system. A standard format for a user story may consist of three elements, i.e., identification of actors and their roles in a business organization, desired functionalities of the system, and a business purpose. However, for big companies and even for medium companies, the user story is not enough, and the system analysis requires studying business plans, regulations, and business transaction documents. Beyond that, the board of directors as well as end-users interviewing is highly required. The process of requirement collecting is a repetitive routine for explaining all information needs of the system end-users, i.e., business decision-makers. In the information system modeling, the focus should be moved beyond requirements to including the reasons for change (i.e., business drivers), the desired effect to achieve (i.e., business goals), identifying what components need to be implemented to reach the objective (i.e., business outcomes), the requirements, as well as the identification what need to be regulated (i.e., business rules) [11].

Figure 1 presents the business information system modeling framework to emphasize the categories important for the BIS environment identification. That conceptual model is visualized in the ArchiMate language. In Fig. 1, the information system requirements are central categories. They should be combined with goals, tasks, constraints, roles, actors and software components. People responsible for providing the requirements are consultants, i.e., business analysts and subject-matter experts. Goals are needs of the target organization, which the system will address. A role is understood as the human responsibility for the tasks. A constraint is a property of the environment that restricts what the information system can do [10]. Constraints are identified with:

- Business rules, i.e., constraints defined by the business organization;
- Physical rules imposed by laws of nature;
- Engineering rules, i.e., decisions being technical choices.

Business rule is a requirement on the conditions expressed in terms of the business enterprise or application domain. Business rules reveal policies, procedures, constraints concerning the use of resources, being compliant with laws and business conventions. According to Wang [12], business rules are classified as follows:

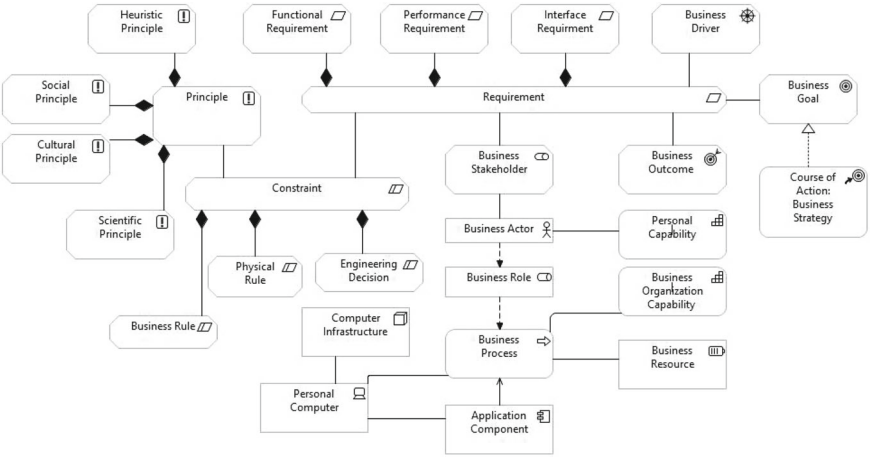


Fig. 1 Business information system modeling framework

- Structural business rules describing relationships and constraints among data elements, e.g., rules of integrity, derivation, reaction, production, or transformation;
- Behavioral business rules describing the principles of process execution;
- Functional business rules concerning the organization actions;
- Enforceable and unenforceable rules.

Business rules are presented graphically in a process model, for instance in the BPMN notation or they are described in a natural language. The processes are structured or ordered to reveal their interdependencies in a hierarchy or a value chain. The modeling of business processes is combined with identification of business rules, which reveal not only organizational constraints, but also obligations, permissions, restrictions, necessities, and possibilities. The business processes are placed in an enterprise system consisting of a purposeful network of various interdependent resources, i.e., supporting technologies, machines, buildings. The resources are needed to coordinate business functions, share information, allocate funding, and make decisions [13]. The business functioning is based on principles, which can be classified as heuristic, social, cultural, and scientific. The information system engineering principles are derived from the principles of that various sources and are based on both practice and theory [14]. Milani [11] has emphasized that the constraints are to be considered as limitations to a solution. They determine the solution realization and they are classified as budgetary, time, technology, legal, and organizational capabilities and competencies limitations.

The business information system modeling for the further designing, implementation and deployment requires further application of modeling languages, e.g., UML, SysML, GoalML. However, the system analysis and design can be supplemented by prototypes, which may be useful for requirements elicitation and requirement

engineering. That various applications of prototypes result from the fact that information system developers use them for different purposes. In general, a prototype provides an alternative approach to the classical questioning of the system stakeholders. With a prototype, the BIS consultant can demonstrate to stakeholders an actual model, program or a selected aspect of the modeled system, to get their opinions, recommendations, and decisions. According to Meyer [10], there are three kinds of prototypes:

- Throwaway prototypes, considered as preliminary version of a system;
- User interface prototypes, enabling discussion with end-user on an ergonomy of work with the software;
- Feasibility prototypes, focused on emphasizing selected aspects of the BIS development.

The open question is what software tools are recommended for the software prototyping. Nowadays, the answer is included in opportunities created by the low coding/no coding platforms, provided by various developers, as commercial tools (e.g., Webcon platform) or open source platforms (e.g., OutSystems). For the further designing and implementation, the BIS prototype should include the business analysis components, i.e., business logic, specification of actions and business rules, identification of actors, their roles, and their responsibilities for the tasks, as well as the business processes automation proposals, forms, reports, user interface versions, tests, and authorization issues.

4 Consultant's Competences

The theoretical background of the BIS modeling should emphasize that the BIS model includes declarative relationships between constructs. The BIS consultant competences are expected to respect the two theories, i.e., the Cognitive Load Theory (CLT) and the Cognitive Fit Theory (CFT). The CLT theory is built on the widely accepted model of human information processing. The theory explains the relationships between cognitive load and understanding performance, indicating that the representation of information should minimize cognitive load in the problem-solving tasks [12]. Wang [12] emphasized that the information processing model has three parts, i.e., sensory memory, working memory, and long-term memory. The sensory information is the information that the human brain collects from the senses (i.e., sight, hearing, touch). Working memory covers information gathered, for example, in the end-user interviewing process, and the long-term memory may include recording the consultant's earlier experiences, practices, and observations. According to Wang [12], the CFT theory provides an explanation of user performance while using different presentation formats of information and as such is used as support in the conceptual modeling process.

According to Milani [11], the BIS consultant is working with the following types of projects:

- Solving a technological or business problem;
- Exploitation of an opportunity concerning technological or marketing innovation;
- Cost saving through the process automation and the Internet of Things (IoT) solution implementation;
- Compliance with regulation;
- Environment management for management the emission of harmful substances;
- Data analytics projects including an integration of transactional and analytical systems for optimal decision-making and predictions;
- Commercial off the shelf (COTS) software customized implementation and deployment projects.

Figure 2 includes an identification of the BIS consultant roles and competences necessary for the system modeling. Mainly, the consultant should be responsible for discovering, synthesizing, and analyzing all information for the requirement elicitation.

The BIS consultants should be investigators of business situations. They are expected to identify and evaluate various options for improving business information systems, and ensure the effective implementation and factual usage. They should know various techniques, languages, and software tools for information visualization and they have abilities to explain to end-users, and deliberate with users on

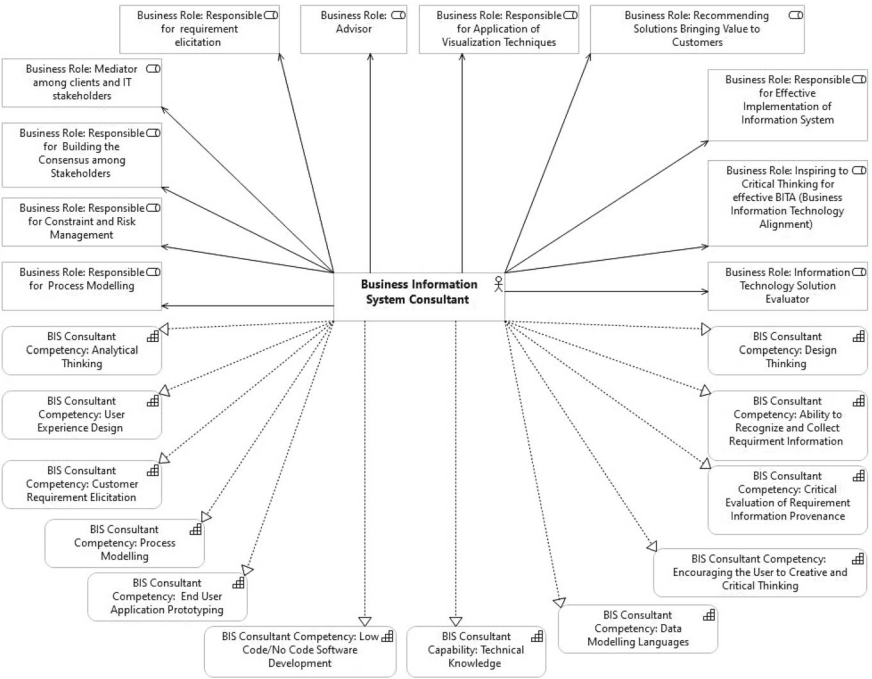


Fig. 2 BIS consultant's roles and competences

various solutions. They ought to know ICTs, e.g., AI, big data, blockchain. They should inspire end-users to creative thinking, perceiving the process improvement opportunities as well as possibilities of process automation, security, and business sustainability assurance.

The BIS consultant competencies are defined at a higher level of abstraction. They do not need to be detailed at that moment. A sufficient level of abstraction might be, for example, "customer requirement elicitation." In the ArchiMate notation model, the consultant competences are considered as the business organization capabilities, which define personal abilities to perform tasks. Capabilities are expected to be unique to ensure an idiosyncrasy of the BIS modeling. Martin et al. [14] argue that the word "capability" is defined in systems engineering as "the ability to do something useful under a particular set of conditions." There are some kinds of capabilities, e.g., organizational, system, operational, or personal. The potential capabilities are identified through an identification of capability gaps that unable realization of vision and IT strategy. The business capability map can be used to identify, which technologies contribute to achievement of the established levels of performance [14]. Effective management of human beings competences assists organizations to better understand the customer needs, and integrate or align the IT solutions with business needs to achieve strategic objectives and better market position through innovative solutions.

5 Discussion

In that wide domain of consulting services, the BIS consulting services have been developed for years for various clients, interested in software selection and implementation, system architecture planning and realization, system security and integration, strategy planning, or system outsourcing. This general view can be further complicated, because of the need to consider various specialized services. In that context, for universities, it is a challenge to cope with those issues and there is an open question on how to educate students to prepare them to consult technology and management projects. Universities may reveal a variety of questions and topics to explore, and opportunities to learn during a single engagement in consultancy. There is a quite different attitude to the consulting services at universities in comparison with business organizations. Big companies organize extensive training programs for their potential consultants prior to the beginning of their first engagement. Candidates are selected and highly motivated to work in a consulting company. They should be able to verify the questions, provide solutions, organize their responses, manage their time and other resources, because their answer should be actual and economically justified. Candidates should demonstrate that they can think quantitatively and that they are comfortable with statistical methods. They should present their knowledge on the newest technologies and be able to propose technical solutions suitable to the client request. They are to be creative and offer up a new and interesting perspective. Candidates should have analytical competencies as well as skills to formulate

conclusions and solve problems. Beyond that, they should have knowledge on business process modeling and optimization, on feasibility study elaboration, and on functional and operational analyses. The consulting companies may require to work under the time pressure, work with stressful clients as well as domain experts. Nissen [3] has noticed that the consultant has to suggest a solution to a client problem in the form of a report; however, he/she is not involved in the implementation process. Hence, the consultant can be like an expert, as a coordinator, a critic, or an initiator of the client's learning process. Consultant provides a solution proposal, but the final decision on the solution acceptance or not acceptance belongs to the client. Typical BIS consulting research methods are self-evaluation and qualitative surveys, i.e., qualitative interviews, participatory observation, case study, and group discussions [3].

At universities, the consulting project issues can be included in plenty of courses, i.e., project management, strategic management and marketing, system engineering, or software engineering. The issues of consultancy are hidden in the course teaching methods. According to Nissen [3] the phenomenon of business consulting is complex and requires different theoretical foundations and teaching methods. For instance, the Action Research method and Design Science Research paradigm provide valuable insights, because they emphasize the need to combine theoretical considerations with solving problems in practice. Although the ICT companies are looking for people competent in consulting, they apply a certain framework to fill the gaps between competency components like mandatory roles and required responsibilities, core knowledge, key skills, certifications, and experiments. The business organization develops their employees competency to engage them in the consulting projects, but university education aims to provide students knowledge on consulting and expand the consultant skills. Although the hard skills are developed during the various courses at university, there is a problem of development of soft competencies. Taking into account student abilities to learn and to create ideas, teachers are able to anticipate the students' willingness to spread the soft competences. To answer questions about what competencies should be developed during the consulting project course at university, a short survey was performed in a student group. The group of thirty students have participated in that investigation. The students were 25–35 years old. They all have worked at various business units and they are part-time students learning on weekends. Just three students have worked at consulting companies and mostly (i.e., 70%) they were not interested in working there. However, they said that if it would be necessary, they choose software implementation and system development positions in a consulting company. They have preferred more general services instead of specialistic services and big companies instead of small ones working for SMEs. Students were mostly interested in business strategy consulting or ICT strategy operationalization consulting. They quite well understood that the BIS consultants are to be highly qualified professionals with the necessary skills to design client specific BIS solutions. However, for them, also a personal attitude toward other stakeholders and people in the work team was important. Therefore, 70% of respondents declared that consultants should be polite and nice, despite the lack of outstanding competences. Hence, 30% of respondents would accept an arrogant consultant, but one

with excellent competencies and perfectly fulfilling the assigned tasks. Those observations allowed for the conclusion that students appreciate not only hard competences, but also consultants' soft skills. Students highlighted that the consultants were able to radically change the way of thinking of their clients and that they had access to the newest ICT knowledge. The consultants were believed to be able to integrate external and internal knowledge, and be helpful in the modern technology implementation process and various BIS environments integration. Students have emphasized that consulting company clients expect during the consulting process customized interviews as well as meetings, events, media conferences, email communication, newsletter dissemination, consulting company advertisements, online and onsite exhibitions and presentations. Surprisingly, students have not supported the thesis that clients are usually requested to reveal their internal business data. They argue that the consulting process could be realized in the testing environment, but not in real production processes.

In this consulting project course, students were asked to play the roles of consultants, simultaneously the external business units, i.e., software development companies were "clients" ordering software applications. Communication process allowed to reveal difficulties in explaining the requirements and ambiguity of definitions. During the discussions with business partners, students were strongly interested in recognizing the consulting project failures. They should learn that the failures implicate both sides simultaneously. By ensuring clear project goals, engaging relevant stakeholders, establishing a well-defined project organization, facilitating effective donor involvement, and appropriate selection of consultants the consulting project can be successful. The consulting projects course seems to ensure the development of hard and soft skills, management skills as well as knowledge on ICT tools and emerging technologies. That mandatory course is included in the program of studies among many other courses, i.e., economics, management, operational research, project management, statistics, econometric prediction, database management, software engineering, programming languages, and cybersecurity. Students are learning programming languages, i.e., python, C++, Java. Those courses are prerequisites to studying the consulting project course. During those studies, students have opportunities to learn about actually important and widely applied methods, i.e., DevOps, user-centered design (UCD), user experience (UX) design, or design thinking. DevOps method results from the observation that the responsibility for the software deliverables terminates with the user acceptance tests. However, the software implementation and exploitation should not be separated. DevOps highlights the necessity of continuous software delivery and its improvement, shared responsibility to deliver high-quality products, automation for development, testing, implementation and operation [15]. User-centered design (UCD) is an iterative method to system design and development [16]. That method focuses on human–computer interaction (HCI) modeling and implementation. The UX method concentrates on a recognition of end-user experiences created by the software product. That experience results from the set of perceptions, observations, and interactions during the use of the digital product. The experiences may concern intuitiveness of use, emotions, ethics, customer support, easiness to learn, or comfortability. Consultants

are expected to understand the context of use, identify the user requirements, produce design solutions, and evaluate the design [16]. Through the UX method, consultants are able to anticipate the end-user needs and understand the software system from the user point of view. Also, the design thinking is an approach requiring the understanding of customers, their behavior, willingness and motivations [17]. The hard skills, developed through the mentioned above courses, are quantifiable and easy teachable things. Those courses include professional knowledge, business software understandings, and abilities for the designing and development of a business information system. However, the soft skills are personal skills. They are intangible and related to people's personal experiences. They are difficult to teach as well as to evaluate. For instance, there is still an open question how to evaluate the student's ability to work in a group. Some group projects realized during studies do not ensure correct evaluation of the competency. The soft skills include good communication, listening, public speaking, critical thinking, emotional intelligence, self-control, problem-solving abilities, creative thinking, conflict management, skills of segregation of duties, and competency of responsibility sharing, design thinking, and customer service skills [18]. Creative thinking is a process to generate new ideas and synergy effects. It allows for a confrontation of different opinions to find a compromise solution. Conflict in consulting projects may result from poor communication and lack of mutual understanding. Although it has a negative interpretation, it can be constructive and lead to change the current situation to get a win-to-win situation for both involved sides. The consulting project course is typical work-in-progress, meaning that the current circumstances can change as the project progresses. Insufficient project requirements' instability is a challenge for students, who learn that in real company the negotiations, conflict management, and discrepancies removal are time-consuming and costly. Providing the BIS solution for clients requires knowledge about accounting, finance, economics and management. The computer science students many times have to admit lack of that knowledge. Hence, writing a user story is a recommended and necessary technique, as well as process modeling and application prototyping. For the business application development the low-code development platforms are advised, because they provide a visual development environment, allow for extraction data from various sources predefined by end-users, and permit end-users to build their own applications [19].

During the one semester course on the consultancy project, students are requested to elaborate a user story, modeling the business process, and provide a prototype of a low-code application. Finally, they have to prepare project documentation and the Pitch Deck. In students' opinion, the low-code platforms support the business application prototyping, and they are needed to enable the requirement identification, requirement gathering, and application validation. That prototyping is useful for requirement engineering complexity reduction. Although students have been requested to present differences among the BPMN process modeling and low-code platform business logic modeling, they argue that the low-code application modeling substitutes the UML design of application. Students have argued that the low-code application development is useful for software architecture holistic consideration, for the enterprise architecture management and its complexity reduction. The low-code

development tools are easy to learn and apply. Their documentations are understandable for students and easy to use. They support creative thinking and facilitate usage of other techniques, i.e., design thinking, DevOps, UCD, or UX, which are strongly required by the commercial software development companies. Students do not share the views that the low-code development platforms (LCDPs) enable construction of all modules of business information systems, although students trust the LCDPs facilitate the modules' integration. Students have noticed some weaknesses of the LCDPs, i.e., lack of access to full knowledge on the commercial tools, hardware capacity limitations, high cost of licensing, and limited functionalities. However, students have highlighted that the LCDPs facilitate consultant-client communication, their mutual trust, and risk-sharing. In the Pitch Deck presentations, students had to reveal why the product, i.e., a software application is useful, what functionalities it includes, what is its potential a total addressable market (TAM), how the requirements were identified, what low-code tools and additional libraries we used, what barriers and challenges were identified by students. Finally, the documentations and the application prototypes were evaluated by the ICT business companies, playing the role of client. In the last two years, the artificial intelligence (AI) researchers and developers have provided tools, which are increasingly used in students' education. In the survey, 30% of students have admitted that they use the ChatGPT in their professional work. Students argue that the ChatGPT facilitates the design thinking and the UX approach applying, as well as the graphical user interface design. Finally, students have admitted that they do not need a special course on consulting services, because the course issues can be provided in many other courses. A similar comment concerns the Generative AI solutions, i.e., ChatGPT. Students perceive opportunities to use GenAI to support learning and use it because it increases their competencies.

6 Conclusions

The consulting services encompass a comprehensive set of rules and procedures mandatory for a consulting project process. Consultants should know technology and business issues, and they should have competency to creatively solve problems. At universities, they are able to develop mainly hard skills, because the soft skills can be developed through direct communication in the project teams and in the cooperation process with a client. The BIS consulting competencies development should be oriented toward strong cooperation universities with ICT business units for better mutual understanding, providing original solutions, and recognizing the talented students.

References

1. Purba S, Delaney B (2003) High-value IT consulting: 12 keys to a thriving practice. McGraw-Hill/Osborne, New York
2. Mullany L (2020) Rethinking professional communication: new departures for global workplace research. In: Mullany E (ed) Professional communication. Palgrave Macmillan, Cham, pp 1–26. https://doi.org/10.1007/978-3-030-41668-3_1
3. Nissen V (2019) Consulting research: a scientific perspective on consulting. In: Nissen V (ed) Advances in consulting research, recent findings and practical cases. Springer Nature, Cham, pp 1–31. https://doi.org/10.1007/978-3-319-95999-3_1
4. Oesterle S, Buchwald A, Urbach N (2019) To measure is to know: development of an instrument for measuring consulting service value. In: Nissen V (ed) Advances in consulting research, recent findings and practical cases. Springer Nature, Cham, pp 79–102. https://doi.org/10.1007/978-3-319-95999-3_4
5. Hillebrand M, Mette S (2019) Systemic consultancy, theory and application. In: Nissen V (ed) Advances in consulting research, recent findings and practical cases. Springer Nature, Cham, pp 193–212. https://doi.org/10.1007/978-3-319-95999-3_9
6. Mauerer Ch (2019) The development of interpersonal trust between the consultant and client in the course of the consulting process. In: Nissen V (ed) Advances in consulting research, recent findings and practical cases. Springer Nature, Cham, pp 273–298. https://doi.org/10.1007/978-3-319-95999-3_13.
7. Kargulowa A (2016) Discourses of counsellogy: toward and anthropology of counselling. Societas Vistulana, Krakow
8. Bodenstein R, Herget J (2022) Consulting governance, implementing guidelines for successful projects. Springer, Berlin
9. ISO20700:2017 (2017) Guidelines for management consultancy services, 1st edn. https://www.iso.org/standard/63501.html. Last accessed 31 Oct 2024
10. Meyer B (2022) Handbook of requirements and business analysis. Springer, Cham
11. Milani F (2019) Digital business analysis. Springer, Cham
12. Wang W (2019) Integrating business process models and rules. Springer, Cham
13. Hutchison N, Pyster A, Cloutier R (2023) Using the systems engineering body of knowledge (SEBoK). In: Verma D (ed) Systems engineering for the digital age. Wiley, Hoboken, pp 791–803
14. Martin J, Fairley D, Lawson B, Fairsandier A (2024) Enterprise systems engineering background. In: A guide to the systems engineering body of knowledge (SEBoK). V.2.10, pp 679–685. https://www.sebokwiki.org/wiki/Guide_to_the_Systems_Engineering_Body_of_Knowledge_(SEBoK). Last accessed 31 Oct 2024
15. Alt R, Auth G, Kogler Ch (2019) Transformation of consulting for software-defined businesses: lessons from a DevOps case study in a German IT company. In: Nissen V (ed) Advances in consulting research, recent findings and practical cases. Springer Nature, Cham, pp 385–404. https://doi.org/10.1007/978-3-319-95999-3_19
16. Deuff D, Cosquer M (2013) User-centered agile method. Wiley, Hoboken
17. Zhang B, Dong N, Rischmoller L (2020) Design thinking in action: a DPR case study to develop a sustainable solution for labor resource management. In: Tommelein ID, Daniel E (eds) Proceedings of 28th Annual conference of the international group for lean construction (IGLC28). Berkeley, California, USA, pp. 25–36. https://doi.org/10.24928/2020/0137
18. Sathish AS, Rajkumar SV, Vijay V, Kathiravan Ch (2024) The significance of artificial intelligence in career progression and career pathway development. In: Khang A (ed) AI-oriented competency framework for talent management in the digital economy, models, technologies, applications, and implementation. Routledge, CRC Press, Boca Raton, pp 28–40. https://doi.org/10.1201/9781003440901-2
19. Di Ruscio D, Kolovos D, de Lara J, Pierantonio A, Tisi M, Wimmer M (2022) Low-code development and model-driven engineering: two sides of the same coin? Softw Syst Model 21:437–446. https://doi.org/10.1007/s10270-021-00970-2

Feasibility of the Cyber-Physical Nurse

Maya Dimitrova and Nina Valchkova

Abstract The paper presents the concept of a "cyber-physical nurse" from a feasibility perspective for wider inclusion in healthcare, in particular in relation to empathic communication with the patient. The results of a pilot study on user perception of two robotic and one human faces are presented and discussed in this context. Users attributed positive features to neutral agents' facial expressions, but not negative, which increases the feasibility of introducing social robots in healthcare. Some guidelines for cyber-physical nurse design are discussed, addressing challenges to its possible implementation in hospitals, rehabilitation centres, and home care settings.

Keywords Cyber-physical nurse · Healthcare · Synthetic sensors · Attachment · Trust

1 Introduction

1.1 A Definition of a "Cyber-Physical Nurse"

The cyber-physical nurse (CPN) is an emerging concept that describes a complex technological system to support—interactively and autonomously—care for vulnerable people in hospitals and at home [1–4]. "Nurse" is a specific category of helping profession that is mandatory in any treatment and rehabilitation centre, as well as in home care for severely chronically ill or elderly patients. In response to the question "What is the unique purpose of nursing?", the following concise definition of a nurse was produced by Google, based on [5]: "Nursing integrates the art and science of caring and focuses on the protection, promotion, and optimization of health and

M. Dimitrova (✉) · N. Valchkova
Institute of Robotics, Bulgarian Academy of Sciences, Sofia, Bulgaria
e-mail: m.dimitrova@ir.bas.bg

N. Valchkova
e-mail: nvalchkova@abv.bg

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_9

human functioning; prevention of illness and injury; facilitation of healing; and *alleviation of suffering through compassionate presence.*" This definition summarizes the most important aspects of the unique professional role of the nurse—to alleviate suffering by medical or physical treatment of the organism and—in parallel—to alleviate the psychological suffering of the patients in relation to their illness. While the first aspect of the nursing role is being modelled on a functional level intensively within the framework of service robots in hospitals [6–9], the second aspect is being addressed in the present paper as the factor, contributing primarily to defining the feasibility of designing effectively a cyber-physical nurse via modelling its *compassionate presence.*

The *perfect* cyber-physical nurse is characterized with the following four features—"pleasant, patient, polite, and physically strong"—according to [2]. The first three characteristics refer to the ability of the nurse to show *empathy* in communicating with the patient, which includes the following qualities—*pleasantness*, *patience*, and *politeness*—necessary to be perceived positively by the patient. In human training, these qualities are acquired after extensive training in the helping professions. The cyber-physical system can be *trained* effectively on numerous examples how to handle critical situations in nursing via implementation of various machine learning techniques, such as in [10, 11]. This brings in the fifth feature—*precision* in fulfilling the manipulation tasks on the patient. Also, knowledge of the patient's psychological profile is an important element of the overall cyber-physical system to support professionals or relatives in daily care, guiding the robot attitude towards the patient. The CPN has to behave in a transparent or *predictive* way to all human counterparts involved (as the sixth characteristic)—from the patient to any representative of the staff or people, involved in caring for the patient. Therefore, the feasibility of the *perfect* cyber-physical nurse depends on complying with at least 6P—"pleasant, patient, polite, physically strong, precise, and predictable."

The methodology for introducing the proposed CPNs, complying with 6P in real-world healthcare settings—hospitals, rehabilitation centres, and home care—with a larger, or more diverse population—is currently being developed [12]. In particular, novel is the implemented differential wheeled robot with encoder sensors to compute the number of rotations of the wheel and to determine the distance travelled [13]. Also, advanced speech-to-text models, such as Whisper, integrated with ROS middleware are implemented to interpret verbal commands accurately, even in noisy hospital environments [14]. The specific technologies to be used in building the cyber-physical nurse (such as sensors, actuators, and machine learning algorithms) are still under development [12, 15].

The paper presents the factors, defining user acceptance based on the degree of anthropomorphism of the CPN in Sect. 1.2, the results of a pilot study on user attitudes towards robots depending on the level of anthropomorphism in Sect. 2, the proposed guidelines for increasing the *feasibility* of designing effectively cyber-physical nurses in Sect. 3, including some future studies aimed at improving the robot's empathic capabilities.

1.2 Factors Defining Acceptance Based on the Degree of Anthropomorphism of a Cyber-Physical Nurse

An important research question is the degree of anthropomorphism as a factor for accepting a robot in the role of a nurse and building trust in the cyber-physical system on the part of the patient. For example, the factors, defining anthropomorphism in a social robot were investigated in [10] and defined as *human-like appearance*, *social intelligence*, *emotional capacity*, and *self-understanding*. While the latter two seem to be difficult to attain soon, the first two—the *human-like appearance* and the *social intelligence* are being studied intensively and seem plausible candidates to contribute to the feasibility of the cyber-physical nurse in the near future.

Social robots are being successfully employed to perform various professional roles, such as teachers or museum guides [11, 16–19], including for people with special needs [20, 21]. By being neutral in behavioural reactions and facial expressions, they seem suitable to perform professional roles such as co-therapists in counselling sessions [22]. However, the anthropomorphism of these robots is a feature, which is influencing the interaction in an unpredictable manner by invoking emotionally charged reactions [21]. In some cases, the emotional reaction to the robot depends on the individual internal criterion for *affinity* with a robot with different degree of human likeness [23–25]. It would be reasonable, therefore, to conduct empirical tests before implementing professional roles in robots to define the subjective range of individual preferences towards the acceptable degree of anthropomorphism of the robot.

In the present paper, we illustrate our proposal for conducting empirical tests by discussing the results of a pilot study on user acceptance of robots with different degrees of anthropomorphism, in order to formulate some guidelines for design of the "cyber-physical nurse."

2 The Study on Perception of Agents with Different Degree of Anthropomorphism

A study was conducted, hypothesizing that humanoid robots are well accepted in professional roles, similarly to their human counterparts, in particular as co-therapists in counselling sessions [22]. The experimental design was based on a methodology proposed in [26], which asked users to attribute personality traits to neutral (human) faces. We tested the trait attribution to two robotic faces and one human face, presented not as a photo, but as a brief video in autonomous mode of silent behaviour [22, 27]. The conducted study confirmed the hypothesis about the appropriateness of the particular robots for professional roles in general.

Here, we present the data and statistical analysis for the different genders, participating in the study (10 female and 6 male participants).

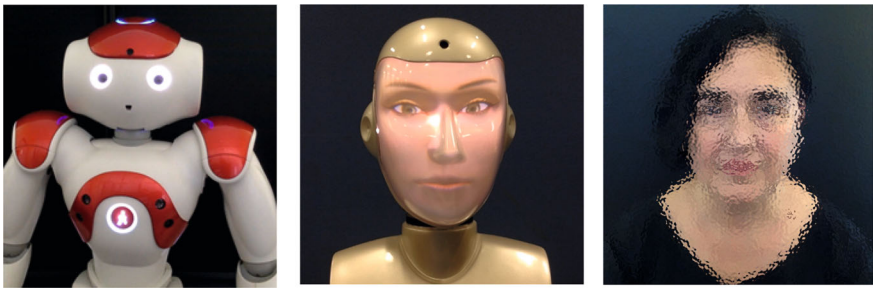


Fig. 1 Three agents, presented for assessment along 6 dimensions, representing 3 positive and 3 negative personality features (the human face is blurred here)

The participants are of age from 24 to 62 with normal and corrected to normal vision. The study was approved by the Ethics committee of IR-BAS (No 5, 8.11.2022).

Procedure. The experimental procedure consisted of the presentation of brief videos of faces of 3 agents (Fig. 1). The agents were—from left to right—a machine-looking robot NAO [28], an android type of robot SociBot [29] and a human face of a female actress. The participants were presented with one of 6 Likert scales to assess the degree, to which each face can be associated with a personality trait, from −3 to + 3. The traits were 3 positive and 3 negative personality features, selected from [26]. The positive dimensions were Emotional stability, Sociability, and Trustworthiness. The negative dimensions were Weirdness, Aggression, and Threat. The responses were made on a sheet of paper by circling the selected number of the Likert scale, presented on the respective slide.

Results. The main results of the study in relation to the overall perception of the agents were presented in [22]. The results show that viewers assess differently the positive and negative features, which can be attributed to the presented faces—human, android or robotic (machine-looking). They generally attributed positive features to the agents, although the facial expressions were neutral. In respect to the negative features, these were scored closer to 0, meaning that users are not *apriori* negative to robotic agents with neutral facial expressions.

In the present paper, the statistical analysis of the scores, given by the female (10) versus male (6) participants, is presented. Figure 2 plots the mean scores for the positive features, associated with the agents, whereas Fig. 3 plots the mean scores for the negative features.

The two way ANOVA with replication on the positive features revealed a main effect of type of agent, $F(2,17) = 9.660$, $p = 0.003$. No main effect of gender was observed, $F(1,17) = 1.324$, $p = 0.272$, nor any interaction of the factors $F(2,17) = 0.921$, $p = 0.424$. As can be seen in Fig. 2, the best accepted agent was the human when presented in a brief video of silent behaviour. This is important in relation to the nursing profession, since patients are being more confident with human caring staff than being alone or with a robot. Both female and male participants attributed

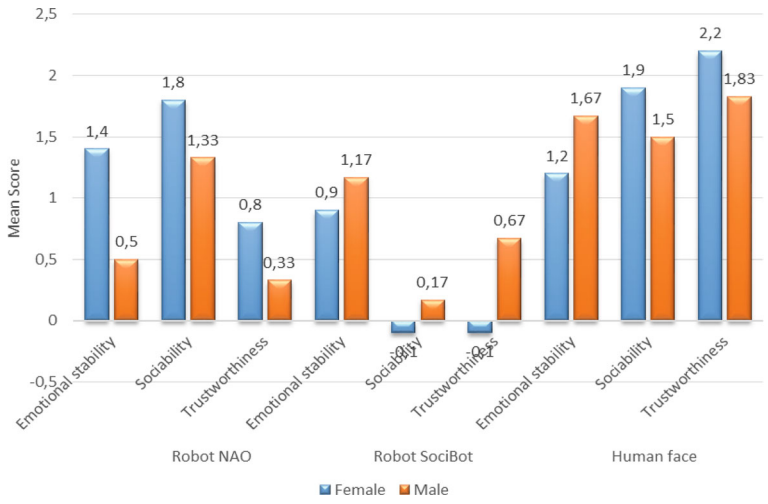


Fig. 2 Mean scores for the positive features, associated with the agents, by the female and male participants (see text)

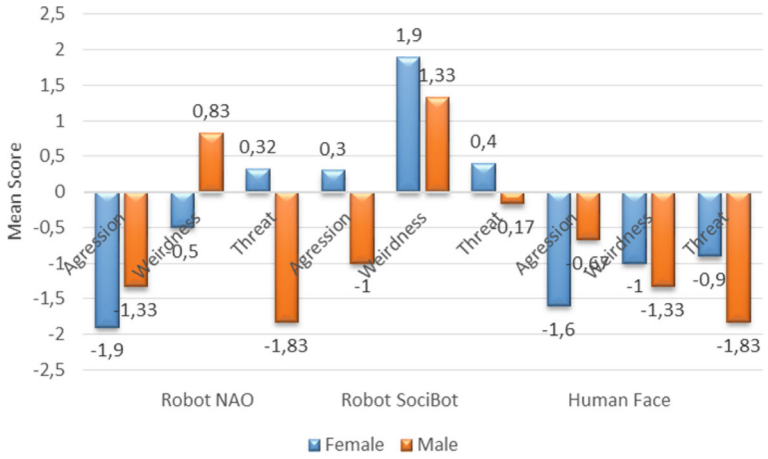


Fig. 3 Mean scores for the negative features, associated with the agents, by the female and male participants (see text)

positive features to the robot NAO, which is being a popular agent in many educational and entertaining scenarios. This confirms the Mori's proposal to seek a balance of human likeness and aesthetics in the design of humanoid robots [23]. In relation to the android robotic agent SociBot, the participants acknowledged its emotional stability as a positive feature. The other two features—sociability and trustworthiness were evaluated around 0, supporting the neutrality of the perception of the android.

Therefore, it is worth considering the possibility that different professional roles may require different level of anthropomorphism in the design of cyber-physical systems.

In relation to the attribution of negative features to neutral agents, the situation is similar to a certain extent.

The two way ANOVA with replication on the negative features revealed a main effect of type of agent, $F(2,17) = 5.303$, $p = 0.022$. No main effect of gender was observed, $F(1,17) = 0.312$, $p = 0.587$, nor any interaction of the factors $F(2,17) = 0.681$, $p = 0.524$. As can be seen in Fig. 3, the human face is attributed *lack* of any negative features and the robot NAO is attributed *lack* of aggression. The robot SociBot is attributed the weirdness trait, in contrast to the other agents. Therefore, the android types of robots can be generally assumed as being individually accepted by the users. The current results confirm the theory of Mori, which postulates a drop in the *affinity* with the increase of the level of anthropomorphism of the robot, especially in dynamic mode of presentation [24, 25].

In general, the fact that users are not certain about the negative features of the robotic agents is a promising outcome of the present study, since it reflects an essential psychological process. Seeing personality traits in neutral facial expressions is a projection of the internal state of the individual. If they are vulnerable people, then it is recommended to initially test their acceptance of agents of different level of anthropomorphism before implementing the professional scenarios. Our proposal is to perform the test implicitly, rather than to ask for explicit response from the patient as a method to increase the feasibility of the proposed cyber-physical nurse.

3 Guidelines for Design of the Cyber-Physical Nurse

The robotic system capability of capturing the so-called "social moments" is proposed in [30] to increase user confidence with the human–robot dialogue. A social robot like Pepper, for example, can take on the task of interviewing patients [31], thereby freeing up the human nurse for other activities. It is important to implement algorithms for a pleasant and polite attitude towards the patient, even more so that the robot is patient and does not get bored or irritated by the possible reactions of the patient.

A robot assistant in healthcare RoNA is described in [32]. She is a robotic nurse with enhanced manipulative abilities, such as lifting the patient and moving the body in space during rehabilitation. The robot's social abilities are more limited. It is necessary to further create the so-called high-level synthetic sensors that respond to behaviours expressing complex feelings such as *attachment* or *trust*.

3.1 Reinstatement of States of Affection and Attachment by the "Cyber-Physical Nurse"

What is attachment? This is a feeling that is not a basic emotion, but an emotional state of a higher abstract level, which develops throughout life and reflects the patients' relationship, especially in old age, with their closest people—family and friends. It is essential that the attachment of the patient to the closest people (or favourite robot) is based on the *affection* emotion. Social robots must be able to cope not only with possible outbursts of negative emotions on the part of the patients, but more importantly—be able to evoke memories of moments of *affection* and *attachment* in their lives—to loved ones, children, grandchildren, etc. Designing such a sensor is not an easy engineering task, but it is achievable [33]. The role of the positive emotions in rehabilitation is well established [34]. Our proposal is to introduce positive emotion by invoking memories of attachment, formed by feelings of affection. The feasibility of the cyber-physical nurse is in direct relation with the ability to design scenarios for the individual patients, based on their own memories of affection and attachment.

3.2 Trust Towards the "Cyber-Physical Nurse"

The role of feeling trust is no less important in rehabilitation than the reinstatement of positive emotions in the process of nurse-patient dialogue. Depending on the level of trust in the caregiver, the patients will adhere to their treatment and rehabilitation regimen. Therefore, a special aspect of scenario design is helping generate trust towards the cyber-physical nurse [35].

The inspection of Fig. 2 reveals that the *trustworthiness* is not expressed in relation to the robot, irrespective of the degree of anthropomorphism. At the same time the human is attributed the highest level of trustworthiness. This is a rather unexpected outcome of the study. Why people would not attribute trust to robotic agents of either kind—machine-looking or android? If this is interpreted in relation to endowing humanoid robots with nursing roles, this may pose a problem to designing the cyber-physical nurse and decrease the degree of its feasibility. One possibility would be to suggest coordination of the roles of the human and cyber-physical nurse towards the individual patient. Knowing that the patients trust the human nurse more would be a useful hint in trying to convince them to comply with the treatment and leave the repetitive and heavy tasks to the robot, which is being perceived as the emotionally stable and sociable agent.

The main shortcoming of the present study is the limited number of involved participants—16. Yet, even with this limited sample some observations are statistically confirmed, such as perceiving differently human and robotic agents on the one hand and not finding any gender difference—on the other.

3.3 Future Studies to Improve the Robot's Empathic Capabilities

In future studies we plan to explore how patients react to robotic utterances, since voice may not be accepted smoothly, even from robots with attractive appearance. On the other hand, it may be possible to engage in a dialogue with the robots of patients, who are hostile to the caregivers. An important aspect in research is designing collaborative human–robot care scenarios tailored to the individual patients for obtaining optimal results. These and other technical and logistical challenges still need to be addressed before the cyber-physical nurse can be deployed in real-world healthcare settings. Promising is the approach to employ learning from patient interactions to improve interactions in the course of time as well as to motivate and engage the patient in the care process.

4 Conclusions

Research has shown that the cyber-physical nurse is a feasible concept, much awaited in present day rehabilitation practice. It can be designed to support the overall rehabilitation of the patient by freeing the staff from heavy and tedious tasks. The presented study provides some guidelines towards the manner of communication of the robot with the patient as well as towards better coordination of the tasks between the human caregiver and the robot. Future studies will focus on improving the abilities to monitor the patient, detect emotional states and engage in a meaningful dialogue in order to effectively reinstate the *compassionate presence* on the part of the cyber-physical nurse.

Acknowledgements This work has received support from the Research Fund of Bulgaria for project No KP-06-ПН57/8, "Methodology for determining the functional parameters of a mobile collaborative service robot assistant in healthcare" (2021–2024).

References

1. González-González CS, Violant-Holz V, Gil-Iranzo RM (2021) Social robots in hospitals: a systematic review. Appl Sci 11:5976. https://doi.org/10.3390/app11135976
2. Dimitrova M (2022) Essential 'human' features of the cyber-physical nurse. Biomed J Sci Techn Res 46(1):36979–36980
3. Gibelli F, Ricci G, Sirignano A, Turrina S, De Leo D (2021) The increasing centrality of robotic technology in the context of nursing care: bioethical implications analyzed through a scoping review approach. J Healthcare Eng 2021(1):1478025. https://doi.org/10.1155/2021/1478025
4. Ragno L, Borboni A, Vannetti F, Amici C, Cusano N (2023) Application of social robots in healthcare: review on characteristics, requirements, technical solutions. Sensors 23(15):6820. https://doi.org/10.3390/s23156820

5. Nursing as a profession. https://www.rnpedia.com/nursing-notes/fundamentals-in-nursing-notes/nursing-profession/. Accessed 20 Nov 2024
6. Eriksen KT, Bodenhagen L (2023) Understanding human-robot teamwork in the wild: the difference between success and failure for mobile robots in hospitals. In: Proceedings of the 32nd IEEE international conference on robot and human interactive communication (RO-MAN), Busan, pp 277–284. https://doi.org/10.1109/RO-MAN57019.2023.10309638
7. van der Putte D, Boumans R, Neerincx M, Rikkert MO, de Mul M (2019) A social robot for autonomous health data acquisition among hospitalized patients: an exploratory field study. In: Proceedings of the 14th ACM/IEEE international conference on human-robot interaction (HRI), Daegu, pp 658–659. https://doi.org/10.1109/HRI.2019.8673280
8. Stokes F, Palmer A (2020) Artificial intelligence and robotics in nursing: ethics of caring as a guide to dividing tasks between AI and humans. Nurs Philos 21:e12306. https://doi.org/10.1111/nup.12306
9. Richert A, Schiffmann M, Yuan C (2020) A nursing robot for social interactions and health assessment. In: Advances in human factors in robots and unmanned systems: proceedings of the AHFE 2019 international conference on human factors in robots and unmanned systems, July 24–28, 2019, Washington DC, USA, pp 83–91
10. Wangpitipanit S, Lininger J, Anderson N (2024) Exploring the deep learning of artificial intelligence in nursing: a concept analysis with Walker and Avant's approach. BMC Nurs 23:529. https://doi.org/10.1186/s12912-024-02170-x
11. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, Wolf-Ostermann K (2021) Application scenarios for artificial intelligence in nursing care: rapid review. J Med Internet Res 23(11):e26522. https://doi.org/10.2196/26522
12. Valchkova N, Zahariev R, Raykov P, Cvetkov V (2023) Methodology for designing a collaborative mobile service robot based on criteria for researching its functional capabilities. In: Proceedings of the 2023 international conference on engineering and emerging technologies (ICEET), pp 1–6
13. Tsvetkov V, Valchkova N, Zahariev R (2024) Sensory system for controlling robot's motion. In: Proceedings XXV international conference "robotics and mechatronics 2024". https://robomed.bg/wp-content/uploads/2024/11/Tsvetkov-V.-N.-Valchkova-R.Zahariev.-Sensory-System-for-Controlling-Robots-Motion.pdf
14. Angelov G, Paunski Y (2024) Voice controlled user interface for ROS service robots in healthcare. In: Proceedings XXV international conference "robotics and mechatronics 2024". https://robomed.bg/wp-content/uploads/2024/11/G_Angelov_Voice-Controlled-Interface-for-ROS-Service-Robots-in-Healthcare.pdf
15. Valchkova N, Zahariev R, Angelov G, Paunski Y (2024) Methodology for design of hydrogen robot for medical needs. Compt Rend Acad Bulgare Sci 77(8):1176–1184
16. Chi OH, Chi CG, Gursoy D (2024) Seeing personhood in machines: conceptualizing anthropomorphism of social robots. J Serv Res 1:7196. https://doi.org/10.1177/10946705241297196
17. Tzampazaki M, Vrochidou E, Papakostas GA (2024) Social robots in education: to select or not to select a robot for a teaching subject at an educational level? In: Proceedings of the 2024 international conference on software, telecommunications and computer networks (softCOM), Split, Croatia, pp 1–6. https://doi.org/10.23919/SoftCOM62040.2024.10721629
18. Song H, Huang S, Barakova E, Ham J, Markopoulos P (2023) How social robots can influence motivation as motivators in learning: a scoping review. In: Proceedings of the 16th international conference on pervasive technologies related to assistive environments, pp 313–320. https://doi.org/10.1145/3594806.3604591
19. Rosa S, Randazzo M, Landini E, Bernagozzi S, Sacco G, Piccinino M, Natale L (2024) Tour guide robot: a 5G-enabled robot museum guide. Front Robot AI 10:1323675. https://doi.org/10.3389/frobt.2023.1323675
20. Bogdanova G, Todorov T, Noev N, Sabev N, Chehlarova N, Todorova-Ekmekci M, Krastev A (2024) An ecosystem for the provision of digital accessibility for people with special needs. Information 15(6):315. https://doi.org/10.3390/info15060315

21. Dimitrova M, Bogdanova G, Noev N, Sabev N, Angelov G, Paunski Y, Todorova Ekmekci M, Krastev A (2023) Digital accessibility for people with special needs: conceptual models and innovative ecosystems. In: Proceedings of the 2023 8th international conference on smart and sustainable technologies (SpliTech), pp 1–5. https://ieeexplore.ieee.org/document/10193524
22. Dimitrova M, Garate VR, Withey D, Harper C (2023) Implicit aspects of the psychosocial rehabilitation with a humanoid robot. In: Kubincová Z, Caruso F, Kim T, Ivanova M, Lancia L, Pellegrino MA (eds) Methodologies and intelligent systems for technology enhanced learning, workshops—13th international conference. MIS4TEL 2023. Lecture notes in networks and systems, vol 769. Springer, Cham, pp 119–128. https://doi.org/10.1007/978-3-031-42134-1_12
23. Mori M, MacDorman KF, Kageki N (2012) The uncanny valley (from the field). IEEE Robot Autom Mag 19(2):98–100
24. Berns K, Ashok A (2024) "You Scare Me": the effects of humanoid robot appearance, emotion, and interaction skills on Uncanny Valley phenomenon. Actuators 13(10):419. https://doi.org/10.3390/act13100419
25. Moore RK (2012) A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. Sci Rep 2(1):864
26. Said CP, Sebe N (2009) Todorov A (2009) Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. Emotion 9(2):260–264
27. Dimitrova M, Chehlarova N, Madzharov A, Krastev A, Chavdarov I (2024) Psychophysics of user acceptance of social cyber-physical systems. Front Robot AI 11:1414853. https://doi.org/10.3389/frobt.2024.1414853
28. NAO the humanoid and programmable robot—Aldebaran. https://corporate-internal-prod.aldebaran.com/en/nao. Accessed 21 Nov 2024
29. SociBot. https://wiki.engineeredarts.co.uk/SociBot. Accessed 21 Nov 2024
30. Durantin G, Heath S, Wiles J (2017) Social moments: a perspective on interaction for social robotics. Front Robot AI 4:24. https://doi.org/10.3389/frobt.2017.00024/full
31. Tulsulkar G, Mishra N, Thalmann NM et al (2021) Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? A thorough study based on computer vision methods. Vis Comput 37:3019–3038. https://doi.org/10.1007/s00371-021-02242-y
32. Hu J, Edsinger A, Lim YJ, Donaldson N, Solano M et al (2011) An advanced medical robotic system augmenting healthcare capabilities: robotic nursing assistant. In: Proceedings of the 2011 IEEE international conference on robotics and automation, pp 6264–6269. https://ieeexplore.ieee.org/abstract/document/598021
33. Jamisola RS (2014) Of love and affection and the gaze sensor. Lovotics 1(1):751
34. Seale GS, Berges IM, Ottenbacher KJ, Ostir GV (2010) Change in positive emotion and recovery of functional status following stroke. Rehabil Psychol 55(1):33–39. https://doi.org/10.1037/a0018744
35. Langer A, Feingold-Polak R, Mueller O, Kellmeyer P, Levy-Tzedek S (2019) Trust in socially assistive robots: considerations for use in rehabilitation. Neurosci Biobehav Rev 104:231–239. https://doi.org/10.1016/j.neubiorev.2019.07.014

Semantic Landscape of Legal Lexicons: Unpacking Medical Decision-Making Controversies

Haesol Kim, Eunjae Kim, Sou Hyun Jang, and Eun Kyong Shin

Abstract This study empirically examined historical trajectories of the semantic landscape of legal conflicts over medical decision-making. We unveiled the lexical structures of lawsuit verdicts, tracing how the core concepts of shared decision-making (SDM)-duty of care, duty to explain, self-determination-have developed and been contextualized in legal discourses. We retrieved publicly available court verdicts using the search keyword 'patient' and screened them for relevance to doctor-patient communications. The final corpus comprised 251 South Korean verdicts issued between 1974 and 2023. We analyzed the verdicts using neural topic modeling and semantic network analysis. Our study showed that topic diversity has expanded over time, indicating increased complexity of semantic structures regarding medical decision-making conflicts. We also found two dominant topics: disputes over health-care providers' liability and disputes over the compensation for medical malpractice. The results of semantic network analysis showed that the rhetorics of patients' right to medical self-determination are not closely tied to the professional responsibility to explain and care. The decoupled semantic relationships of patients' right and health professionals' duties revealed the barriers of SDM implementations.

Keywords Medicine and law · Shared decision-making · Patients-doctors relationship · Medical court verdicts · Computational methods

H. Kim
Department of Sociology, Stanford University, Stanford, California, USA

E. Kim · S. H. Jang · E. K. Shin (✉)
Department of Sociology, Korea University, Seongbuk-Gu, Seoul, Republic of Korea
e-mail: eunshin@korea.ac.kr

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_10

1 Introduction

Shared decision-making (SDM) has gained significance in healthcare for enhancing patient satisfaction, treatment adherence, prognosis, and quality of life [1–4]. This approach emphasizes effective communication between patients and medical professionals to prioritize patients' interests in critical medical decision-making. Medical professionals not only explain benefits and risks of treatment options but also incorporate patient opinions and preferences [5, 6]. This patient-centered communication can foster patient safety, promote patient autonomy, and improve quality of healthcare [6, 7].

Despite its strengths, the institutional adoption and practical application in healthcare settings of this model is still at an early stage in most societies [5, 8, 9]. Effective SDM was less likely to occur due to lack of time and resources among medical professionals [10]. A lack of communication skills and awareness to include patients as an important agent in decision-making has been identified as a major hindrance [11]. Paternalism and power imbalance in communications are also critiqued for ineffective communication [12–14].

Ineffective medical communication can escalate into lawsuits. From a legal perspective, SDM is composed with three key concepts: duty of care (enhancing treatment outcome), duty to explain (delineating treatment options), and self-determination (integrating patient preference) [15, 16]. Legal data can provide a unique and empirical source to identify the lexical structure of the key concepts in medical litigation. This data allows us to observe the existing patterns of communication conflicts over medical decision-making. The effectiveness of communication can be assessed based on the connections among the three concepts [15]. Examining the relationships among the key concepts enables the development of better strategies and policies to implement SDM in practice.

This study analyzed historical trajectories of the semantic landscape of medical conflicts. Unveiling the critical themes emerging from the rhetoric of disputes and interconnectedness among the concepts, we traced how the core concepts of SDM have developed and been contextualized in legal discourses. We take South Korea (hereinafter Korea) as the case—a country with a short history of adopting SDM but a long history of legal conflicts over medical decision-making [17, 18]. Using all publicly available court verdicts over medical decision-making in Korea (N = 251, 1974–2023), a country where SDM is not yet prevalent with relatively limited patient autonomy [19], we created semantic networks to observe rhetorical changes in medical-legal conflicts over time. We further analyzed ego-networks of the lexicons, "duty of care," "duty to explain," and "self-determination" to examine their association within the context of medical decision-making.

2 Data and Method

2.1 Data

In order to obtain court verdicts related to medical decision-making, we collected publicly available court judgments. Using "patient" as a search keyword, we retrieved data from the legal information database provided by the Supreme Court of Korea through automatic web crawling. The initial dataset included court cases both related and unrelated to doctor-patient communications ($N = 1608$). We therefore manually screened relevant verdicts. Three coders independently coded the assigned documents into three labels, "relevant," "irrelevant," and "unclear." To ensure the reliability of the manual coding, the coders cross-checked the coded labels. The intercoder reliability was 88.6%, showing good reliability and consistency. 251 documents were selected for the final dataset. Details on the filtering process for court verdicts are provided in Fig. 1.

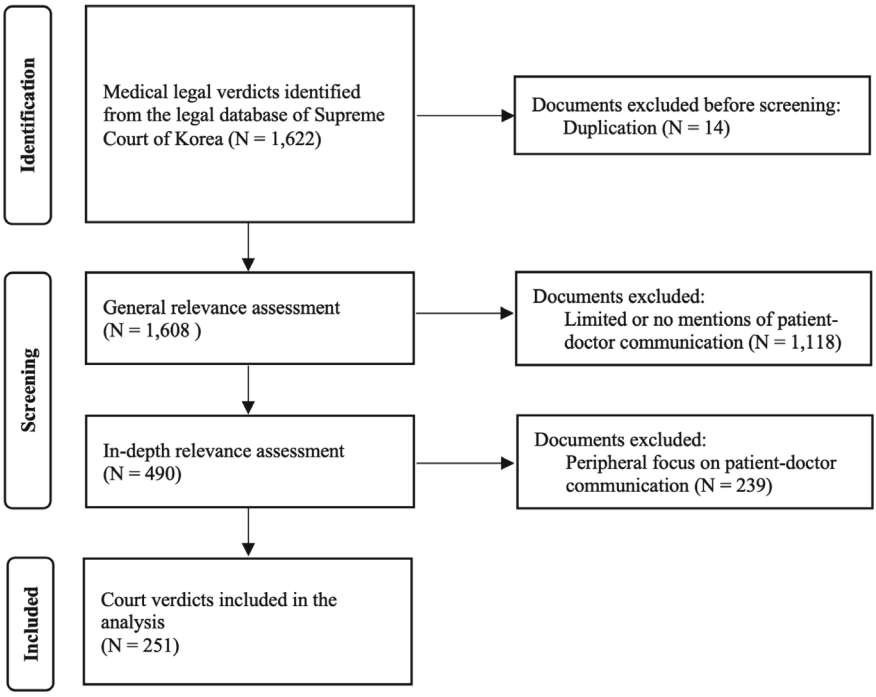


Fig. 1 Flow diagram of court verdicts filtering

2.2 Methods

Neural Topic Modeling. We used neural topic modeling analysis to discover dominant topics in the medical verdict corpus. Specifically, we applied BERTopic using the LEGAL-BERT, a pre-trained language model specialized for legal domain [20]. Compared to traditional topic modeling methods, BERTopic uses the embedding representations from an advanced pre-trained language model to effectively capture the contextual semantics of the words or sentences within the corpus [21]. It can provide topic models that reflect contextualized keywords. We removed repetitive legal terms such as "plaintiff," "court," or "pleading" to avoid overestimation of connectivity.

Semantic Network Analysis. We applied semantic network analysis to explore the meaning clusters generated within the data. To construct shared meanings within data, lexical units are represented as nodes and their co-occurrences within a predefined window are represented as links [22, 23]. We employed the Louvain community detection algorithm that efficiently works on large networks by maximizing between-cluster differences and within-cluster cohesiveness. We treated each verdict as a window and used the top 500 words based on their co-occurrence frequencies to identify dominant themes. We primarily used nouns and adjectives to ensure clear interpretation of the networks.

2.3 Analysis

We created semantic networks using medical verdict data to capture changes in semantic structures over time. First, we conducted neural topic modeling to develop meaning clusters—the networked maps of connectedness between semantic entities—corresponding to each of the emerging themes. Neural topic modeling analysis can reveal dominant topics so we can categorize the documents based on extracted topics. We then created two sets of semantic networks: time-collapsed and longitudinal. The time-collapsed network can show overall meaning structures of the verdicts, while the longitudinal network reveals evolution of semantic networks, illustrating how semantic clusters are entangled or separated over time. The semantic networks were created using CorText Manager and Gephi [24, 25].

3 Results

The number and issues addressed in court verdicts on medical disputes have increased and diversified over time. Figure 2 displays the yearly distribution of the final medical lawsuits from 1974 to 2023 by the line graph. The figure additionally shows the

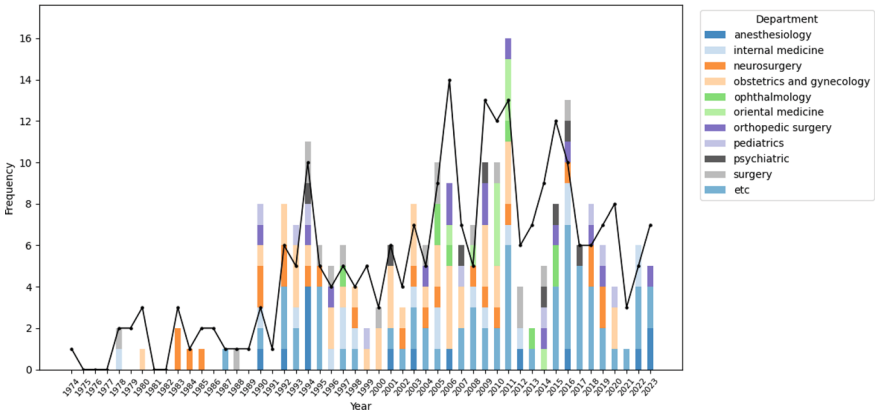


Fig. 2 Yearly distribution of medical court verdicts (line) and mentions of medical departments (bar). *Note* Some court verdicts mention more than one medical department

annual frequency of department mentions in the medical-legal verdicts by the bar graph, focusing on the top ten most mentioned departments. The number of court verdicts continued to rise rapidly for the last five decades. This indicates that the number of medical lawsuits increased.

The legal lexicons of medical disputes have dramatically diversified. Figure 3 displays the semantic network of the 483 most-used words. The word appears on the map when its cumulative occurrence reaches 20% of its total frequency. The colors of nodes denote their cluster memberships. The growth of diversity expanded dramatically through the second half. The results show the increased complexity of semantic network structures regarding medical decision-making. It suggests that patient-doctor communication problems cover wider areas of the medical services and practices over time.

We found two major topics (Table 1). Seven documents were automatically dropped. The first topic concerned disputes over physicians' liability for malpractices ($N = 133$). This topic included terms such as "negligence," "surgery," "accused," "treatment," "procedure," and "death," which addresses the extent of the responsibility of medical practitioners. It discussed whether medical practitioners fully fulfilled their duty of care and duty to explain.

The second topic ($N = 111$) addressed the extent of compensation. This topic represents the financial compensation from medical practitioners (or hospitals) for health damages caused by their negligence. It included terms such as "amount," "pay," "expenses," and "damages."

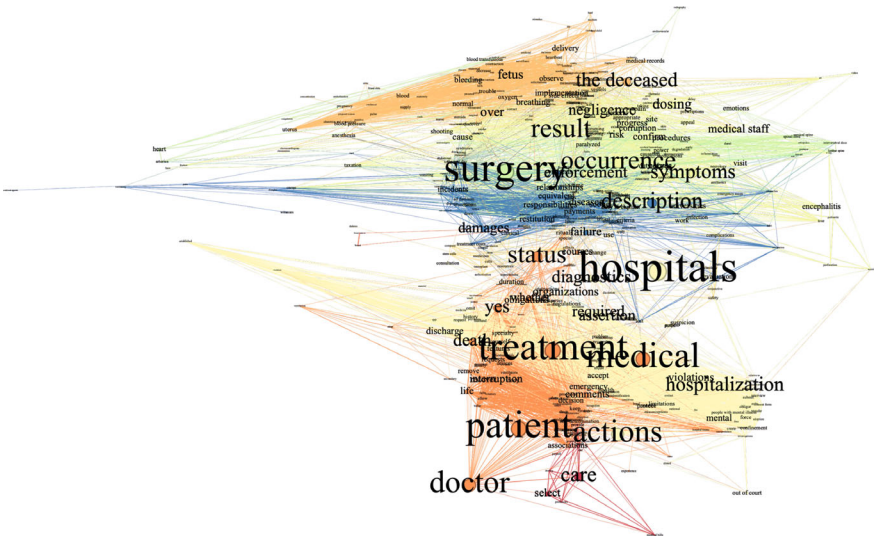


Fig. 3 Historical semantic network of the entire final corpus

Table 1 Topic modeling result

Topic	*N*	Words	Example sentences
Topic 1 liability	133	Negligence, surgery, circumstances, treatment, procedure, medical, hospital, death, patient	……the plaintiff would not have undergone the procedure in question if she had been properly informed by the defendant about the exact state of affairs in clinical medicine
Topic 2 compensation	111	Annum, amount, pay, expenses, damages, rate, claims, claim, per	……the court may take into account the victim's factors that contributed to the cause or aggravation of the damage by applying the doctrine of contributory negligence while determining the amount of damages……

3.1 Topic 1 Liability Semantic Network Analysis

To dissect the contexts where key terms appeared, "duty of care," "duty to explain," and "self-determination," we analyzed semantic networks of each topic. In the semantic network of Topic 1 (Fig. 4), six clusters were identified. Overall network statistics indicate that the network is well-divided into cohesive groups, in particular the modularity of 0.665 and the average clustering coefficient of 0.559 (Table 2).

The node level statistics indicate the significance of key SDM concepts in Topic 1 semantic network (Table 3). All three concepts had high raw and weighted degree compared to the network averages. "Self-determination" showed moderately high eigenvector centrality, reflecting solid links to central nodes. "Duty to explain" and

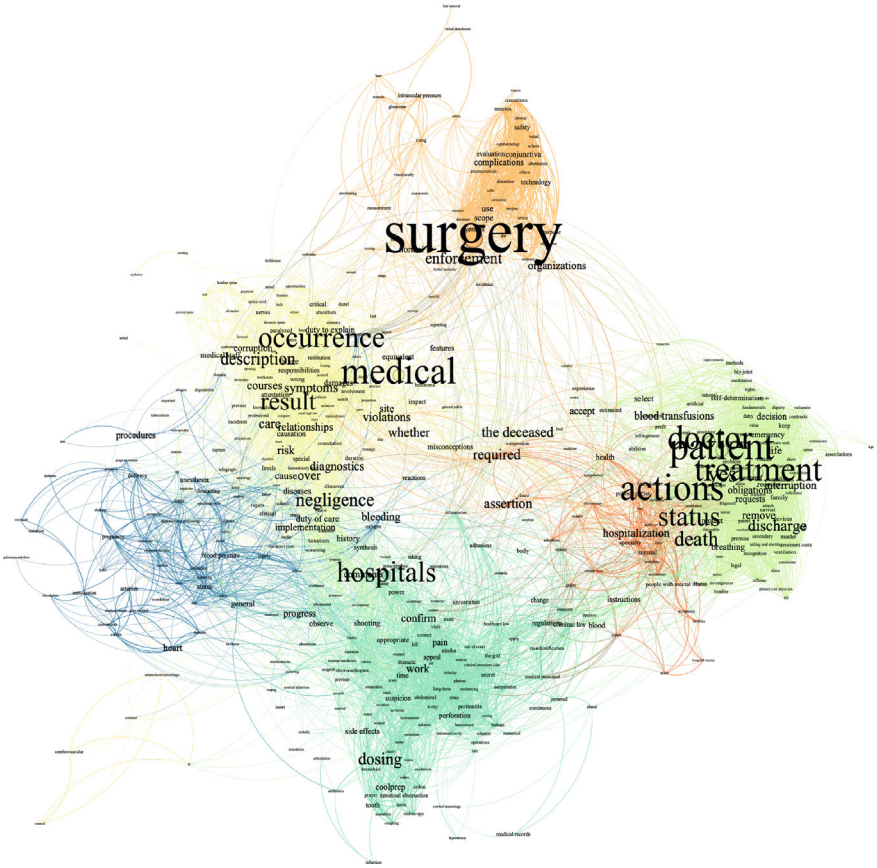


Fig. 4 Topic 1 semantic network

Table 2 Descriptive statistics of the topic 1 semantic network

Measures	Value
Number of nodes	496
Number of edges	7161
Average degree	28.875
Average weighted degree	22.371
Graph density	0.058
Modularity	0.665
Average clustering coefficient	0.559
Average path length	2.878

"duty of care" stood out in betweenness centrality, highlighting their role as core bridges between nodes.

Yet, while individually significant, the semantic interconnectedness between the three concepts is minimally found. Table 4 exhibits a thematic summary of the meaning clusters. The terms "duty of care" and "duty to explain" were included in one of the largest clusters, "Providers' Duties and Responsibilities" with 23.99% of node ratios. The two concepts were described in a context that highlights the patient risks and harms due to medical negligence. The cluster underscored the responsibility of medical professionals for proper treatments and explanations during medical procedures. "Self-determination" was included in the third cluster, "Patients' Rights" showing 20.97% of node ratio. This cluster identified key agents: "patient," "family," "parents," and "doctor." It suggests that patients, family members as caregivers, and doctors as medical providers play integral roles in medical decision-making.

Closer examination on three key concepts' neighbors also demonstrates that the concepts were articulated in distinct contexts. Figure 5 presents extracted ego-networks of three SDM concepts. For "duty of care," the words including "incidents," "worsening," "symptoms" and "practice" addressed detrimental impacts on patients from medical malpractices. For "duty to explain," the words including "damages," "aftereffects," "invasions," "risk" and "result" indicate that medical practitioners are required to provide explanation on possible adversities. For "self-determination," the words including "dignity," "choose," "patient," "consultation," "treatment," and "methods" illustrated that patients should make decision with caregiver engagements and doctor consultations. Each embedded subgraph of the three concepts revolved around their definitions in the SDM framework. However, they lack direct links and shared neighbors, which indicate minimal semantic ties.

Table 3 Node statistics of key SDM concepts in the topic 1 semantic network

Measures	Duty of care	Duty to explain	Self-determination
Degree	52	38	46
Weighted degree	31.143	29.293	42.579
Eigenvector centrality	0.186	0.148	0.488
Closeness centrality	0.422	0.378	0.352
Betweenness centrality	2008.929	1115.129	243.089

Table 4 Meaning clusters from the topic 1 semantic network

Cluster	Key nodes	*N* of nodes	Node ratio (%)
Symptom tracking and mortality risk	Work, pain, perforation, appropriate, time, dosing, side effects, hospitals, kill, delay, nursing, peritonitis, long-term, notices	125	25.20
Providers' duties and responsibilities	Restitution, damages, *duty of care*, losing, result, knowledge, *duty to explain*, negligence, causation, aftereffects, consultation, diagnostics	119	23.99
Patients' rights	Treatment, requirements, voluntary, protect, allow, requests, life, ventilators, remove, interruption, discharge, parent, family, *self-determination*, patient, doctor	104	20.97
Maternal and fetal health	Obstetrics and gynecology, maternity, fetus, cesarean section, blood pressure, embolization, choking, preterm birth	57	11.49
Surgical and pharmaceutical practice	Surgery, complications, scope, chronic, safety, effects, necrosis, pharmaceuticals, resection, vessels, ophthalmology, conjunctiva	57	11.49
Psychiatric hospitalization	Required, justification, limitations, hospitalization, force, mental, mental illness patient, unjust, facilities, consent form	29	5.85

3.2 Topic 2 Compensation Semantic Network Analysis

The semantic network of Topic 2 is provided in Fig. 6. Eight clusters are found here, with a 0.741 of modularity and a 0.51 of average clustering coefficient (Table 5). The statistics indicate that the Topic 2 semantic network was more explicitly clustered compared to Topic 1.

The node level statistics show the peripheral status of key SDM concepts in Topic 2 semantic network (Table 6). The concepts' low degrees, eigenvector centralities, and closeness centralities reflect the marginality of the three concepts. Exceptionally, 'self-determination' and 'duty to explain' demonstrate moderate betweenness centrality, indicating that they function as mediators despite not being central.

Fig. 5 Key neighbors of SDM concepts in the topic 1 semantic network

The cluster-level contexts confirm that the key SDM concepts are largely peripheral in the Topic 2 network. Table 7 presents a thematic summary of the meaning clusters. Compared to the Topic 1, no conspicuous major cluster was found. "Duty of care" appeared in the second largest cluster, "Surgical Procedures and Complications" with a 15.3% of node ratio. On the other hand, "duty to explain" and "self-determination" appeared in the same minor cluster, "Duty to Explain and Self-Determination" with a 10.48% of node ratio. The importance of "duty to explain" was tied to patients' decision-making right, reflected in their shared cluster membership and direct node connection. However, "duty of care," requiring medical practitioners to provide optimal care, was not discussed in the same context. The divergence underscores the lack of integration between the broader duty of care and the specifics of patient autonomy and communication. This gap suggests opportunities for healthcare frameworks to better align these essential aspects of SDM.

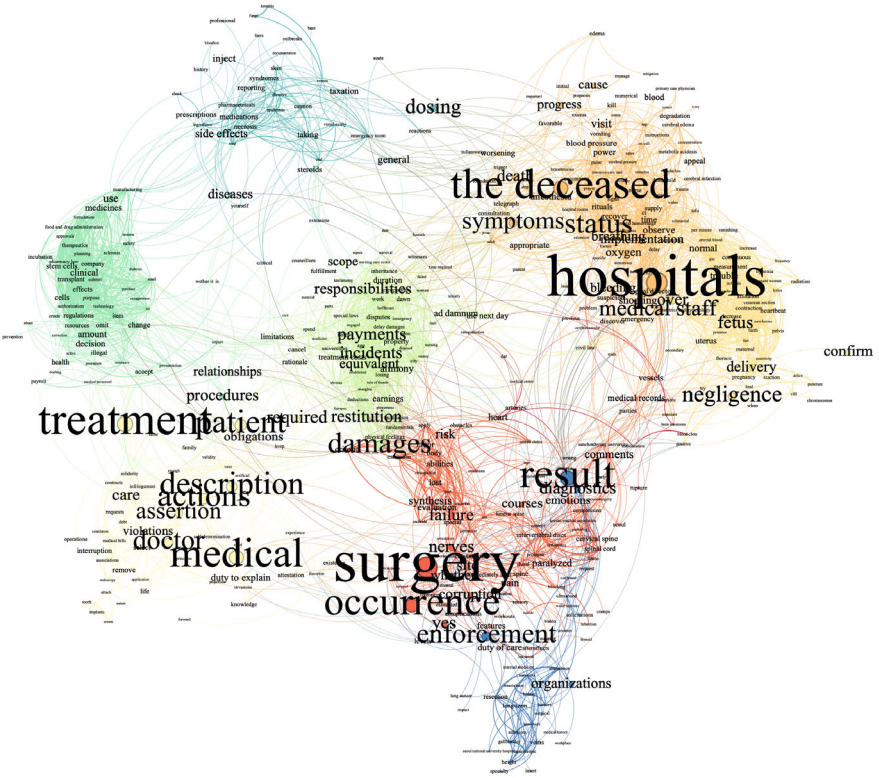


Fig. 6 Topic 2 semantic network

Table 5 Descriptive statistics of the topic 2 semantic network

Measures	Value
Number of nodes	477
Number of edges	3557
Average degree	14.914
Average weighted degree	11.508
Graph density	0.031
Modularity	0.741
Average clustering coefficient	0.51
Average path length	3.671

Figure 7 represents important meaning clusters around key terms related to SDM. It shows that "duty to explain" and "self-determination" were directly connected, sharing two neighbors: "doctor" and "violations." This implies that Topic 2 documents addressed the role of medical professionals' explanation in ensuring patients'

Table 6 Node statistics of key SDM concepts in the topic 2 semantic network

Measures	Duty of care	Duty to explain	Self-determination
Degree	1	8	12
Weighted degree	0.612	5.597	7.134
Eigenvector centrality	0.002	0.016	0.025
Closeness centrality	0.212	0.273	0.279
Betweenness centrality	0	937.013	1049.968

Table 7 Meaning clusters from the topic 2 semantic network

Cluster	Key nodes	*N* of nodes	Node ratio (%)
Vital sign tracking	Status, neurology, implementation, observe, breathing, oxygen, airway, signs, cerebral hemorrhage, CT, vitality, pulse, reflex	95	19.92
Surgical procedures and complications	Surgery, paralyzed, aftereffects, occurrence complications, failure, nerves, spine, intervertebral discs, risk, occupation, abilities, *duty of care*	73	15.3
Medical expenses and income loss compensation	Damages, restitution, payments, responsibilities, earnings, age, duration, equivalent, Hoffman, life expectancy, nursing	65	13.63
Transplant medicines and risks	Effects, company, clinical, safety, resources, technology, transplant, medicines, therapeutics, sclerosis, cells, approvals	53	11.11
Maternal and fetal health	Obstetrics and gynecology, measurement, fetus, newborns, uterus, late, trouble, hypoxia, cesarean section	51	10.69
Duty to explain and self-determination	Doctor, patient, treatment, obligations, description, *duty to explain, self-determination*, violations, medical bills, ventilators	50	10.48
Pharmaceutical side effect	Medications, necrosis, emergency room, side effects, pharmaceuticals, syndromes, caution, cornea, skin, epidermis	39	8.18
Surgical interventions	Result, diagnostics, resection, long-term, tumor, lung cancer, breast, internal medicine, adhesions, veins	38	7.97

decision-making rights. On the other hand, “duty of care” was minimally discussed, with no direct or indirect connection to other two key concepts. This limited engagement suggests that discussions in Topic 2 focus more on patient rights and communication than on the broader obligations of optimal care. This disparity points to a potential gap in integrating holistic care principles with patient-centered communication practices.

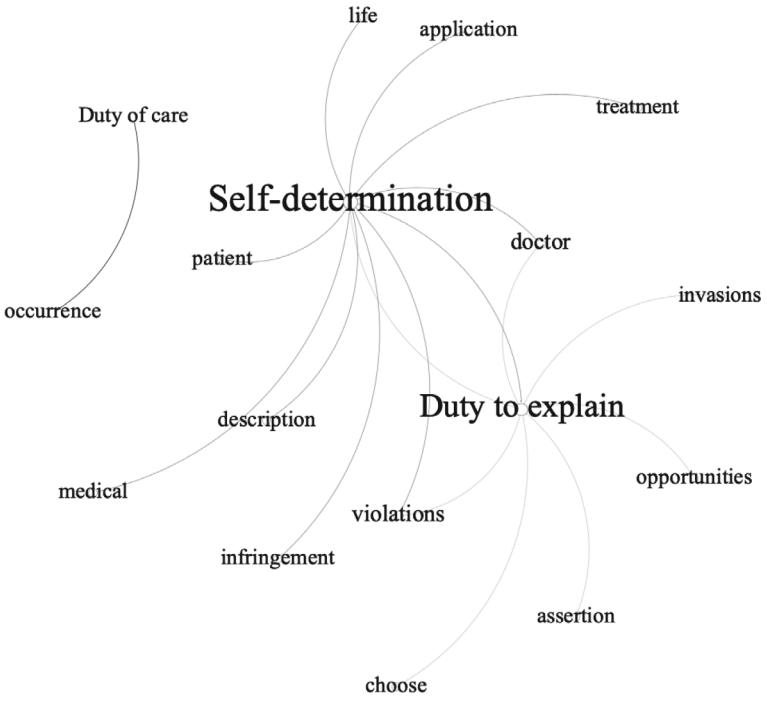


Fig. 7 Key neighbors of SDM concepts in the topic 2 semantic network

4 Conclusion

This study examined the unfolding of lexical structures of the medical lawsuit verdicts. We unveiled semantic networks of the key concepts of SDM. The topic diversity in medical conflicts has expanded over time. We also showed that the verdicts have two dominant topics: disputes over physicians' liability (Topic 1) and the disputes over the compensation related to malpractice (Topic 2). The semantic network analysis showed that not all core concepts of the shared decision-making, "duty of care," "duty to explain," and "self-determination," were closely connected.

From the SDM perspective, the three concepts should be closely associated in the meaning clusters for effective communications [15, 16]. In order to make decisions that maximize patients' interests reflecting their preferences and priority, medical professionals should provide enough explanations to support the optimal treatment options. However, from medical-legal conflict Topic 1, we found no direct link between "self-determination" and "duty of care," "duty to explain." The medical-legal Topic 2 semantic network included link between 'self-determination' and 'duty to explain,' however, they remained largely peripheral in the entire network. The rhetorics of patients' right to medical self-determination are not closely tied to the professional responsibility to explain and care.

This study has several limitations, some of which could be addressed in future research. The findings should be cautiously interpreted in terms of generalizability. We mostly dealt with legal verdicts from supreme courts, which may not cover a wide range of medical-legal conflicts. The data also may underrepresent disputes resolved in lower courts. Second, the exclusion of settlements, policies, and expert testimonies limits the scope of our analysis. Including these additional sources in future studies could provide a more comprehensive understanding of medical decision-making in Korea.

Despite these limitations, this study makes two significant contributions. First, we demonstrate the heuristic utility of medical verdict data through the implementation of semantic network analysis. Second, we unveil the semantic networks underlying medical-legal conflicts, highlighting the longitudinal changes in their lexical structures. Each meaning cluster has merged into a cohesive global structure, indicating the increased complexity and expanded scope of medical conflicts. In this study, we revealed barriers of SDM implementations. Through the semantic network analysis, we uncovered decoupled meaning clusters of healthcare professionals' duties and patients' rights.

Acknowledgements This research was supported by a grant of the Korea Health Technology R&D Project through the Patient-Doctor Shared Decision-Making Research center, funded by the Ministry of Health and Welfare, Republic of Korea (HI20C1234). This research was supported by a National Research Foundation of Korea grant funded by the Korean government (No. 2022R1A4A1033856).

References

1. Siebinga VY, Driever EM, Stiggelbout AM, Brand PLP (2022) Shared decision making, patient-centered communication and patient satisfaction: a cross-sectional analysis. Patient Educ Counsel 105:2145–2150
2. Ong LML, de Haes JCJM, Hoos AM, Lammes FB (1995) Doctor-patient communication: a review of the literature. Soc Sci Med 40:903–918
3. Joosten EA, DeFuentes-Merillas L, de Weert GH, Sensky T, van der Staak CP, de Jong CA (2008) Systematic review of the effects of shared decision-making on patient satisfaction, treatment adherence and health status. Psych Psych 77:219–226
4. Hughes TM, Merath K, Chen Q, Sun S, Palmer E, Idrees JJ, Okunrintemi V, Squires M, Beal EW, Pawlik TM (2018) Association of shared decision-making on patient-reported health outcomes and healthcare utilization. Am J Surg 216(1):7–12
5. Stiggelbout AM, Pieterse AH, De Haes JCJM (2015) Shared decision making: concepts, evidence, and practice. Patient Educ Counsel 98:1172–1179
6. Godolphin W (2009) Shared decision-making. Healthc Q 12:e186–e190
7. Sandman L, Granger BB, Ekman I, Munthe C (2012) Adherence, shared decision-making and patient autonomy. Med Health Care Philos 15:115–127
8. Waddell A, Lennox A, Spassova G, Bragge P (2021) Barriers and facilitators to shared decision-making in hospitals from policy to practice: a systematic review. Implement Sci 16:74
9. Steffensen KD, Vinter M, Crüger D, Dankl K, Coulter A, Stuart B, Berry LL (2018) Lessons in integrating shared decision-making into cancer care. J Oncol Pract 14:229–235

10. Boland L, Graham ID, Légaré F, Lewis K, Jull J, Shephard A, Lawson ML, Davis A, Yameogo A, Stacey D (2019) Barriers and facilitators of pediatric shared decision-making: a systematic review. Implement Sci 14:7
11. Covvey JR, Kamal KM, Gorse EE, Mehta Z, Dhumal T, Heidari E, Rao D, Zacker C (2019) Barriers and facilitators to shared decision-making in oncology: a systematic review of the literature. Support Care Cancer 27:1613–1637
12. Gwyn R, Elwyn G (1999) When is a shared decision not (quite) a shared decision? Negotiating preferences in a general practice encounter. Soc Sci Med 49:437–447
13. Joseph-Williams N, Elwyn G, Edwards A (2014) Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. Patient Educ Counsel 94:291–309
14. Timmermans S (2020) The engaged patient: the relevance of patient–physician communication for twenty-first-century health. J Health Soc Behav 61:259–273
15. Szalados JE (2021) The ethics and laws governing informed decision-making in healthcare: informed consent, refusal, and discussions regarding resuscitation and life-sustaining treatment. In: Szalados JE (ed) The medical-legal aspects of acute care medicine: a resource for clinicians, administrators, and risk managers. Springer, Cham, pp 43–73
16. Cave E (2020) Selecting treatment options and choosing between them: delineating patient and professional autonomy in shared decision-making. Health Care Anal 28:4–24
17. Jo KH (2012) Development and evaluation of shared medical decision-making scale for end-of-life patients in Korea. J Korean Acad Nurs 42:453–465
18. Lee I (2023) South Korea's end-of-life care decisions act: law for better end-of-life care. In: Cheung D, Dunn M (eds) Advance directives across Asia: a comparative socio-legal analysis. Cambridge University Press, Cambridge, pp 57–74
19. Mo HN, Shin DW, Woo JH, Choi JY, Kang J, Baik YJ, Huh YR, Won JH, Park MH, Cho SH (2012) Is patient autonomy a critical determinant of quality of life in Korea? End-of-life decision making from the perspective of the patient. Palliat Med 26:222–231
20. Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I (2020) LEGAL-BERT: the muppets straight out of law school. arXiv preprint arXiv:2010.02559
21. Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794
22. Doerfel ML (1998) What constitutes semantic network analysis? A comparison of research and methodologies. Connections 21:16–26
23. Christensen AP, Kenett YN (2023) Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. Psychol Methods 28:860–879
24. Breucker P, Cointet J, Hannud Abdo A, Orsal G, de Quatrebarbes C, Duong T, Martinez C, Ospina Delgado JP, Medina Zuluaga LD, Gómez Peña DF, Sánchez Castaño TA, Marques da Costa J, Laglil H, Villard L, Barbier M (2016) CorTexT Manager (version v2). https://docs.cortext.net
25. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: International AAAI conference on weblogs and social media

Understanding ENSO Teleconnections' Influence on Drought in Southern Africa: A Machine Learning Approach

Jimmy Katambo, Gloria Iyawa, Lars Ribbe, and Victor Kongo

Abstract The vulnerability of Southern Africa to climate variability, especially drought, places substantial pressure on agriculture, water systems, and the economy. This study explores how El Niño-Southern Oscillation (ENSO)-related Sea Surface Temperature (SST) variations influence drought patterns across the region using machine learning methods. Two approaches were taken: (i) a feature ranking of SST in comparison to twelve other climate variables and (ii) drought model performance comparisons with and without SST data. Results reveal SST's significant and consistent impact across all climate zones, with both methods indicating that SST data, particularly in connection with ENSO phases, strongly influences drought variability, despite slight variations in its order of effect with respect to climatic zonal divisions. This underscores the value of incorporating SST in climate models for enhanced drought prediction and adaptation planning. Although limited by a focus on SST and not fully accounting for interactions with other climate factors, this research provides a solid foundation for understanding regional climate dynamics. Adding more climate indicators and studying SST's interactions with land-based factors could help future studies make drought predictions more reliable and better prepare vulnerable areas.

Keywords SST · ENSO · Machine learning · Transformer architecture · Climate zones · Data

J. Katambo (✉)
Namibia University of Science and Technology, Windhoek, Namibia
e-mail: jimmy.katambo@cs.unza.zm

G. Iyawa
University of Salford, Salford, UK
e-mail: g.e.iyawa@salford.ac.uk

L. Ribbe
TH Köln, University of Applied Sciences, Cologne, Germany
e-mail: lars.ribbe@th-koeln.de

V. Kongo
Global Water Partnership Tanzania, Dar es Salaam, Tanzania

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_11

1 Introduction

Southern Africa is one of the most climate-sensitive regions in the world, particularly vulnerable to drought, which has far-reaching impacts on agriculture, water resources, and livelihoods. Climate variability in this region is strongly influenced by global ocean–atmosphere phenomena, with the El Niño-Southern Oscillation (ENSO) being a primary driver of climatic fluctuations [1]. ENSO events [2], characterized by shifts in sea surface temperatures (SST) in the Pacific Ocean, lead to teleconnections that impact rainfall and temperature patterns across Southern Africa. Zhao et al. [3] observed that ENSO affects the world's major river basins. Understanding the complex influence of ENSO on regional drought conditions is critical for developing accurate predictive models that support early warning systems and inform adaptation strategies.

A large body of research [4–8], has demonstrated that Sea Surface Temperature (SST), particularly in relation to ENSO phases, plays a significant role in modulating climate variability in Southern Africa. Despite these advances, there is a need to refine our understanding of how SST variations interact with local climatic processes to influence droughts at a more detailed level. Machine learning offers promising tools for analyzing complex and high-dimensional climate data, enabling the exploration of intricate relationships between SST patterns and climate responses across different zones. However, studies specifically leveraging machine learning approaches to assess the role of SST in regional drought dynamics remain limited, particularly in Southern Africa, where diverse climate zones may exhibit varying responses to SST fluctuations and ENSO events.

In analyzing the drought-ENSO relationship, [4, 9] suggest the Standardized Precipitation Evapotranspiration Index (SPEI) as one of the preferred indices. Manatsa et al. [9] argues that SPEI, which factors in both precipitation and potential evapotranspiration, captures increased water demand resulting from rising temperatures and incorporates both precipitation and drought severity linked to temperature variability. Gore et al. [4] also supports this view, noting that SPEI's foundation in climate water balance makes it effective for quantifying drought with both variables in mind.

The selection of the machine learning algorithm was guided by a combination of prior research insights and a preliminary practical test evaluation of three options: Random Forest [10, 11], Feedforward Neural Network (FFNN) [12], and transformer architecture [13–15]. The algorithm that showed the best results was then utilized for all further research trials. This initial evaluation ensured that the chosen model was best suited to handle the dataset's complexity and variability, providing an optimal balance of accuracy and computational efficiency. By using an evidence-based approach to model selection, the study focused on maximizing predictive performance while minimizing potential overfitting and resource use during the experiments.

The main objective of this paper was to investigate the teleconnection relationships between El Niño Southern Oscillation (ENSO) and SPEI drought across the

different regions of Southern Africa. To address this objective, machine learning was used to quantify and rank the influence of SST data on drought conditions across Southern Africa's climate zones. Two primary novel methods are employed: (i) a feature importance analysis to evaluate SST's relative influence compared to twelve other climatic variables, and (ii) a comparative model approach that contrasts setups with and without SST data to assess its overall impact. Through these methods, this study seeks to determine which aspects of ENSO-related SST changes most strongly correlate with drought, and to identify the most sensitive regions for these effects.

This study provides valuable insights into the primary climate drivers influencing drought in Southern Africa, with a particular emphasis on the consistent role of Sea Surface Temperature (SST) across climate zones. By demonstrating SST's impact on drought variability and underscoring its utility in predictive climate models, this research enhances the accuracy of drought forecasting, which is essential for effective planning and adaptation.

The findings hold significant implications for decision-making, especially for farmers and other stakeholders who rely on seasonal forecasts to guide their actions. With improved predictions, stakeholders can plan strategically: for instance, investing in crop diversity and additional inputs in favorable seasons to boost yields, or minimizing investments in anticipated poor seasons to reduce financial risk. This research thus establishes a foundation for understanding climate variability in Southern Africa and highlights the need for further exploration into how global phenomena, such as ENSO, interact with local climatic factors to shape drought risk. Overall, this research provides practical knowledge that supports both scientific growth and effective strategies for adapting to changing climates.

2 Methodology

2.1 Study Area

Southern Africa, spanning from the equator to 40° South, experiences extreme climate variability, including frequent floods and droughts. The Kalahari Desert, covering parts of South Africa, Namibia, and Botswana, reaches temperatures over 40 °C, with wide diurnal variations [16]. The region has a mild summer wet season (November–March) and a dry winter (April–October), while the southwest exhibits a Mediterranean climate with winter rains. Climate variability here is influenced by atmospheric pressure shifts, the Intertropical Convergence Zone (ITCZ), and ocean currents from the Atlantic and Indian Oceans, creating diverse climatic zones across the area [17, 18].

Southern Africa's Okavango, Limpopo, Orange, Save, and Zambezi River basins are vital for the region's economy, hosting industries and attracting tourism. The Orange River basin spans ~1,000,000 km^2 across Lesotho, South Africa, Botswana,

and Namibia, with extreme temperature variations, high rainfall in the highlands, and high evaporation rates [16, 19, 20].

Covering over 415,000 km^2, the Limpopo River basin spans Zimbabwe, South Africa, Botswana, and Mozambique. Rainfall varies from 200 mm in the west to 1500 mm in the east, mostly falling between October and April, with high variability leading to both floods and droughts. Temperatures range between 0 and 36 °C, and annual evaporation averages 1.970 mm [21, 22].

Encompassing Namibia, Botswana, Angola, Zambia, and South Africa, the Okavango basin has a sandy savannah. Its water from the Okavango Delta in Botswana evaporates in cooler months. The Zambezi River basin, covering 1.37 million km^2 across eight countries, is the largest in Southern Africa, with annual rainfall ranging from 700 to 1200 mm and discharges to the Indian Ocean [16, 20, 23].

Southern Africa features diverse climates, ranging from desert heat in the Kalahari to Mediterranean zones in the southwest. The region experiences a summer wet season (November–March) and a dry winter (April–October). Influenced by atmospheric pressure, the ITCZ, and ocean currents, it faces significant climate variability, including floods and droughts. Key river basins like the Okavango, Limpopo, Orange, and Zambezi support economies, agriculture, and tourism. The Limpopo basin spans four countries, with rainfall between 200 and 1500 mm and evaporation averaging 1970 mm annually. The Zambezi, Southern Africa's largest basin, covers 1.37 million km2, discharging to the Indian Ocean. Figure 1 below displays the map of Southern Africa on which this research was based.

2.2 Data Sources

Climate station historical data on a monthly time scale was downloaded from 1972 to 2022 from the SASSCAL Information and Data portal [24]. Variables such as relative humidity (RH) (%), wind speed at 2 m (m/s), sunshine hours (SS) (h), solar radiation (SR) (W/m^2), maximum temperature (T_{max}), minimum temperature (T_{min}), average temperature (T_{avg}) (°C) and precipitation amount (P) (mm) were collected. Station coordinates from the 164 weather stations were used as selected data points.

Then a procedure of mapping the stations was done using QGIS software by adding a layer of coordinates on the shape file representing the Southern African region. The climatic zones for Southern Africa were accessed from global maps of the Köppen-Geiger climate classification [25]. To create the climatic zones displayed in Fig. 2, the process was carried out using QGIS. Initially, a shapefile of Southern Africa was added. Next, a TIFF layer generated from [25] representing the climatic zones was incorporated. Finally, a CSV file containing the station coordinates was added as another layer. The original climate zone divisions are based on threshold values and the seasonality of monthly air temperature and precipitation [25].

Figure 2 under the results section shows the station coordinates and the eight climatic zones onto which these were mapped. The rest of the experiments were then done according to the eight climatic zones. Then using the coordinates for each zone,

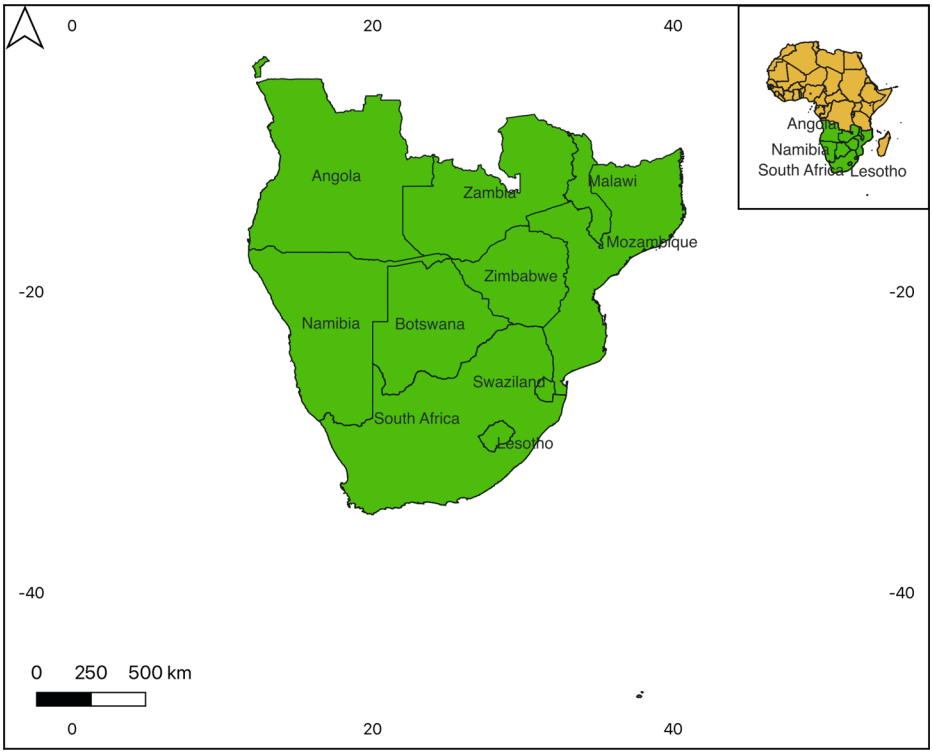


Fig. 1 Geographic representation of Southern Africa, highlighting its constituent countries

python codes were used to extract data from NETCDF files that were downloaded from satellite products available such as ERA5 from Copernicus [26, 27]. While variables such as wind speed, relative humidity, solar radiation, precipitation, potential evapotranspiration, minimum, maximum, and average temperature were from ERA5, two additional variables namely: soil moisture, and sunshine hours were also added to the feature set. Data from 1981 to 2022 for soil moisture was available as satellite observed data from NASA [28]. The deficit from 1972 to 1980 was generated synthetically using the Generative Adversarial Networks (GANs). For sunshine hours, additional data was generated by using the available insitu data from some stations and then applying the GANs to generate the data for the remaining cells.

It must be emphasized that the initial SASSCAL [24] in situ station data played a role only in validating the satellite data. Once validated, the satellite data was used exclusively for all remaining experiments. This approach was essential, as the SASSCAL data had substantial gaps and missing values, making satellite data the more reliable option.

Concerning the calculation of SPEI values, this research determined SPEI across 1-, 3-, and 6-month intervals for each station, with coordinates represented within

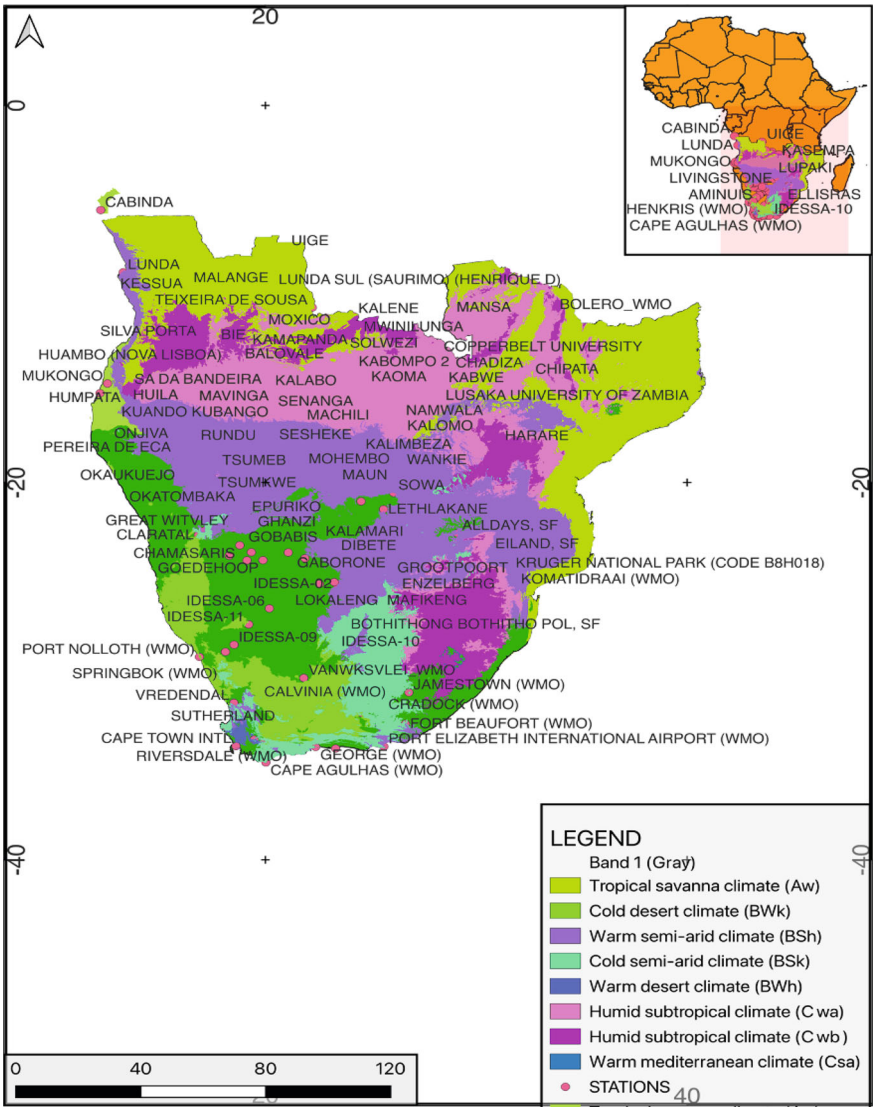


Fig. 2 Distribution of coordinates based on climatic zonal mapping using Köppen-Geiger climate classification

each zone. Following this, the average SPEI for all stations in each zone was calculated to facilitate further analysis, such as machine learning experiments. These calculations were performed using the SPEI package in the R software.

2.3 Machine Learning Algorithms

In the first stage of this section of the study, three machine learning models (Random Forest [10, 11], Feedforward Neural Networks (FFNNs) [12] and transformer architecture [13–15]) were evaluated to determine which would be best suited for assessing the effect of SST data. This preliminary test also aimed to identify the most effective Standardized Precipitation Evapotranspiration Index (SPEI) scale (SPEI 1, SPEI 3, or SPEI 6) for the data used. One of the reasons for choosing the SPEI index is because it accounts for both precipitation and potential evapotranspiration. As seen in Fig. 4, the transformer model paired with SPEI 6 was selected for further experimentation.

The transformer model utilized in this current study featured a simple architecture with three fully connected layers. The first layer took in input features and produced 32 units, the second layer reduced it to 16 units, and the final layer output a single unit for binary classification. ReLU activation in the first two layers enabled the model to capture complex patterns. The model used the Adam optimizer with a learning rate of 0.01 and a weight decay of 1e−5 to control overfitting. To further stabilize training, a learning rate scheduler reduced the rate every 20 epochs, and gradient clipping kept gradients within a safe range, preventing them from growing too large. Training was conducted over 50 epochs with a batch size of 8, and efficient data handling was achieved with PyTorch's DataLoader. Adjustments in epoch count and batch size also reduced memory demands, preventing kernel crashes and optimizing resource use.

Exploring How El Niño Southern Oscillation (ENSO) Teleconnections Affect Drought in Southern Africa: Sea Surface Temperature (SST) data was downloaded as a netCDF file from National Oceanic and Atmospheric Administration (NOAA) [29]. A basic Python algorithm was used to extract SST values from the netCDF file for all 164 points in the original station dataset. Figure 8 in the results section shows which stations had SST data.

After downloading the SST data from 1972 to 2022 for the selected stations containing SST, an average was computed across all stations. This averaged data was then incorporated as a new variable in the feature set for each of the modeling experiments conducted in the eight zones. As a result, the feature set expanded from 11 variables to 12. The modeling experiment involving SST was performed in all the zones.

This analysis involved the use of permutation importance function applied to neural networks. Permutation importance function from sklearn calculates the feature importances correctly, ensuring that the metric is computed on the entire dataset for each class. This method evaluates the contribution of each feature by measuring how the model's performance decreases when the values of that feature are randomly shuffled. By disrupting the relationship between the feature and the target variable, permutation importance provides insight into how much the model relies on each feature for making accurate predictions.

The process began with training a model on the dataset using the original feature values. After the model was trained, its performance was evaluated using a suitable

metric, such as accuracy, F1 score, or mean squared error, tailored to the classification task in this study. Once the baseline performance was established, permutation importance was calculated by systematically permuting the values of each feature, one at a time, and measuring the change in model performance.

For each feature, the values were randomly shuffled, breaking the association between the feature and the target. The model was then re-evaluated using the permuted dataset. The decrease in performance compared to the original model indicates the importance of that feature; a significant drop suggests that the feature plays a crucial role in the model's predictions. This process was repeated multiple times to account for randomness and variability, and the results were averaged to obtain a stable measure of feature importance.

The final output of the permutation importance calculation is a set of importance scores for each feature. These scores can be visualized in a ranking format, allowing for easy identification of the most influential features. Features with high importance scores are critical to the model's predictive power, while those with low scores contribute less to the model's performance. This technique was realized by embedding segments of code to compute permutation importances within the core code of the transformer algorithm.

The second analysis was performed on the dataset for each zone and consisted of two experiments. The first experiment used the feature set without SST data, while the second extended the feature set to 12 variables by adding SST data. The performance metrics of the transformer model were compared between the two experiments to investigate the role of SST data in drought modeling. This process involved calculating the differences for each metric across zones, comparing outcomes when SST was included versus excluded. Figure 8 in the results section represents these values as variances for R, NSE, and accuracy.

3 Results and Discussion

Southern Africa was divided into eight climatic zones by adding a coordinate layer for station locations as demonstrated in Fig. 2. This zonal framework provided a structured basis for calculating the SPEI and carrying out the modeling experiments.

Figure 3 displays the coordinates where Sea Surface Temperature (SST) data was retrieved. Out of 164 stations, only 18 had complete SST data, while the remaining 146 were either marked with zeros or NaN values. This indicates that SST data is mostly available near the sea or ocean. Additionally, the map emphasizes that the data collected from [24] mainly covered locations close to the Atlantic Ocean on the left side of Southern Africa, with no stations near the Indian Ocean due to the nature of the original station dataset.

Results from the preliminary experiment aimed at selecting the optimal machine learning algorithm and SPEI index scale are presented in Fig. 4. The figure shows that the transformer model outperformed the other two algorithms, and SPEI6 emerged as the most effective scale. Across all initial tests for SPEI1, SPEI3, and SPEI6, the

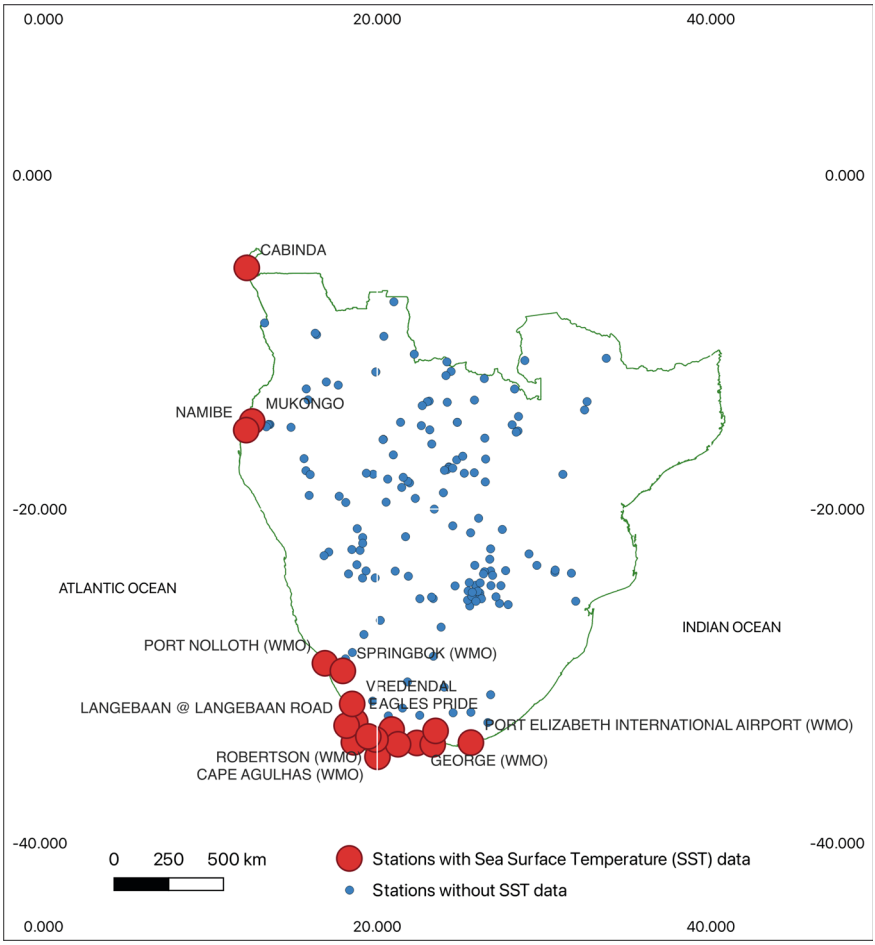
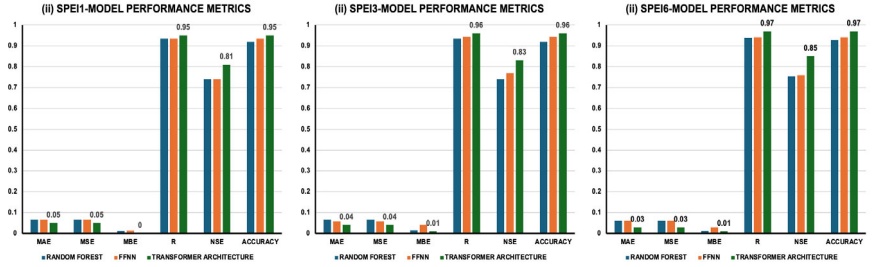


Fig. 3 Spatial arrangement of 18 weather stations with sea surface temperature data from a pool of 164 stations

transformer consistently delivered superior results, with SPEI6 achieving the highest performance metrics among all scales.

The experimental outcomes for each zone, depicted in Fig. 5a(i–iv) and b(v–viii), were generated using a neural network permutation importance approach with SST data included in the feature sets. Each figure illustrates SST's relative ranking among twelve variables. A final, combined ranking of SST across zones is provided in Fig. 6 after comparing the charts in Fig. 5a and b.

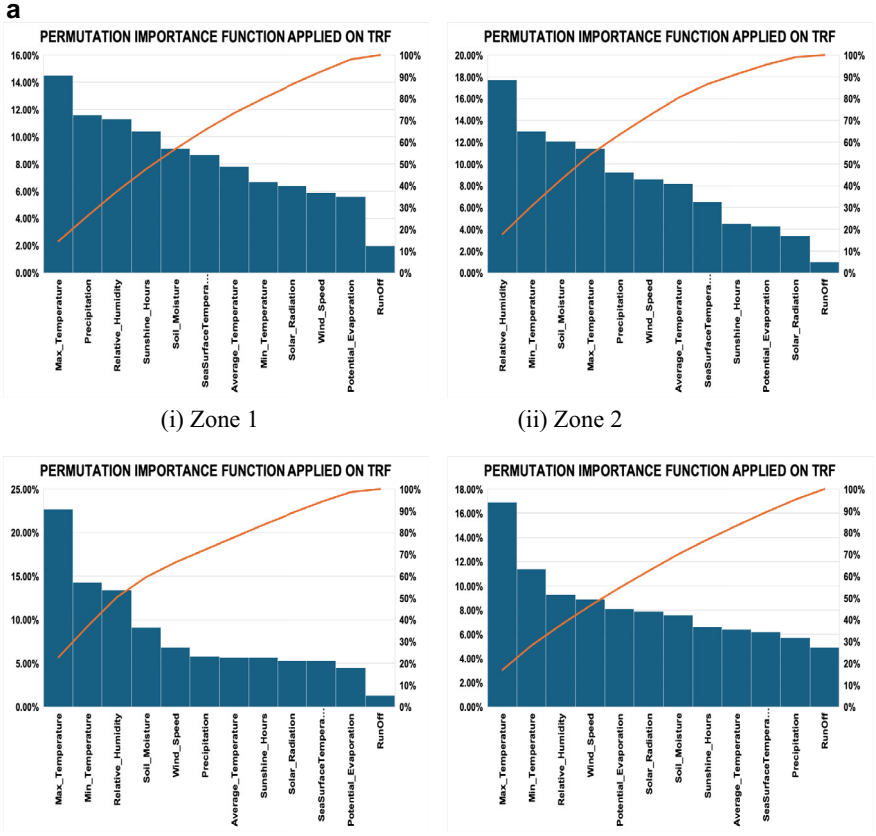
According to Fig. 6, SST data achieved its highest ranking in Zone 7 [Cold Semi-Arid Climate (BSk)] and Zone 8 [Warm Mediterranean Climate (Csa)], both at fourth place among 12 variables. Zone 6 [Cold Desert Climate (BWk)] ranked SST fifth, Zone 1 [Humid Subtropical Climate (Cwa)] ranked it sixth, and Zone 5 [Humid



(a) (b) (c)

Fig. 4 Zone 1's performance metrics for **a** SPEI1, **b** SPEI3, and **c** SPEI6, ranked from least to most effective according to the metrics

a



(i) Zone 1 (ii) Zone 2

(iii) Zone 3 (iv) Zone 4

Fig. 5 **a**(i–iv) Results of the permutation importance analysis from the transformer algorithm after adding SST data to each zone's feature set. **b**(v–viii) Results of the permutation importance analysis from the transformer algorithm after adding SST data to each zone's feature set

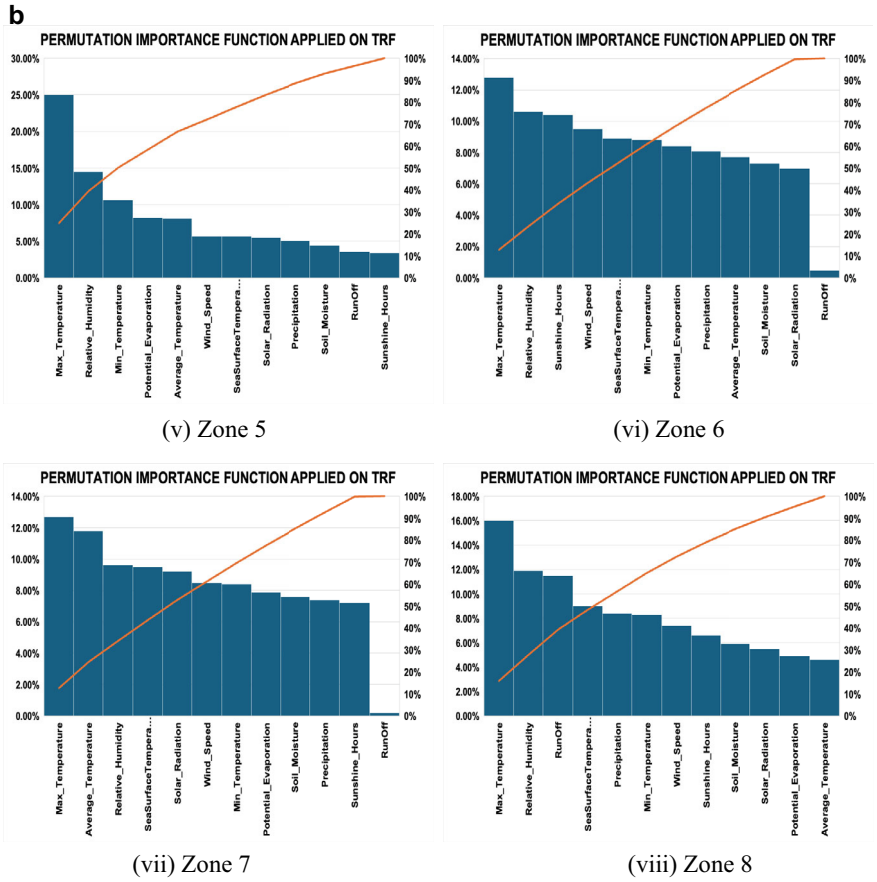


Fig. 5 (continued)

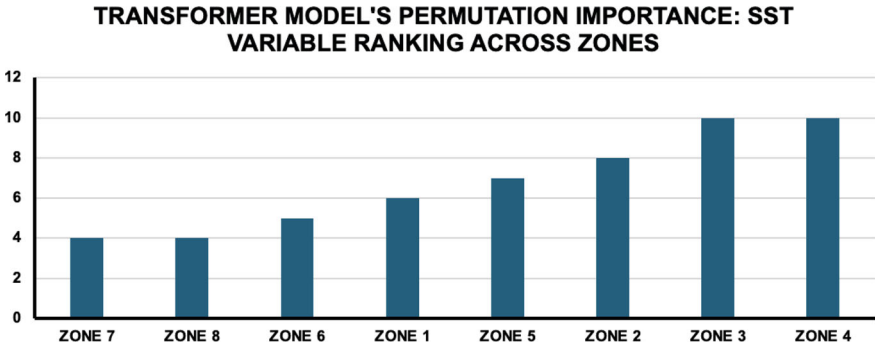


Fig. 6 Position of SST variable among 12 variables by transformer model permutation importance, by zone

Subtropical/Subtropical Oceanic Highland (Cwb)] placed it seventh. Zone 2 [Tropical Savanna Climate (Aw)] ranked SST eighth, while Zones 3 [Warm Semi-Arid (BSh)] and 4 [Warm Desert (BWh)] both placed it at tenth.

These findings illustrate SST's impact on drought within each Southern African climate zone. Zones 7 and 8 were the most influenced by El Niño, followed by Zone 6. Zone 1 was the third most affected, Zone 5 the fourth, Zone 2 the fifth, while Zones 3 and 4 shared the lowest level of impact.

In the second approach shown in Fig. 7, the model demonstrates El Niño's effects across different zones, with clear improvements in most performance metrics following the inclusion of SST data compared to the initial model without SST data. The charts illustrate that all metrics showed gains when comparing results from experiments conducted with and without SST data.

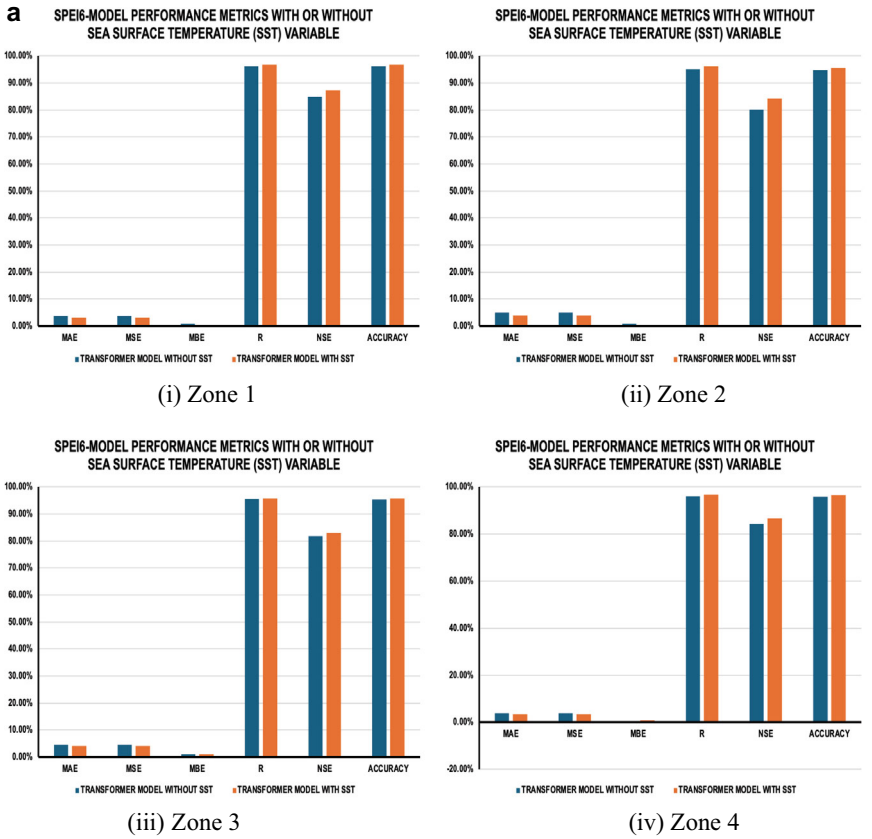
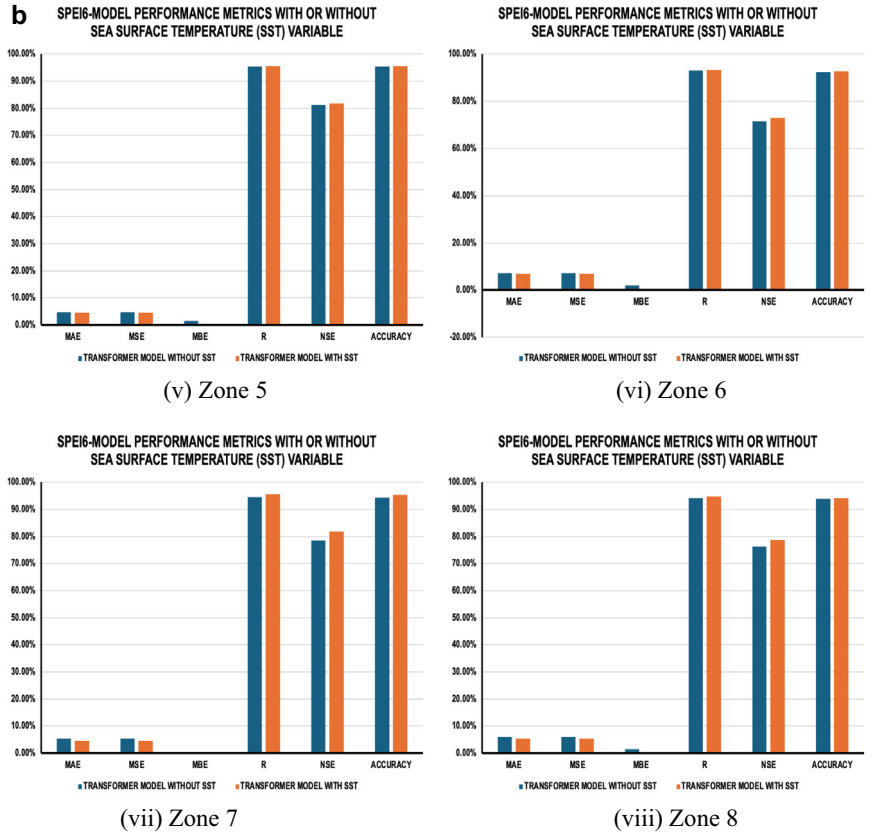


Fig. 7 **a**(i–iv) Transformer model performance for each zone, shown for two cases: (1) without SST and (2) with SST data included in the feature set. **b**(v–viii) Transformer model performance for each zone, shown for two cases: (1) without SST and (2) with SST data included in the feature set



(v) Zone 5 (vi) Zone 6

(vii) Zone 7 (viii) Zone 8

Fig. 7 (continued)

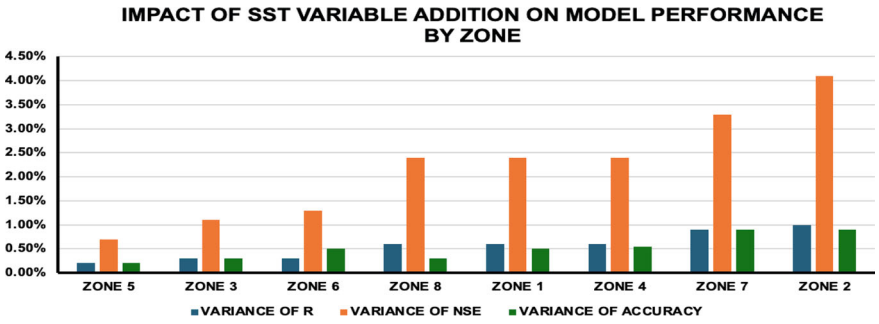


Fig. 8 Performance comparison of models across zones following the addition of the SST variable

Generally, in modeling, the rule of the thumb is that performance improves with higher Correlation Coefficient (*R*), accuracy, and Nash–Sutcliffe Efficiency (NSE) values, ideally close to 1 (1 = perfect, >0.75 = very good, 0.64–0.74 = good, 0.5–0.64 = satisfactory, <0.5 = unsatisfactory; [11]). In contrast, lower Mean Bias Error (MBE), Mean Squared Error (MSE), and Mean Absolute Error (MAE) values near 0 indicate better performance. According to Fig. 7, in Zone 1, the *R* metric rose from 96.2 to 96.8%, NSE increased from 84.9 to 87.3%, and accuracy improved from 96.2 to 96.7%. Likewise, MBE dropped from 0.9 to 0.08%, while both MSE and MAE improved from 3.8 to 3.2%.

For Zone 2, improvements included an increase in *R* from 95.1 to 96.1%, NSE from 80.2 to 84.3%, and accuracy from 94.7 to 95.6%, alongside a reduction in MBE from 0.9 to 0.4% and in MSE and MAE from 4.9 to 3.9%. Zone 3 saw R rise to 95.8% (from 95.5%), NSE to 83.0% (from 81.9%), and accuracy to 95.7% (from 95.4%), with MBE unchanged at 1% and MSE and MAE declining to 4.2% (from 4.5%). Zone 4 recorded increases with *R* reaching 96.7% (from 96.1%), NSE 86.7% (from 84.3%), and accuracy 96.5% (from 95.9%), while MBE moved from −0.5 to 0.9% and MSE and MAE reduced to 3.3% from 3.9%.

Zone 5 showed *R* increasing to 95.5% from 95.3%, NSE to 81.9% from 81.2%, and accuracy to 95.5% from 95.3%. MBE declined from 1.5 to 0.3%, and MSE and MAE both improved from 4.7 to 4.5%. Zone 6 recorded an *R* increase to 93.2% (from 92.9%), NSE to 72.9% (from 71.6%), and accuracy to 92.7% (from 92.2%). MBE reduced from 1.9 to −0.4%, while MSE and MAE dropped to 6.8% from 7.1%. Zone 7 saw R rise to 95.5% from 94.6%, NSE to 81.9% from 78.6%, and accuracy to 95.3% from 94.4%, with MBE being held at 0.5% and MSE and MAE reduced to 4.5% from 5.3%. Zone 8 recorded gains with R up to 94.7% (from 94.1%), NSE to 78.7% (from 76.3%), and accuracy to 94.2% (from 93.9%). MBE changed from 1.5 to 0%, with MSE and MAE both dropping to 5.3% (from 5.9%).

To make sense of the findings in Fig. 7, a comparative analysis between the setups—(i) without SST data and (ii) with SST data added to each zone's features—is essential, as shown in Fig. 8. This figure details the variance in key metrics like *R*, NSE, and accuracy across the scenarios.

According to Fig. 8, the variance across metrics increases sequentially from zone 5 up through zones 3, 6, 8, 1, 4, 7, with zone 2 showing the largest effect. These results suggest that SST data significantly impacts zone 2, while zones with lower variance, such as zone 5, experience less effect. This trend illustrates an increasing impact of SST, moving from the least affected zone (zone 5) up to the most impacted (zone 2), underscoring the overall significance of SST on each zone's performance and its influence on all analyzed zones.

The two methods employed—(i) a feature importance function to rank SST data among twelve variables per zone and (ii) a model approach comparing setups with and without SST—yielded similar insights. Both approaches demonstrate SST's impact across Southern Africa's climate zones. Although the order of effect varies slightly between the methods, both clearly indicate that SST data, and consequently El Niño, influence all these zones.

The feature importance function ranked El Niño's impact in this order: Zones 7 and 8 with identical outcomes as most affected, followed by Zone 6, then Zone 1, with Zone 5 fourth, and Zone 2 fifth. Zones 3 and 4 with identical values showed the lowest effects overall. In contrast, the model approach, which used setups with and without SST data, indicated an impact order from Zone 5 through Zones 3, 6, 8, 1, 4, 7, to Zone 2 as most affected. Therefore, it is important to note that these differences in results may stem from randomness in algorithms, such as permutation importance, where feature values are shuffled multiple times to capture variability.

Other studies, such as [7, 30], found a strong link between SST data and rainfall. Likewise, [4] employed the SPEI to show that ENSO plays a significant role in predicting southern Africa's climate. Moreover, studies such as [6] indicate that SST effects can differ across various regions and sub-regions. While most similar studies have been limited to country-level scales, our work examined a large geographical region divided into climatic zones. This research was unique in using two approaches: (i) a feature importance analysis to rank SST data against twelve additional variables for each climate zone, and (ii) evaluating comparative machine learning models with and without SST data to determine its influence on Southern Africa's diverse climates.

In this study, we focused on machine learning (ML) methods due to their ability to handle large datasets and capture non-linear relationships, which are often challenging for traditional physical models. While ML models have shown great promise in improving drought forecasting, it is important to recognize the strengths of traditional physical models, which excel in representing the underlying physical processes of drought events. Although a direct comparison with physical models was beyond the scope of this study, future research could benefit from integrating these approaches to benchmark ML models against established methodologies.

While the methods employed in this study demonstrated strong predictive capabilities in Southern Africa, their applicability to other regions depends on the presence of enough data, the reliability of that data, and the unique regional climate and environmental characteristics. Machine learning models, by design, are highly adaptable and can be retrained with region-specific data, which suggests that the approach could be extended globally to other drought-prone areas. However, variations in climatic conditions, soil types, vegetation, and water resource management practices across regions may affect model performance. By tailoring the model's inputs and fine-tuning its parameters to account for region-specific variables, the framework presented here could serve as a foundation for broader application.

4 Conclusion

In conclusion, this study employed two robust methods—(i) a feature importance analysis to rank SST data against twelve other variables per climate zone, and (ii) a comparative model analysis assessing setups with and without SST data—to evaluate the impact of SST on Southern Africa's diverse climate zones. Both approaches

yielded consistent findings, affirming SST's significant influence across all zones despite minor variations in the order of effect. These results underscore the consistent and widespread role of SST data, and by extension, El Niño events, in shaping regional climate dynamics. This reinforces the critical need to incorporate SST-related data into climate models for Southern Africa, particularly in predictive efforts related to droughts and rainfall variability.

Despite the robustness of these findings, certain limitations should be acknowledged. While SST was analyzed among other variables, interactions between SST and other climate drivers were not exhaustively modeled, leaving room to further explore complex interdependencies that may also influence climate variability. Additionally, the spatial and temporal resolutions of the SST and climate data used may limit the ability to capture localized or short-term climate impacts. Another limitation lies in the model's generalizability, as it may not fully represent variations across all sub-regions within Southern Africa, where microclimates or other local factors might moderate or amplify the effects observed.

Future research could enhance these findings by integrating additional climate indicators or exploring SST interactions with land-based climatic factors to deepen our understanding of climate drivers in this region. The study uses a limited dataset in terms of years analyzed. Extending the temporal range could provide more long-term insights into climate patterns and trends. Increasing data granularity or expanding the dataset to include additional years could also provide more comprehensive insights. The analysis might benefit from higher spatial and temporal resolution of the data. This would enable the capture of localized or short-term climate impacts that might be lost in broader-scale models. Altogether, this study provides a strong foundation for understanding and modeling climate impacts, including drought, in Southern Africa, which is essential for developing effective adaptation and mitigation strategies in response to increasing climate variability and drought risk.

Acknowledgements The authors acknowledge funding for this project, by the Southern African Science Service Centre for Climate Change and Adaptive Land Management (SASSCAL) sponsored by the German Government through the Federal Ministry of Education and Research (BMBF) with funding No: 01LG2091A.

References

1. Franchi F, Mustafa S, Ariztegui D, Chirindja FJ, Di Capua A, Hussey S, Loizeau JL, Maselli V, Matanó A, Olabode O, Pasqualotto F, Sengwei W, Tirivarombo S, Van Loon AF, Comte JC (2024) Sci Total Environ 924
2. Coughlan de Perez E, Anderson W, Han E, Masukwedza GIT, Mphonyane N (2024) Clim Serv 36
3. Zhao T, Li X, Li Y, Zhang B, Zhang Y (2024) J Hydrol 644
4. Gore M, Abiodun BJ, Kucharski F (2020) Clim Dyn 54:307–327
5. Gobie BG, Miheretu BA (2021) Model Earth Syst Environ 7:2733–2739
6. Rouault M, Tomety FS. https://doi.org/10.1175/JPO-D-21

7. Information Z, Technology C (2018) Forecasting seasonal rainfall in Zambia: an artificial neural network approach 2
8. Chou C, Marcos-Matamoros R, González-Reviriego N, Miravet AS (2024) Sci Total Environ 951
9. Manatsa D, Mushore T, Lenouo A. https://doi.org/10.1007/s00704-015-1632-6/Published
10. Elbeltagi A, Kumari N, Dharpure JK, Mokhtar A, Alsafadi K, Kumar M, Mehdinejadiani B, Ramezani Etedali H, Brouziyne Y, Towfiqul Islam ARM, Kuriqi A (2021) Water 13:1–18
11. Mokhtar A, Jalali M, He H, Al-Ansari N, Elbeltagi A, Alsafadi K, Abdo HG, Sammen SS, Gyasi-Agyei Y, Rodrigo-Comino J (2021) IEEE Access 9:65503–65523
12. Mumbi AW, Li F, Bavumiragira JP, Fangninou FF (2022) Mar Freshw Res 73:292–306
13. Veltman A, Pulle DWJ, De Doncker RW (2017) Power Syst 2017:47–82
14. Minixhofer C, Swan M, Mcmeekin C, Andreadis P (2021)
15. Wen Q, Zhou T, Zhang C, Chen W, Ma Z, Yan J, Sun L (2022)
16. Abiodun BJ, Makhanya N, Petja B, Abatan AA, Oguntunde PG (2019) Theor Appl Climatol 137:1785–1799
17. Davis-Reddy CL, Vincent K (2017) Climate risk and vulnerability: a handbook for Southern Africa, 2nd edn. CSIR, Pretoria
18. Yim BY, Yeh SW, Kug JS (2017) Clim Dyn 48:3799–3811
19. Li J, Wang Z, Wu X, Xu CY, Guo S, Chen X, Zhang Z (2021) Water Resour Res 57
20. Chisanga CB, Mubanga KH, Sichigabula H, Banda K, Muchanga M, Ncube L, van Niekerk HJ, Zhao B, Mkonde AA, Rasmeni SK (2022) J Water Clim Change 13:1275–1296
21. Mosase E, Ahiablame L (2018) Water 10
22. Kalu I, Ndehedehe CE, Okwuashi O, Eyoh AE (2021) Remote Sens 13.
23. Hamududu BH, Ngoma H (2020) Environ Dev Sustain 22:2817–2838
24. SASSCAL Data and Information Portal (2023) Open data and information on climate change and adapted land management in Southern Africa. https://data.sasscal.org/. Accessed 18 Sep 2023
25. Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF (2018) Sci Data 5.
26. Copernicus (2022) Home. https://www.copernicus.eu/en. Accessed 14 May 2024
27. Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, Nicolas J, Peubey C, Radu R, Rozum I, Schepers D, Simmons A, Soci C, Dee D, Thépaut J-N
28. National Aeronautics and Space Administration (2022) NASA prediction of worldwide energy resources: data access viewer. https://power.larc.nasa.gov/beta/data-access-viewer/. Accessed 14 May 2024

29. Huang B, Thorne PW, Banzon VF, Boyer T, Chepurin G, Lawrimore JH, Menne MJ, Smith TM, Vose RS, Zhang HM (2017) J Clim 30:8179–8205
30. Lap TQ, Bang NL (2024) Water Resour 51:764–779

Evaluation Study of an Adaptive Appointment Booking System

Massimo Carlini, Giuseppina Anatriello, and Elisabetta Cicchiello

Abstract The modern business context and the amount of data available to companies and organizations has made decision-making processes even more complex and articulated. This pushes companies to provide a better product or service for customers, reasoning in terms of quality, flexibility, and responsiveness to their requests and needs. In this context, the concepts of Customer Centricity and satisfaction are placed, or the need for companies to try to satisfy demand by offering efficient and quality treatment aimed at satisfying customer needs based on a deep and solid knowledge of them. This paper reports on the activities carried out by Anas S.p.A., by Customer Service, over the last few years, to improve the Digital Customer Experience, making available to customers the knowledge and experience acquired over the years. The objective, in terms of Customer Centricity, was to put the customer at the center of the offer, providing them with more modern, innovative, intelligent and efficient dialogue tools.

Keywords Adaptive system · Intelligent system · Digital communication

1 Introduction

Anas S.p.A, a company of the Ferrovie dello Stato Italiane Group, is an efficient, innovative, transparent and internationally open industrial company that manages assets of approximately 32 thousand kilometers of Italian roads and highways.

M. Carlini (✉) · G. Anatriello · E. Cicchiello
Anas Ltd., Statistical Analysis and Reputation Management (DCOM), Rome, Italy
e-mail: massimo.carlini@stradeanas.it

G. Anatriello
e-mail: g.anatriello@stradeanas.it

E. Cicchiello
e-mail: e.cicchiello@stradeanas.it

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_12

The company's commitment is aimed at those who use road infrastructures for their travel (work, leisure, vacation, etc.) and is implemented through the design, construction, management, and maintenance of the roads that connect every location in the country.

The basis of this commitment lies in the awareness that the emergence of the digital economy, through better connectivity and access to information, has made it easier and cheaper to communicate with customers. In fact, the organizational strategy of companies has changed, moving from the centrality of the product to the centrality of the customer.

Customer Services along the road and motorway network are a strategic activity of the company's business.

2 Customer Centricity

Anas Customer experience has become an essential concept for companies and is closely linked to another important concept, Customer Centricity, which became established in the late 1990s.

The first to support the importance of customer orientation were the managers of General Electric, who in 1950 introduced the concept of Marketing Concept. They argued that it was necessary for the entire company organization to revolve around the customer: the consumer must be the fulcrum around which the entire organizational system rotates.

> "The Marketing Concept is based on the idea that the customer must be the focal point in planning the company's activities and that company resources must be organized keeping in mind the needs and requests of customers" [1].

Therefore, adopting the Customer Centricity approach involves implementing a strategy that places the customer at the center of the company's processes.

Customer Centricity is a business approach that aims to achieve the competitive advantage that comes from positive Customer Experience. It is a strategy that integrates products, services and experiences inside and outside the company to provide solutions to customer needs [2].

For this reason, it is essential that companies really know their customers, using tools that allow them to define common characteristics to specifically design the best possible experiences. Supporting this need are Customer Relationship Management (CRM) platforms, technologically advanced platforms that allow organizations and businesses to track everything that is done with their customers to manage, analyze and optimize relationships with them. This is an inter-functional strategy for managing business processes, oriented toward the customer, integrated with technology that aims to maximize relationships and includes the entire organization.

3 Anas Customer Service

To meet the needs of external users, in 2006 Anas established a centralized public relations service "Pronto Anas," capable of responding to requests and reports from external customers. The main objective is to facilitate the customer receiving correct information about the company, the activities and services offered, responding to requests and evaluating reports or complaints.

Those who contact "Pronto Anas" receive an immediate response directly from the Customer Service consultants (Contact Center). If the complexity of the request does not allow for an immediate resolution, it is forwarded to the territorially competent Public Relations Office, which provides the information within thirty calendar days of receiving the request.

Customers can contact the company through various channels, such as telephone, email, PEC, live chat, Twitter, WhatsApp, Telegram or by going directly to one of the offices located throughout the country.

Over the last few years, Customer Service has further grown in line with the innovative processes that have affected the entire company, perfecting the transformation already underway both in terms of processes and contact tools as well as the way of communicating, offering more transparent, effective and innovative services. In fact, the digital communication channels WhatsApp and Telegram were introduced at the end of 2022 to make the search for information on the road world in line with the most modern needs related to digital communication. This communication strategy has confirmed that new technologies can offer users the possibility of interacting with the company at any time, significantly simplifying the management of communications, offering the possibility of also having "asynchronous" channels with which to interact quickly and directly.

The objective of these activities carried out by the company is to improve the Digital Customer Experience of Anas, enhancing the Know-how acquired over the years, which is not lost within the company, but rather made available to the customer.

3.1 The Renewal of Customer Service

For the staff of the internal Anas Contact Center, a real phase of renewal of the Customer Service was started which led to a cultural revolution, in line with the innovation processes that affected the entire company.

The objective, in compliance with the strategic directions expressed by the top management, was to put the customer at the center of the offer, providing him with more modern, innovative and efficient dialogue tools. The Lean Six Sigma methodology was therefore introduced into the company, which led—among other things—to an efficient and effective involvement of its collaborators, a reduction in company costs, the optimization of processes and improving service levels.

The Lean Six Sigma methodology places emphasis on reducing variability and optimizing processes, improving their quality. Using the Define, Measure, Analyze, Improve, Control (DMAIC) cycle approach—a structured data-driven problem-solving process—companies can identify and manage the key factors that influence their reputation.

With the DMAIC approach we tried to enhance the performance of every single resource in the team to achieve better experience for the customer.

4 The Appointment Booking System

In this context, Anas, to streamline its project on user experience, has added an appointment booking system to the traditional contact channels.

Starting from June 3, 2020, during the Covid pandemic, in compliance with the Legislative Decrees, Anas has launched a new online appointment booking system to regulate and manage customer access to the Public Relations Offices (URP) located throughout the country. To guarantee accessibility to citizens, in record time, the company was able to transform a long-standing global problem into an opportunity for both users and the company itself.

A process was built, regulated by an internal procedure approved by the various company bodies, ensuring that citizens can go—in complete safety—to the Anas Public Relations Offices by appointment.

Below are the technical details of the designed process, an intelligent and adaptive system both with respect to the needs of the customers receiving the service and of the Anas operators who manage the process itself.

Customer area: through a dedicated page on the institutional website, the customer will be able to access the "appointments" section via a dedicated link. To log in, the customer must indicate the email address or mobile number after which they will receive a temporary token to insert in the dedicated field.

After logging in, the customer will find themselves on the booking page through which they can send a new booking request or modify or cancel an existing booking.

To send a new booking it will be necessary to enter the requested personal data, select the Anas URP structure of interest, the subject of the request, the reference road and the method through which the appointment will be managed (telephone or in person). With a view to Customer Centricity and improving the user experience, the system provides the customer with the possibility of consulting the calendar of the relevant URP structure and independently selecting both the day and the desired time from the available time slots. The customer is also provided with the possibility of adding attachments to the booking request via a corresponding button. After filling in all the fields, the customer must accept the authorization to process their personal data to proceed with sending the request. To confirm the delivery of the booking, the customer must enter a token received on their email address or on their mobile number.

Once the booking has been sent, the appointment will be sent to the Anas CRM and the customer will receive a confirmation message on their email address or on their mobile number containing the date and time of the appointment as well as the booking code.

As mentioned above, the system also allows you to independently modify or cancel the booked appointment: the customer must access the system and enter the booking code and their surname. The application will match the information with the Anas CRM and, in the event of a coincidence, will return the booking details to the customer, with the possibility of modifying the appointment based on the free slots, exactly with the same procedure as the first booking.

This is an intelligent and adaptive system that considers the days when the offices are closed, the opening hours of the offices and is directly connected to the hardware, firmware and software infrastructure of the Anas CRM platform. In fact, the system has been designed in such a way as not to imply either structural or functional changes to the company information systems, interfacing directly with the CRM.

Anas's operator area: this is an area accessible via a dedicated link reserved for employees of the Anas URP structures in charge of managing appointments at their headquarters.

After logging in, the operator will find themselves on the "Local Area Manager" home page of their office, from which they can access one of the two main sections:

- **Calendar management**: by accessing this section, the operator will find the calendar of their office. By clicking on a working date, the list with the times and reservations opens. From the perspective of an intelligent system, it is possible to define the available time slot with a specific time range based on the type of request that is the subject of the appointment requested by the customer. It will be possible to add more than one choice, among the available time slots, following the same procedure.
- **Appointment management**: by accessing this section, the operator will find the list of booked appointments, both past and future, with the search at the top right and the filter by day/time. Also on this page, you will find the edit/delete buttons. To modify an appointment, simply select it from the list and press and modify. The operator also can modify the time of an appointment or delete it. Whenever the appointment time is changed or canceled, the customer will receive a notification via their email address or mobile number.

5 Evaluation of the Appointment Booking System

This process designed for citizens has been constantly monitored since its inception, to identify customer perceptions regarding their satisfaction and to anticipate new and possible needs.

The concept of Customer Satisfaction refers to the degree of customer satisfaction with a product, service and/or company. It represents the perception or evaluation of what a company offers.

The degree of Customer Satisfaction is measured with specific statistical techniques and analyses, applied to data collected through questionnaires built ad hoc that can help companies identify their strengths and the aspects of business and services that need to be improved.

The results of Customer Satisfaction surveys are indicators of the propensity of customers to continue using a certain service in the future and/or to repurchase a certain product [3].

Customer Satisfaction can be defined as management discipline and a behavioral style that characterizes the company. In fact, it defines the manifestation of the company's ability to generate value for customers and to be able to anticipate and manage their expectations, demonstrating skills and responsibility in responding to and satisfying the expressed needs [4].

In the current economic context, it is necessary for companies to constantly monitor the level of customer satisfaction, as they are no longer uninformed individuals who buy products and/or use services blindly, but have accumulated experience and capabilities, including financial ones, that allow them to request products and services that satisfy their needs and expectations.

Starting from this assumption, having satisfied customers is fundamental for a company: a satisfied customer is the premise for having a loyal customer, who represents a very important resource. From this, follows the need to measure the value of this resource to manage it in the best possible way.

Knowledge of the level of customer satisfaction is therefore a fundamental indicator that allows the company to have a precise idea of the value of the customer resource and allows it to identify concrete actions that can lead to an improvement in the performance perceived by customers.

Measuring the Customer Satisfaction process underlines and enhances both the company's attitude toward listening to the customer and its orientation toward service quality.

Starting from 2020, the year the system was introduced, annual surveys were launched with the support of a structured questionnaire with the aim of analyzing satisfaction with the quality of the service offered.

In detail, the sample under analysis was identified through random sampling among customers who booked an appointment at Anas URPs in the reference period and who gave their consent to participate in the research, in compliance with GDPR 679/2016.

Through the support of a structured questionnaire submitted to customers who booked an appointment at Anas URPs, a survey was launched with the aim of analyzing satisfaction with the quality of the service offered.

The following indicators were identified, recorded and analyzed:

- Appointment booking procedure;
- Courtesy of the URP employee;
- Quality of the response provided to the customer;
- Usefulness of the response provided to the customer.

Starting from these dimensions, a global index of satisfaction with the appointment booking process was constructed using additive media.

5.1 The Results of the Survey

Since the launch of the new appointment booking system, over 1.000 appointments have been pre-booked at the territorial URPs located throughout the country.

The latest 2023 annual survey revealed that approximately 41% of customers who participated in the survey booked their appointment through the Anas telephone channel. For approximately 56% of those interviewed, it was very easy to book their appointment at the local URPs and approximately 87% declared that they would use the service again with the same method. One of the strengths of the analyzed process is the courtesy of the URP staff who manage the requests: the interviewees gave an average score of 8.5 on a scale of values between 1 and 10 (Table 1).

The dimensions that make up the global satisfaction index of the appointment management process were also the subject of a gap analysis, a technique that allows identifying the deviations between the expected quality and the quality perceived by customers, in relation to a service.

Customer satisfaction, in fact, depends both on the size and the direction of the discrepancy between perceived performance and the comparison standard [5].

The objective of gap analysis consists in analyzing the contingent situation and governing the process of evolution toward the desired condition.

From the comparison between the expectations and perceptions of users in relation to the service received, the existing differences (gaps) are identified, based on which it is possible to identify any actions necessary to achieve the set objectives (Fig. 1).

Expectations represent a subjective reference level, present in the consumer's mind before purchasing or using.

The gap is estimated as the difference between expectation and satisfaction; therefore, a positive difference indicates that the provision of the service has generated a level of satisfaction below market expectations. On the contrary, a negative gap indicates that satisfaction has gone beyond customer expectations.

From the gap analysis conducted on the dimensions identified for the evaluation of the appointment booking process, it emerged that the interviewees were, overall, satisfied with the service received even compared to the initial expectations (Table 2).

Table 1 Dimensions of the global satisfaction index

Index	% positive reviews (6–10) (%)	Average rating (scale 1–10)
Booking procedure	86	7.7
Courtesy of the URP employee	93.8	8.5
Quality of response	87	7.9
Usefulness of the response	80	7.4

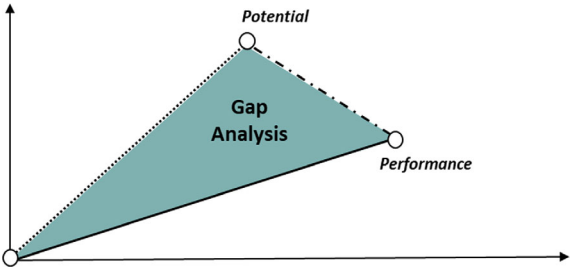


Fig. 1 Gap analysis

Table 2 La gap analysis

Index	Expectation	Satisfaction	Gap
Booking procedure	7.2	7.7	−0.5
Courtesy of the URP employee	7.5	8.5	−1.0
Quality of response	7.3	7.9	−0.6
Usefulness of the response	7	7.4	−0.4
Global appointment management process index	**7.4**	**7.8**	**−0.4**

Anas' commitment to citizens is to maintain, at least, the level of satisfaction expressed, which in the 2023 annual survey was equal to 7.8/10.

These results, together with the general level of satisfaction recorded over the years, have confirmed the efficiency of the current booking system and process, for which, therefore, it was not deemed necessary to make substantial changes in the short term. In fact, already from the first annual survey conducted in 2020, an overall satisfaction of 7.7/10 was recorded, a result substantially confirmed in the 2023 annual survey, with an increase of 1.3%.

In addition to gap analysis, quadrant analysis was also carried out, a technique whose objective is to represent, through a graph, the positioning of a company/service, as perceived by customers, within a quadrant composed of areas characterized by a different priority of intervention. It is a useful tool in orientation and action planning.

To place every aspect of the Anas appointment booking system within the matrix it was necessary to define the performance values, understood as user satisfaction, and relative importance.

The quantification of the importance of each single aspect in the user's perception was calculated using the Partial Least Square–Path Modeling (PLS-PM) structural equation statistical model which allowed the "latent importance" to be estimated, therefore not declared, but inferred from the data structure.

PLS-PM uses structural equation models (SEMs) that are used to test hypotheses and measure perceptions of impacts. SEM is a statistical modeling technique frequently used in the behavioral sciences. It can be seen as a combination of factor analysis and multiple regression or as a combination of factor analysis and path analysis.

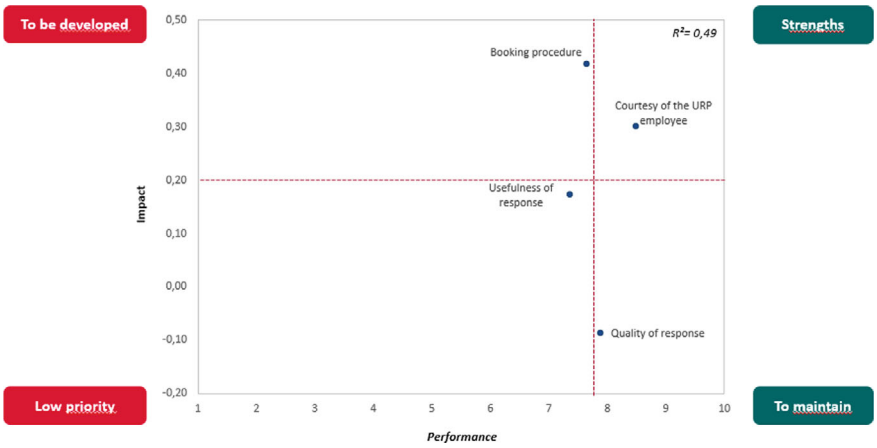


Fig. 2 Quadrant analysis

In the social sciences it is common to study perceptions that are not directly measurable. They are referred to as "constructs" or "latent variables." SEMs are models that can model complex structures of causal relationships between latent variables starting from a set of real variables, called "manifest" [6].

As highlighted previously, the analysis revealed that the strong point of the new system put into the service of customers is the courtesy of the URP employee who manages the appointments; while, in the quadrant called "to be developed" we find the "booking procedure" dimension (Fig. 2), made up of two indicators: the booking process and the appointment management method. This index is placed in this area as, despite recording a positive value of 7.7/10, it was the one with the greatest impact, or importance, on the overall satisfaction with the appointment booking system.

This means that to further increase overall customer satisfaction with the appointment booking process, we need to work on the activities inherent to the booking procedure.

6 Conclusions

In line with the concepts of Customer Centricity and Customer Experience, Anas Customer Service, over the last few years, has implemented a transformation of the contact processes and tools to offer users more transparent, effective, immediate and innovative services also from a communication point of view.

These processes are constantly monitored by the company with the aim of identifying concrete actions that can lead to an improvement in the performance perceived by customers.

This monitoring also occurs through Customer Satisfaction surveys, which allow:

- to know the opinions of customers;
- to understand the needs, requirements and expectations of the customer;
- to identify the actions and methods to overcome the deviations between the perceived quality and the quality delivered;
- to establish performance standards;
- to understand in which direction to orient future choices.

The appointment booking system, the subject of this paper, was built, implemented and subsequently measured to make it increasingly adaptable to the needs of users.

The results of the evaluation study conducted showed a general picture of satisfaction, with an overall satisfaction of 7.8/10, confirming the efficiency and adaptability of the appointment booking system and process. In fact, approximately 87% stated that they would use the service again in the same way.

In line with customer satisfaction, the system designed by Anas is intelligent and adaptive for the following characteristics:

- it considers the closing days and opening hours of the offices;
- it is directly connected to the hardware, firmware and software infrastructure of the Anas CRM platform, in such a way as not to imply structural and functional changes to the company information systems;
- it allows customers to book and manage their appointments independently;
- it allows the operator, through calendar access to their workplace, to define the available time slot with a specific time range defined based on the type of request that is the subject of the appointment.

This last feature fully expresses the intelligence of the designed system, as it allows an optimization of the management of the resources that work at the territorial URPs of Anas and of the activities they carry out. In fact, by scheduling the day, the time slot and the time range useful for managing an appointment, it was possible to better manage and schedule the activities of each URP resource.

References

1. Myers JH (1999) Measuring customer satisfaction: hot button and other measurement iusses, USA
2. Lamberti L (2013) Customer centricity: the construct and the operational an-tecedents. J Strateg Market 21. https://doi.org/10.1080/0965254X.2013.817476
3. Farris PW, Bendle N, Pfeifer PE, Reibstein D (2010) Marketing metrics: the definitive guide to measuring marketing performance. Pearson Education
4. Valdani E (1995) Marketing strategico, un'impresa produttiva per svi-luppare capacità market driving e valore, RCS libri & Grandi Opere, Milano
5. Guido G, Bassi F, Peluso A (2010) La soddisfazione del consumatore: la misura della Customer Satisfaction nelle esperienze di consumo, FrancoAngeli, Milano
6. Bollen KA (1989) Structural equations with latent variables. John Wiley and Sons Inc., New York

Continuous Learning System for Detecting Anomalies in Daily Routines Using an Autoencoder

Dominic Gibietz, Daniel Helmer, Eicke Godehardt, Heiko Hinkelmann, and Thomas Hollstein

Abstract The ongoing demographic change towards an ageing population increases the need for effective solutions to support independent living and ensure the safety of elderly people living alone. Detecting anomalies in the daily routines of these people is a critical task in order to address these challenges and maintain their well-being. This paper proposes an unobtrusive method for anomaly detection using binary sensor data and machine learning. The approach involves a neural network in form of an autoencoder, which evaluates hourly data of each room, including the accumulated residence time, the activity time and the number of room entries. The system learns individual normal behaviour through online learning and detects deviations from it. Testing and evaluation of the system was carried out using a publicly available dataset and comparing different configurations for the model. A comparison was also made between the use of individual maximum values for each room to normalize the data and uniform values for all rooms, with the former performing significantly better. The results demonstrate that the system can effectively identify the majority of unusual daily routines with a high accuracy, offering potential for improving safety measures for people living alone.

Keywords Anomaly detection · Daily routines · Ambient assisted living · Autoencoder · Machine learning · Time series data

D. Gibietz (✉) · D. Helmer · E. Godehardt · H. Hinkelmann · T. Hollstein
Frankfurt University of Applied Sciences, Frankfurt, Germany
e-mail: dominic.gibietz@fra-uas.de

D. Helmer
e-mail: daniel.helmer@fra-uas.de

E. Godehardt
e-mail: godehardt@fra-uas.de

H. Hinkelmann
e-mail: hinkelmann@fra-uas.de

T. Hollstein
e-mail: hollstein@fra-uas.de; thomas@ati.ttu.ee

T. Hollstein
Tallinn University of Technology, Tallinn, Estonia

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_13

1 Introduction

The demographic change brings new challenges for the healthcare system and social care in many countries. Elderly people who live alone are at an increased risk of suffering health emergencies or experiencing a deterioration in their physical or mental condition. For example, falls, acute illnesses or a gradual deterioration in health can often only be detected when medical intervention is required. In such cases, the opportunities for early support or preventive measures are not used, which not only affects the quality of life of the resident, but also has an impact on the healthcare system.

Another factor associated with demographic change is the increasing shortage of skilled workers in the care and health sector. As the number of elderly people increases, so does the number of people in need of care and therefore the need for skilled workers. According to current calculations, the number of nursing staff shortages in Germany is expected to rise to 690,000 by 2049 if employment in the nursing professions does not increase further [1]. Not only are caregivers and medical staff under greater demands [2], they are also increasingly overloaded, which limits the possibilities for regular and preventative health monitoring and care. The shortage of skilled workers means that home care and support services do not have sufficient capacity and therefore preventative measures are often neglected.

Against this trend, ambient assisted living (AAL) technologies have become increasingly important in recent years. By combining sensor technology, data analysis and artificial intelligence (AI), these technologies offer the potential to detect deviations at an early stage and identify potential health risks through continuous monitoring and analysis of behaviour patterns. Individual behaviour patterns and habits require a high degree of sensitivity and adaptability of AAL systems in order to enable meaningful anomaly detection. Such technologies could not only help to relieve the demands on skilled workers and family members through automation, but also give individuals living alone a certainty of being safe.

In this paper, our first results of a neural network system are described that learns the daily routines and normal behaviour of a person living in an unsupervised manner and detects unusual deviations from what has been learned. Unusual behaviour, which otherwise goes unnoticed and may be the first sign of a deterioration in health, should be identified at an early stage. Examples of unusual behaviour include signs of the restless legs syndrome [3, 4], sundown syndrome [5] or the negative impact of new medication. By identifying potential health problems at an early stage, they can be addressed in time, allowing residents to live safely within their own four walls. At the same time, the aim is for the system to be able to be integrated into any previously unknown living environment without time-demanding pre-configuration and to work completely locally for data protection purposes.

This paper is organized as follows. The following section reviews previous studies related to this research. Section 3 outlines the proposed method for anomaly detection. In Sect. 4, the data format and preprocessing steps are described. Section 5 details the setup and design of the system. Section 6 presents the performed tests, their

configurations and key results. In Sect. 7, limitations and potential improvements are discussed. Finally, the conclusion and future work are summarized in Sect. 8.

2 Related Works

The detection of anomalies in AAL environments, especially for elderly people living alone at home, is an active field of research with significant social relevance due to its potential to enhance quality of life and ensure safety alike. Various approaches have been developed to reliably identify changes in people's routines or behaviour. One such approach is the use of statistical methods.

Susnea et al. [6] propose a statistically based solution that uses binary sensor data to create activity maps that show both the spatial and temporal distribution of activities. Deviations from normal behaviour are detected by comparison with a manually defined reference interval. This manual selection of the interval prevents automated adaptation to dynamic or minor changes in behaviour. De Paola et al. [7] developed a context-aware AAL system that uses dynamic Bayesian networks to fuse sensor data and detects anomalies in user behaviour with a rule-based decision maker. The system evaluates activities, sends alerts to caregivers when necessary, and adapts the environment using actuators. The results show that anomalies in predefined activities are reliably detected, while detection of anomalies in non-predefined ("other") activities is limited. A continuous health assessment system by Merten et al. [8] compares the movement patterns of the last 24 h with a reference day, which is built up of at least seven days of normal behaviour. Anomaly detection is performed by calculating the Hamming distance between the current day and the reference day. PIR sensors are used that must not overlap during installation to ensure that movement is correctly attributed. The anomaly detection threshold is set statically and does not dynamically adapt to minor or long-term changes in behaviour. In addition, the reference tag is not automatically updated, which affects the accuracy of the system in the event of long-term changes. Chifu et al. [9] propose a system that uses a smartwatch to detect a person's location and activities. The smartwatch transmits Bluetooth data to beacons distributed throughout the apartment. Deviations from normal daily routines are detected by using a Markov model and analysing entropy rates. The authors point out concerns regarding data security, as the beacons are vulnerable to abuse if no protective measures such as time-variable IDs are used. The activities that can be detected are predefined and based on fixed rules. Yao et al. [10] present a non-invasive system for activity recognition based on RFID tags. The system uses compressed representations of received signal strength (RSSI) in combination with a dictionary-based approach to classify activities. Their study also compares the proposed method with other statistical classifiers, including Support Vector Machine (SVM) and Random Forest. The system achieves 95 % accuracy in person-specific validation, where the model is trained and tested on the same person's data. However, with person-independent validation, the accuracy drops to around 70 %, which illustrates that individual behaviours and routines vary greatly.

Another approach is the use of clustering methods to identify anomalies on the basis of outliers in grouped data. For this, [11] propose a low-cost system for monitoring the movement of elderly people in smart homes, based on RFID sensors and a NodeMCU microcontroller. The RSSI data is used to detect movement patterns and analysed using K-means clustering to identify anomalous behaviour. The system requires an active RFID tag to be worn, as motion data can't be recorded without it. Shahid et al. [12] also use K-means clustering to analyse movement data collected over several weeks or months, which can be used to detect long-term changes in behaviour, such as reduced activity or increased time spent in certain areas. A limitation of this approach is that the data needs to be aggregated over a longer period of time before anomalies can be identified, making it impossible to take action if problems occur earlier. Zekri et al. [13] developed a framework for long-term behavioural analysis that combines DBSCAN clustering with a fuzzy logic-based decision module. Sensor data is used to model behavioural patterns and detect deviations such as longer stays in rooms or irregular activities. To do this, data must be collected over several weeks to create a reference of normal behaviour. However, this static reference does not dynamically adapt to changes, which limits the flexibility of the system.

In addition, neural network-based machine learning and deep learning models are increasingly being employed for anomaly detection in smart homes. These models are particularly effective as they are able to analyse high-dimensional data and identify complex patterns that are challenging to detect using conventional methods. Zhang et al. [14] use network cameras for data acquisition and employ Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) for analysis. By combining CNNs for skeleton extraction with LSTMs for temporal sequence analysis, their system detects falls or other unusual behaviour patterns with an accuracy of over 85 %. Although the processing is performed locally using powerful but cost-intensive hardware, the use of cameras can be considered intrusive and thus may limit user acceptance. Kim et al. [15] developed a system to detect unusual behaviour in dementia patients. The approach uses autoencoders to detect anomalies and subsequently applies an LSTM model to classify these anomalies. The system only distinguishes between insomnia and repetitive behaviour. At least 30 d of data is required to train the system with the system relying on a pre-stored weekly schedule against which the person's activities are compared. Gonzalez et al. [16] present a system based on the analysis of household appliances power consumption. Autoencoders and Variational Autoencoders (VAE) are compared in order to learn typical behaviour patterns and detect deviations. A separate neural network is trained for each appliance, which increases accuracy but also increases the complexity of the system. In addition, the models remain static after training and do not adapt any further to the user.

The above-described research works provide various ways in which data can be collected and then analysed to detect anomalies and deviations from normal behaviour. They are based on deterministic or statistical models, as well as on neural network learning approaches, which are unsupervised or, in some cases, supervised. Since life habits can be very different, in fact a complex supervised initialization

phase is needed for the majority of these systems, which is a real obstacle for practical application, since health care staff can typically not support this process.

Therefore, this paper presents a novel approach which is based on unsupervised learning and that detects deviations from the normal behaviour of a person living alone by evaluating their daily routine. The aim is to detect changes and unusual behaviour in daily data such as bedriddenness or night-time activity, which may indicate illness or similar, at an early stage so that preventive measures can be taken by caregivers or family members. The proposed system employs machine learning in the form of an autoencoder. This allows the system to independently learn the person's individual normal behaviour without any complex pre-configuration. The characteristic of the autoencoder makes it possible to compress the input data and then reconstruct it in order to identify anomalies. While in some research projects, as described above, no further adaptation is made once the system has been trained, the system described here is designed to ensure that small variabilities in a person's routines do not lead to false alarms through daily training. Data from binary sensors, which can be integrated discretely, unobtrusively and easily into homes, is used for this purpose. With the data, it is possible to generate accumulated residence and activity times as well as the number of room entries for every hour, which can be evaluated by the neural network. At the same time, the system takes into account the individual maximum normal values for each room in the apartment. This preprocessing of the data also has the advantage that the number of binary sensors is scalable at a later stage. For example, the system can be expanded to include bed sensors in order to document and evaluate restlessness during the night.

3 Proposed Approach

This paper proposes a system for anomaly detection in ambient assisted living (AAL) environments using an autoencoder. Autoencoders have various applications, including anomaly detection [e.g. 15–17], as in this study, which is about detecting deviations from learned normal behaviour patterns. These neural networks are used for tasks such as dimensionality reduction and data reconstruction, making them useful for detecting irregularities in data caused by reconstruction errors. An autoencoder consists of two primary components: the encoder, which compresses the input data into a compact internal representation, and the decoder, which reconstructs the input data from this internal representation [17]. During training, the reconstruction error between input and output, which is calculated using loss functions such as the mean square error (MSE), is continuously minimized. As the autoencoder learns to reconstruct normal data patterns, it becomes sensitive to anomalous inputs, as these will result in higher levels of reconstruction error. There are various autoencoder architectures, each of which fulfils specific tasks, such as the denoising autoencoder for error correction or the VAE for image generation, among other things. Each variant is tailored to specific applications and data features, making autoencoder a versatile tool.

In the context of this study, a simple autoencoder architecture with a single hidden layer is applied. The compression and reconstruction of the data described above enables the detection of anomalies due to increased reconstruction errors. Also, through continuous learning, the model adapts to the individual normal behaviour of the resident so that it can be used in previously unknown environments and adapts to different living conditions. This approach uses the autoencoder's ability to model normal data patterns and is therefore suitable for detecting anomalies where deviations from learned patterns may indicate irregular or concerning behaviour.

To learn the resident's daily routine the proposed system analyses binary data. For this study, the input features include accumulated residence time, activity time and the number of room entries for each hour, taking into account different maximum normal values for each room, as further detailed in the following sections. By continuously analysing these data points, the system can detect unusual deviations, which can serve as early indicators of potential health problems, the negative effects of new medications or safety risks. Early detection of such irregularities enables timely intervention, which can minimize health risks and improve the general well-being of the resident.

Binary sensors can be integrated unobtrusively into the home without compromising the resident's privacy. At the same time, different variants of these sensors, such as bed sensors, can be easily added to obtain detailed data on activity times or night-time restlessness, making the system scalable. Furthermore, to ensure privacy and data protection, all data processing and analysis takes place locally on the system.

This combination of adaptability, scalability and focus on data protection makes the system a robust tool for continuous monitoring in assisted living environments and a valuable new approach.

4 Data Acquisition and Processing

The present study utilized public smart home datasets from the Centre for Advanced Studies in Adaptive Systems (CASAS) at Washington State University [18]. These datasets consist of a list of sensor measurements with time stamps from motion sensors, temperature sensors and door sensors, although only motion sensor data were employed in this study. The data also record various activities of the resident, such as watching TV, sleeping, cooking and guest visits. When selecting the datasets, particular attention was paid to ensure that they represented a single-person household and that the proportion of data involving visitors was not excessive.

For data preprocessing, the rooms of the apartment were extracted from the sensor data and the corresponding motion sensors were assigned to each room. The data differentiates between two types of motion sensors. "MotionArea" sensors cover a larger area or even entire rooms, whereas the field of view of "Motion" sensors is limited to a specific area, such as a bed or a sofa. The motion sensors were assigned to the respective rooms or areas of the apartment. Sensors that could not be clearly assigned to a room were excluded, for instance, if their orientation covered transi-

tional areas such as hallways. The number of sensors per room may therefore vary. Based on the selected sensors, the list of sensor data was filtered and the active time of the binary sensors was calculated using the timestamps. This way, a simplified activity level of the person can be extracted.

With the filtered data, the information for each room can be reconstructed by hour throughout the day, allowing the creation of a daily schedule. The key features used for this are the length of time spent in a room, the duration of activity during the stay, and the number of room entries. To be recognized as a valid room stay or entry, specific criteria must be met. For example, sensors in the entered room must remain active for a minimum duration to filter out false activations, while still capturing very short stays, such as unusual back-and-forth movements between rooms. The number of room entries results from the number of confirmed entries in a room in the respective hour. Tables 1 and 2 illustrate how individual sensor activations in the data were converted to reconstruct a daily schedule, with data separated at full hours. For each stay, the corresponding residence time (RT) and activity time (AT), both in seconds, are added, which are used in the next step as input data of the neural network. A graph showing the residence times of an example day is shown in Fig. 1.

Currently, the presence of additional people is not considered. This can lead to an increased number of sensor activations, resulting in a higher recorded activity time and a greater number of detected room changes. Furthermore, leaving the apartment is registered as a stay at the entrance door (referred to as "OutsideDoor" in the data), and no activity time is recorded outside of the apartment.

Table 1 Sample section of the raw data, with milliseconds removed

Date	Time	Sensor	Room	Status
26.07.2012	10:59:09	M008	LivingRoom	ON
26.07.2012	10:59:10	M008	LivingRoom	OFF
26.07.2012	11:00:46	M008	LivingRoom	ON
26.07.2012	11:00:52	M008	LivingRoom	OFF

Table 2 Sample section of a reconstructed routine, with data separated at full hours

Date	Time	Room	RT	AT
26.07.2012	10:24:06	Bedroom	178	80
26.07.2012	10:27:04	LivingRoom	1976	62
26.07.2012	11:00:00	LivingRoom	2084	96
26.07.2012	11:34:44	Bathroom	172	39

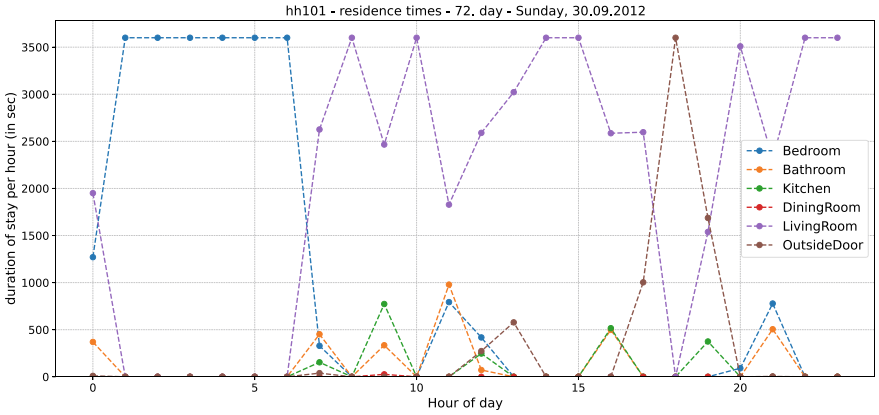


Fig. 1 Example visualization of hourly residence times for a single day

5 System Structure and Implementation

As mentioned before, this study employs an autoencoder as the neural network model. A simple autoencoder with a single hidden layer is used, utilizing Leaky-ReLU as the activation function in the encoder and sigmoid in the decoder. The mean squared error (MSE) serves as the loss function, with the Adam optimizer applied for optimization. The network is not pre-trained and is intended to adapt to normal behaviour of a person during the first few days of operation and then continues to update incrementally over time through online learning.

The architecture of the autoencoder depends on the floor plan of the apartment or the distribution of the sensors, as the number of rooms determines the number of input neurons in the network. For each room and each hour, the three features mentioned before are extracted from the data: residence time, activity time and number of room entries. In an apartment with six rooms (or defined areas), this results in 432 neurons in the input layer of the autoencoder. Before the system is trained for the first time, the model is automatically built using these information. The data presented in Table 2 is converted into a pure numeric format and stored as a sequence of values, representing each room and each hour as a consistent set of values, as shown in Table 3. The table illustrates the first hour (0:00 to 1:00) with 18 data points for the six rooms (3 features per room), starting with the data for the bedroom. The order of the rooms and the features is consistent throughout each hour and each day, maintaining consistency in the input structure.

Since the values of the features differ significantly in both scale and units, data normalization is essential before feeding it into the neural network. Normalization aims to bring all feature values into a comparable range, thereby improving the stability and efficiency of the training process. Therefore, all values are scaled to a consistent range between 0 and 1, which not only simplifies the learning task for the model but also reduces the risk of numerical instability due to large discrepancies

Table 3 Room data for the first hour of the day

Day	Data
1	662, 56, 1, 229, 42, 1, 0, 0, 0, 0, 0, 0, 2703, 17, 1, 6, 5, 1, ...
2	3600, 47, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
3	2991, 59, 1, 609, 108, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...

in value range of the features. The maximum values required for normalization are extracted from the dataset and updated after each day, provided the day has not been declared as anomaly. Each room in the apartment is assigned individual maximum values for the three key features: residence time, activity time and number of room entries. This room-specific normalization is crucial to accurately capture differences in usage patterns between rooms, allowing the system to detect, for example, unusually long stays or increased activity in a specific room that might indicate a deviation from typical behaviour. In cases where a feature value exceeds its corresponding maximum, the normalized value will also exceed the upper limit of 1. This results in an increased error during reconstruction by the autoencoder, as the Sigmoid activation function in the output layer restricts the values to a maximum of 1. Consequently, values beyond this range indicate potential anomalies.

Since the system is not pre-trained, a "warm-up" phase was implemented to allow the model to adapt gradually to the resident's typical behaviour patterns. During this phase, the model is trained without anomaly detection. This training period also enables the extraction of initial maximum values for normalization. To facilitate a faster and more robust adaptation, the days within the warm-up phase are trained with an increased number of epochs. This approach accelerates the network's convergence towards a stable representation of normal behaviour. It is worth noting that during the training phase no anomaly detection is performed, meaning that any anomalous days in this period may also be included in training. Through continuous training, these anomalies are gradually minimized and their influence mitigated over time, as the model adapts to the data.

After completing the warm-up phase, the subsequent days are analysed for anomalies or deviations from the learned behaviour. The autoencoder compresses the input data during the encoding process, extracting the essential information while discarding less critical details. During decoding, the neural network attempts to reconstruct the original input data from the compressed representation. By comparing the reconstructed data to the original input, deviations can be detected using the MSE. If the average reconstruction error for a given day is too high and exceeds a certain threshold, the system flags the day as anomaly. At the time of this study, the threshold was calculated based on the average MSE of the preceding days where no anomalies were detected, augmented by the average standard deviation of those same days. This

threshold is designed to account for typical variations in behaviour while remaining sensitive to significant deviations. A day that does not exceed the threshold and is therefore not classified as anomalous is then used for training through online learning. This allows the system to continuously refine its understanding of normal behaviour by learning from the most recent representative data. The ability to train daily ensures that the neural network remains up-to-date, adapting to the slight variations in behaviour that naturally occur over time. This continuous enhancement not only improves the system's ability to model typical patterns, but also results in a steady lowering of the threshold. As the threshold decreases, the system becomes more sensitive to less significant deviations from normal behaviour, which can then be detected as potential anomalies. This progressive improvement makes the system even more effective in detecting changes that might otherwise go unnoticed in earlier phases.

Additionally, as described earlier, the maximum values used for normalization are updated only if new peaks are observed in the data for a day that is not classified as anomalous. This process ensures that the system remains adaptive to gradual, minor changes in normal behaviour patterns over time, while disregarding significant deviations.

6 Experimental Results

Several tests were done using the approach described in Sect. 5. The tests were based on the CASAS dataset HH101 [18], which is based on the data of an apartment with six rooms or areas of a person living alone. By extracting three features from the data of each room at each hour, this results in a total input dimension of the autoencoder of 432 data points per day. The first 150 d of the dataset were used for the experiments. To validate the results of the system, 140 of these days were previously classified as either abnormal or normal using mean values, the standard deviation and a manual review. No classification is required for the first ten days as the system does not check for anomalies during the initial training phase anyway. Out of the 140 d, 25 d were classified by us as anomalies or strong deviations.

To investigate the effects of the capacity of the hidden layer (HL) on anomaly detection and to find a balance between underfitting and overfitting, the number of neurons in the HL was tested with 42, 128 and 256 neurons. This corresponds to approximately $\frac{1}{10}$, $\frac{1}{4}$ and $\frac{1}{2}$ of the input data. During the training phase, in which the system only learns and does not detect any anomalies yet, the data was learned over ten epochs to allow a fast adaptation to the individual routines of the person. After this initial training phase, each new day was trained with only one epoch. The number of training days was varied slightly depending on the number of neurons in the HL. A higher number of neurons enables faster adaptation to new or unknown patterns, which means that fewer training days are required and overfitting is avoided at the same time. Specifically, the system was trained for 14 d with 42 neurons, 12 d with 128 neurons and only for 10 d with 256 neurons in the HL.

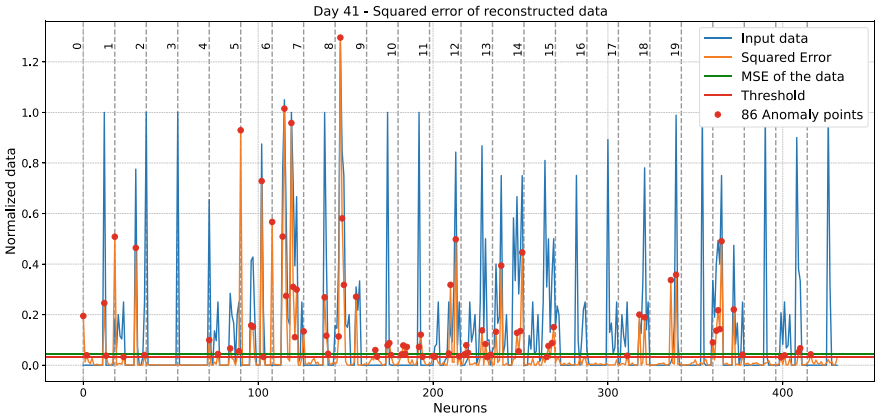


Fig. 2 Visualization of the quadratic reconstruction errors from an anomalous day

In addition, three different initial values for the normalization (maximum values) of the three features were tested for each of the HL sizes. As described in Sect. 5, the system automatically adjusts these maximum values if new maximum values occur in the data on a given day and there is no anomaly. Initially, the normalization values were set to 0, so that the system had to extract the maximum values completely by itself right from the start. The second approach used small initial values: The maximum values were set to 600 s for the residence time, 100 s for the activity time and 0 for the number of room entries. The idea behind these values was to avoid anomalies that might occur if the person spends little or no time in a room during the initial training phase. In the third approach, the starting values were set as high as possible: For each room the maximum residence time was set to 3600 s, the activity time to 1500 s and the number of room entries to 10. The 1500 s for the activity time was chosen as this value was not exceeded in the data.

During the test runs, corresponding graphical visualizations are created for each day to illustrate the deviations in the data. Figure 2 shows the squared error of the data reconstructed by the autoencoder for one day. In the example shown, the person was only in the bedroom from about 1:00 a.m. to about 3:40 a.m. that night. There are also larger errors in the following hours, as the system continues to assume that the person is in the bedroom, which would be the learned behaviour, but is actually in the living room. An unusually long stay in the bathroom around 8:00 a.m. results in a value that exceeds the current maximum value of the bathroom and therefore reaches a normalized value greater than 1, as described in Sect. 5. Figure 3, on the other hand, shows the squared error of the day previously shown in Fig. 1, which largely corresponds to the normal behaviour of the person learned by the system and was therefore not classified as an anomaly. The observable errors here are due to normal variations in the person's daily routine. Both visualizations of the quadratic error were obtained from the run with 128 neurons in the HL and the initial maximum values of 600, 100 and 0.

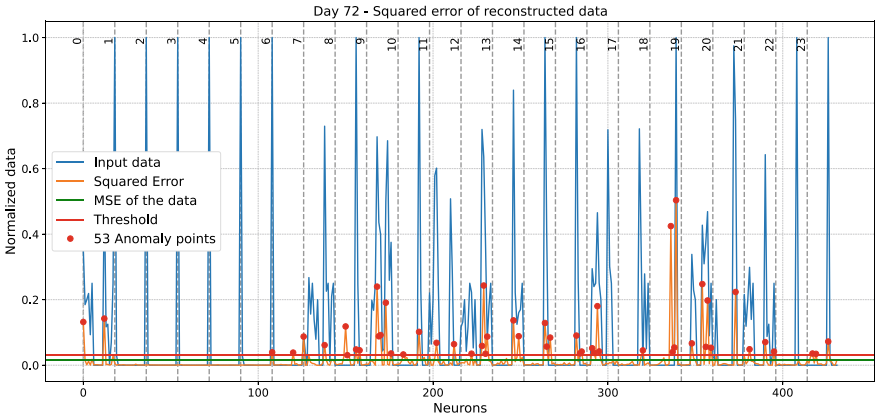


Fig. 3 Visualization of the quadratic reconstruction errors from a non-anomalous day

To evaluate the performance of the system, the metrics accuracy, recall, precision and F1 score were used. Accuracy represents the overall correctness of the model by indicating the proportion of correctly classified days relative to all days. Recall measures the ability of the system to correctly identify anomalies among all actual anomalies. Precision indicates the proportion of correctly identified anomalies among all days classified as anomalies. The F1 score provides an overall evaluation of the model by calculating the harmonic mean of precision and recall.

Table 4 shows the results of the tests described above in the form of confusion matrices. These values can be used to calculate the mentioned metrics for evaluating the performance of the system and are shown in Table 5.

Table 4 Confusion matrix of the results

	Initial max values **0, 0, 0**			Initial max values **600, 100, 0**			Initial max values **3600, 1500, 10**		
		p	n		p	n		p	n
42 HL Neurons	p'	**22**	**8**	p'	**23**	**8**	p'	**10**	**2**
	n'	**2**	**104**	n'	**1**	**104**	n'	**14**	**110**
		p	n		p	n		p	n
128 HL Neurons	p'	**23**	**8**	p'	**24**	**10**	p'	**17**	**7**
	n'	**1**	**106**	n'	**0**	**104**	n'	**7**	**107**
		p	n		p	n		p	n
256 HL Neurons	p'	**21**	**7**	p'	**21**	**6**	p'	**20**	**10**
	n'	**4**	**108**	n'	**4**	**109**	n'	**5**	**105**

Table 5 Evaluation of the autoencoder results with the best values in bold

Neurons	Initial max values	Accuracy (%)	Recall (%)	Precision (%)	F1 score (%)
42	0, 0, 0	92.65	91.67	73.33	81.48
42	600, 100, 0	**93.38**	**95.83**	74.19	**83.64**
42	3600, 1500, 10	88.24	41.67	**83.33**	55.56
128	0, 0, 0	**93.48**	95.83	**74.19**	**83.64**
128	600, 100, 0	91.30	**100.00**	66.67	80.00
128	3600, 1500, 10	89.86	70.83	70.83	70.83
256	0, 0, 0	92.14	**84.00**	75.00	79.25
256	600, 100, 0	**92.86**	**84.00**	**77.78**	**80.77**
256	3600, 1500, 10	89.29	80.00	66.67	72.73

The results show that with 128 neurons in the HL, the variant without specified maximum values for normalization (0, 0, 0) achieves the best results. In comparison, with 42 and 256 neurons in the HL, the variants with small initial values for each room perform best (600, 100, 0). Due to the longer training phase of 12 and 14 d for 128 and 42 neurons respectively, a maximum of 24 anomalies are possible, as a potential anomaly falls within the training phase. On the other hand, the variant in which all rooms are initialized with the highest possible maximum values (3600, 1500, 10) and in which the system does not adjust them during runtime, consistently delivers the worst results. This indicates that the use of room-specific maximum values to normalize the data makes more sense than the use of uniform maximum values for all rooms. The precision value of 83.33 % for 42 neurons is due to the fact that only a few anomalies were detected in this test overall.

Comparing the different numbers of neurons in the HL, 42 and 128 neurons perform slightly better than 256 neurons with an F1 score of 83.64 %. In terms of accuracy, the test results are almost equal, while 256 neurons deliver significantly worse results for recall. The advantage of 256 neurons is that a shorter training phase is required to achieve comparable results. However, fewer neurons in the HL allow a more targeted feature extraction by the autoencoder during data compression, which makes the model more robust against strongly varying data.

Overall, most anomalies are correctly detected in the variants with no or small initialization of the maximum values. The number of false positives should be further reduced in future work.

7 Discussion

The results presented in Sect. 6 demonstrate that anomaly detection is feasible with the selected data preprocessing and the extraction of hourly and room-specific features—in the form of accumulated residence time, activity time and room entries. In the tests, the majority of anomalies were successfully detected. However, the

system also incorrectly classified some days that were not defined as anomalies, as such. It should be noted that neural networks identify complex patterns in the data that may not have been noticed when the data was analysed manually. As a result, it is possible that the system correctly classified days as anomalies that we had not previously considered as such. In this context, a more detailed analysis of the results is required.

In general, the system has potential for improvement in certain areas to achieve more accurate results. Currently, the system evaluates the data of an entire day. Gradual changes over longer periods of time cannot currently be identified this way. While it is intended that minor adjustments in daily routines are learned by the system, these can also indicate changes due to illness or age. Addressing this limitation would require the system to compare current data with historical data.

Additionally, the system could benefit from the integration of rule-based methods. For example, anomaly detection could be paused when more than one person is detected or by appointments specified in the system in order to avoid false alarms.

The length of the training phase currently depends on the selected number of neurons in the hidden layer. However, the complexity and variability of the data is also of significant relevance here. Therefore, the system should be optimized to dynamically adapt the training phase to the given conditions in order to enable implementation in different and unknown environments.

Future work will focus on further developing the system with the points mentioned and testing it with additional datasets to ensure improved performance.

8 Conclusion

This research describes an unobtrusive self-learning method for detecting anomalies in the daily routines of people living alone. For this purpose, the hourly data of each room in the apartment is used in form of the accumulated residence time, the activity time and the number of room entries. The binary sensor data from a publicly available dataset served as the basis for testing and validation. The data preprocessing described in this work enables the system to be subsequently expanded with additional binary sensors to improve data collection, thereby ensuring the scalability of the system. To evaluate the daily data, a neural network in form of an autoencoder was used, which is suitable for anomaly detection due to its ability to find deviations in the compressed and later reconstructed data. The model continuously learns even after an initial training phase in order to adapt to normal variability and smaller changes in daily routines.

Several tests were performed in which the number of neurons in the hidden layer of the autoencoder was varied in order to determine the optimal configuration for balancing underfitting and overfitting. In addition, different initial maximum values for the normalization of the room data were tested. The results indicate that the system performs better when no or only small initial maximum values are set at the start of the training and allowing the system to independently extract these values from the

data for each room individually. With an accuracy of 93.48 % and an F1 score of up to 83.64 %, the system was able to reliably detect most days that deviated noticeably from the learned normal behaviour. However, there were also false alarms, which future work will aim to reduce.

The integration of rule-based methods could further improve the system and enhance its robustness against false alarms. For example, the presence of other people who might distort the data could be recognized, or upcoming appointments could be stored in the system to prevent false positives. Furthermore, regular events that occur at varying times, such as visits by a care service, should also be identified. Currently, the system analyses only data from the current day, which limits its ability to detect gradual changes over a longer period of time that may indicate health issues or age-related problems. Future work should extend the system to compare and evaluate current data with historical data of the daily routines of the person. This would enable caregivers to take any necessary preventive measures at an early stage to maintain the persons health.

References

1. Eppers N (2024) The nursing and care labour market and demographic change-methodology and results of nursing and care staff projection. WISTA-Sci J (Statistisches Bundesamt) 2(2024):44–54
2. Plöthner M, Schmidt K, De Jong L, Zeidler J, Damm K (2019) Needs and preferences of informal caregivers regarding outpatient care for the elderly: a systematic literature review. BMC Geriatr 19(82). https://doi.org/10.1186/s12877-019-1068-4
3. Fritz RL, Cook D (2017) Identifying varying health states in smart home sensor data: an expert-guided approach. In: World multi-conference of systemics, cybernetics and informatics: WMSCI
4. Kim KY, Kim EH, Lee M, Ha J, Jung I, Kim E (2023) Restless leg syndrome and risk of all-cause dementia: a nationwide retrospective cohort study. Alzheimer's Res Therapy 15(1):46. https://doi.org/10.1186/s13195-023-01191-z
5. Menegardo CS, Friggi FA, Scardini JB, Rossi TS, Vieira TDS, Tieppo A, Morelato RL (2019) Sundown syndrome in patients with Alzheimer's disease dementia. Dementia Neuropsychol 13:469–474. https://doi.org/10.1590/1980-57642018dn13-040015
6. Susnea I, Pecheanu E, Sandu C, Cocu A (2021) A scalable solution to detect behavior changes of elderly people living alone. Appl Sci 12(1):235. https://doi.org/10.3390/app12010235
7. De Paola A, Ferraro P, Gaglio S, Re GL, Morana M, Ortolani M, Peri D (2017) An ambient intelligence system for assisted living. In: 2017 AEIT international annual conference, pp 1–6. https://doi.org/10.23919/AEIT.2017.8240559
8. Mertens M, Debard G, Davis J, Devriendt E, Milisen K, Tournoy J, Croonenborghs T, Vanrumste B (2021) Motion sensor-based detection of outlier days supporting continuous health assessment for single older adults. Sensors 21(18):6080. https://doi.org/10.3390/s21186080
9. Chifu VR, Pop CB, Demjen D, Socaci R, Todea D, Antal M, Cioara T, Anghel I, Antal C (2022) Identifying and monitoring the daily routine of seniors living at home. Sensors 22(3):992. https://doi.org/10.3390/s22030992
10. Yao L, Sheng QZ, Li X, Gu T, Tan M, Wang X, Wang S, Ruan W (2017) Compressive representation for device-free activity recognition with passive rfid signal strength. IEEE Trans Mob Comput 17(2):293–306. https://doi.org/10.1109/TMC.2017.2706282

11. Nisar K, Ibrahim AAA, Park YJ, Hzou YK, Memon SK, Naz N, Welch I (2019) Indoor roaming activity detection and analysis of elderly people using rfid technology. In: 2019 1st international conference on Artificial Intelligence and Data Sciences (AiDAS). IEEE, pp 174–179. https://doi.org/10.1109/AiDAS47888.2019.8970780
12. Shahid ZK, Saguna S, Åhlund C (2023) Outlier detection in iot data for elderly care in smart homes. In: 2023 IEEE 20th Consumer Communications and Networking Conference (CCNC). IEEE, pp 1066–1073 (2023). https://doi.org/10.1109/CCNC51644.2023.10060085
13. Zekri D, Delot T, Thilliez M, Lecomte S, Desertot M (2020) A framework for detecting and analyzing behavior changes of elderly people over time using learning techniques. Sensors 20(24):7112. https://doi.org/10.3390/s20247112
14. Zhang Y, Liang W, Yuan X, Zhang S, Yang G, Zeng Z (2024) Deep learning based abnormal behavior detection for elderly healthcare using consumer network cameras. IEEE Trans Consum Electron 70(1):2414–2422. https://doi.org/10.1109/TCE.2023.3309852
15. Kim K, Lee S, Kim S, Kim J, Shin D, Shin D (2020) Sensor-based deviant behavior detection system using deep learning to help dementia caregivers. IEEE Access 8:136004–136013. https://doi.org/10.1109/ACCESS.2020.3011654
16. Gonzalez D, Patricio MA, Berlanga A, Molina JM (2022) Variational autoencoders for anomaly detection in the behaviour of the elderly using electricity consumption data. Expert Syst 39(4):e12744. https://doi.org/10.1111/exsy.12744
17. Bank D, Koenigstein N, Giryes R (2023) Autoencoders. In: Rokach L, Maimon O, Shmueli E (eds) Machine learning for data science handbook. Springer, Cham, pp 353–374. https://doi.org/10.1007/978-3-031-24628-9_16
18. Cook DJ (2012) Learning setting-generalized activity models for smart spaces. IEEE Intell Syst 27(1):32–38. https://doi.org/10.1109/MIS.2010.112

Sentiment Analysis on the Young People's Perception About the Mobile Internet Costs in Senegal

Derguene Mbaye, Madoune Robert Seye, Moussa Diallo, Mamadou Lamine Ndiaye, Djiby Sow, Dimitri Samuel Adjanohoun, Tatiana Mbengue, Cheikh Samba Wade, De Roulet Pablo, Jean-Claude Baraka Munyaka, and Jerome Chenal

Abstract Internet penetration rates in Africa are rising steadily, and mobile Internet is getting an even bigger boost with the availability of smartphones. Young people are increasingly using the Internet, especially social networks, and Senegal is no exception to this revolution. Social networks have become the main means of expression for young people. Despite this evolution in Internet access, there are few operators on the market, which limits the alternatives available in terms of value for money. In this paper, we will look at how young people feel about the price of mobile Internet in Senegal, in relation to the perceived quality of the service, through their comments on social networks. We scanned a set of Twitter and Facebook comments related to the subject and applied a sentiment analysis model to gather their general feelings.

Keywords Sentiment analysis · Social media · Social web · Language models · Low-resource · African languages · Low-resource languages · Wolof

D. Mbaye (✉) · M. R. Seye · M. Diallo · M. L. Ndiaye
Polytechnic School (ESP), Dakar, Senegal
e-mail: derguenembaye@esp.sn

M. R. Seye
e-mail: robertseye@hotmail.fr

M. Diallo
e-mail: moussa.diallo@ucad.edu.sn

D. Sow · D. S. Adjanohoun · T. Mbengue · C. S. Wade
Gaston Berger University (UGB), Saint Louis, Senegal

D. R. Pablo · J.-C. B. Munyaka · J. Chenal
Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_14

1 Introduction

Social networking has taken off in leaps and bounds around the world, and Africa is no exception. People share their opinions on all kinds of subjects, share achievements in their lives or simply engage in chit-chat. In 2024, the Digital Global Overview Report[1] recorded more than 5 billion active users on social networks, representing 62.3% of the world's population. In Senegal, 20.6%[2] of the population is on social networks, according to the same report. With a median age of around 18, this is a particularly young population, suggesting that they represent the vast majority of social network users. `Facebook`, `Twitter` and, more recently, `Tiktok` are the flagship platforms most used by the population. However, despite a growing mobile Internet penetration rate in Africa (one of the key factors in the rise of social networks in Africa), only two countries (South Africa and Mauritius) have achieved the 'advanced' status in the 2023 GSMA Connectivity Index.[3] The report shows that 42% of adults in low-income countries are still not using mobile internet, despite being covered by a mobile broadband network. Several factors were identified, including a lack of the necessary knowledge and skills, and the inability to afford an internet-connected phone, data plans and other service fees. To emphasize the latter, it is common to find posts on social media in Senegal about the cost of mobile internet, and end-users' perception of network quality, among other things. `Twitter` in particular is an ideal source because of its audience, the variety of its users and its micro-blogging nature [1] facilitating the sharing of opinions through short messages.[4]

There are five (05) operators active in Senegal, including `Orange`, which has the largest market share, as shown in Fig. 1.

`Orange` is the brand name of `SONATEL`, the country's incumbent operator, with `France Telecom` as its majority shareholder since its privatization in 1997. It therefore occupies a dominant position, having own most of the infrastructure in place. Although it has the highest number of active users, it is not uncommon to see negative reviews on social networks about the cost of accessing its services, suggesting inequalities in coverage and quality of service between operators.

In this article, we study the opinions of young people on the cost of the mobile Internet in Senegal, using social networking platforms such as `Twitter` and `Facebook`. We show how a corpus can be built up through these platforms and how it can be used as a pressure tactic on telecoms operators. We collected a corpus of more than 10.000 text posts distributed between three types of sentiments:

1. texts containing `positive` emotions, such as happiness, amusement or joy;
2. texts containing `negative` emotions, such as sadness, anger or disappointment;
3. objective or `neutral` texts that only state a fact or do not express any emotions.

[1] https://datareportal.com/reports/digital-2024-global-overview-report.

[2] Social media users may not represent unique individuals.

[3] https://www.mobileconnectivityindex.com/index.html#year=2023.

[4] Twitter offers longer messages since 2023 through its premium feature, but short messages are still the favorite format among users.

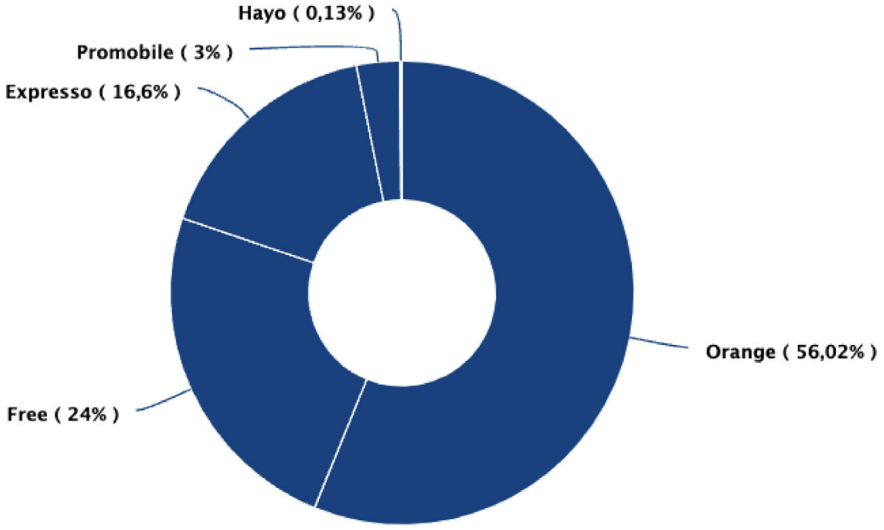


Fig. 1 Operator market shares as of March 31, 2024, published by the Senegalese Telecommunications Regulatory Agency (ARTP)

We perform a linguistic analysis of our corpus and use a multilingual language model (LM) as a sentiment classifier to illustrate users' feelings. The paper is therefore structured as follows:

- We begin by presenting some work done on sentiment analysis applied to various fields, including telecoms in Sect. 2.
- Our data collection approach is presented in Sect. 3.
- Analysis of collected data is performed in Sect. 4.
- In Sect. 5, we present an analysis of the extracted sentiments and the approach adopted.
- We present some limitations of our methodology in Sect. 6
- Conclusion and perspectives are presented in Sect. 7.

2 Related Work

Opinions on social networks have been the subject of many studies in the literature, on subjects ranging from politics to health issues and natural disasters, among others. Researchers in [1] propose an efficient way to collect text data from Twitter for sentiment analysis and opinion mining. Their method allows automatic text collection based on sentiment (positive or negative) in such a way that human intervention is not required for classification. They worked on the English language, although the method is reusable in other languages. A sentiment analysis benchmark on African

languages was proposed in [2] covering 14 languages from 04 different families with data sourced from Twitter. Tweets were manually labeled by native speakers, highlighting the challenges of working with African languages. Regarding the telecommunications field, the customer feedback and review on mobile telecommunication services in Malaysia has been studied in [3] using a Naïve Bayes sentiment analysis approach on Twitter data. Researchers in [4] studied user complaints about internet quality during the COVID-19 pandemic in Indonesia with a CNN-based classifier to classify feelings about telecom operators. Data were collected on Twitter and underwent pre-processing and weighting of Word2Vec embeddings. To enhance the quality of service provided by Mobile Phone operators working in Pakistan, Twitter data has been analyzed in [5] to perform sentiment analysis in order to enable organizations to gain better insights regarding quality of service improvement. A framework has been proposed in [6], built on top of the Hadoop ecosystem, for analyzing data from Twitter using a domain-specific Lexicon in Greek.

Twitter has been extensively studied in the literature, but due to the limitations imposed since its takeover by Elon Musk, it has become extremely restrictive to limit oneself to this platform.[5] Since then, hundreds of research projects have been canceled, halted, or pivoted to other platforms as a result of these changes,[6] and a significant decline in the commitment of researchers has been noted [7]. Facebook is a viable alternative, and similar work to that done on Twitter has been carried out there. Researchers in [8] used Facebook for tracking the evolution of COVID-19 related trends. They collected a multilingual corpus covered 07 languages (English, Arabic, Spanish, Italian, German, French, and Japanese) and proposed an exhaustive analytics process including data gathering, pre-processing, LDA-based topic modeling and a presentation module using graph structure. A case study on text mining for Facebook and Twitter unstructured data analysis has been conducted in [9] to help financial institutions in Nigeria analyze their competitor's social media sites and improve their decision-making. A sentiment analysis of Facebook comments was carried out in [10] concerning the presidential election in Indonesia in 2014. The authors targeted two official accounts among those of the candidates and used a Naïve Bayes classifier for sentiment analysis. A similar study was carried out in [11] on the June 2017 local government campaign in the central state of Mexico. Researchers collected nearly 4,500 Facebook posts and performed a sentiment analysis on the text including emoticons raising a surprising paradox between perceived sentiment toward candidates and the actual election outcome. The perception of UBER[7] users has been studied in [12] on the basis of Facebook posts published between July 2016 and July 2017. Corpus collection and sentiment analysis were carried out using a proprietary tool relying on a lexicon-based approach to categorizing sentiment. Using Facebook and Twitter accounts of the top three telecommunication companies in Ghana, researchers in [13] reveal insights from unstructured texts. They studied the

[5] What's really going on with Twitter?—Data Reportal.

[6] Q&A: What happened to academic research on Twitter?—CJR.

[7] American multinational transportation company that provides ride-hailing services, courier services, food delivery, and freight transport.

customers feelings about operators products or brand using a lexicon-based approach. The opinions of the Internet service in Sudan have been studied in [14] in the Arabic language and the Sudanese dialect, which is a low-resource setting. They applied an SVM classifier and another one based on Naïve Bayes on a corpus of a thousand Facebook comments.

3 Data Collection

An important characteristic of Twitter is its real-time nature. For example, when a disaster occurs, people make many Twitter posts (tweets) related to it, which enables its following easier simply by observing the tweets. However, due to limitations within the twitter platform, we were only able to collect a very limited extract of `250 tweets`. These tweets date back to a window of one week from the date of collection (September 23, 2024), and concern comments on a post made by the Orange operator on the lowering of mobile internet package prices. These types of posts are very rich in information, as users naturally bounce off them to share their opinions on package costs and how they feel about various aspects of the operator. We therefore adopted a similar approach to collecting information on Facebook, by targeting operators' posts on their Internet packages and then collecting the corresponding comments. To do this, we first performed a keyword search to identify these posts, collected the URLs to them and used Bright Data's[8] scraping API to retrieve the corresponding snapshots. These snapshots represent all the data collected from the URLs provided, which we then downloaded and stored in `csv` format. This last step was skipped for the Twitter scraping, where we used the Twitter API directly to download the data without using snapshots. We thus collected more than `10,000` Facebook and Twitter comments spread over 4 operators. The overall approach is illustrated in Fig. 2.

We then proceeded to a data pre-processing stage in which we used regular expressions to remove all the hashtags, usernames and links of collected tweets. We also removed stop words, images and unnecessary columns returned during scraping. Stop words are a set of commonly used words in a language (e.g., "a," "the," "is"...) used in text mining and natural language processing (NLP) to eliminate words that are so widely used that they carry very little useful information. Stopwords lexicons for resource-rich languages like English are easily accessible on the Internet, but not for a language like Wolof (present in the collected data), making pre-processing difficult. Moreover, Wolof is written on social media in a way that differs from standard writing as illustrated in [15]. This phenomenon results from the fact that Wolof is not taught at school, due to the fact that French has been adopted as the official language since colonization. As a result, people don't have a good grasp of its script, and tend to write it without regard to any writing rules, even though the language's alphabet is established. This difference in script tends to reduce the performance of existing language processing tools for Wolof, which are generally designed for the standard script.

[8] https://brightdata.com/.

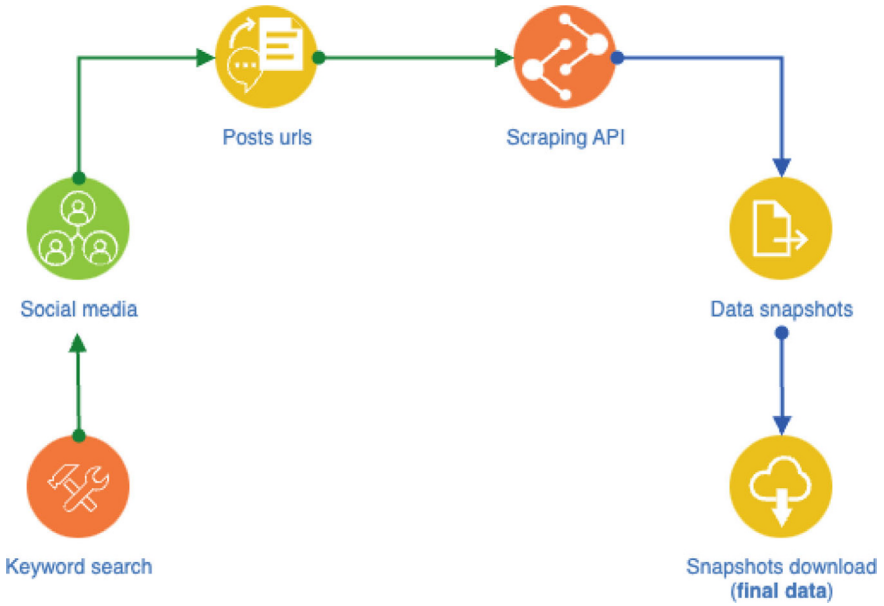


Fig. 2 Diagram of the scraping strategy

4 Data Analysis

We were able to collect a total of `10184 posts`, originating mainly from Facebook, as illustrated in Fig. 3.

The data has been collected from the 04 main operators in Senegal, and the distribution of their data is shown in Fig. 4.

The scraped data range from 2019 to 2024 for Facebook data and around September 2024 for Twitter data. Our access level to the Twitter API did not allow us to obtain the publication dates of the comments, therefore only those from Facebook are shown in Fig. 5. It's worth noting that the majority of comments relating to Orange were posted around 2019, while those for Free (its main competitor) were posted between 2021 and 2024. Promobile has the most recent comments (2024), while Expresso's comments are concentrated between 2023 and 2024. The period is therefore very important to consider, as external circumstances can influence user sentiment toward a particular operator.

With two dominant languages, French (the official language) and Wolof (the lingua franca), it's not uncommon to see comments in both languages on social networks. People also tend to mix the two in the same speech, a phenomenon known as code mixing or code switching [16], typical of countries where several languages are present. To identify the languages of the collected comments, we used the GlotLID library [17], capable of identifying over 1,500 languages, including a wide coverage of low-resource ones. We were thus able to identify a higher proportion of texts

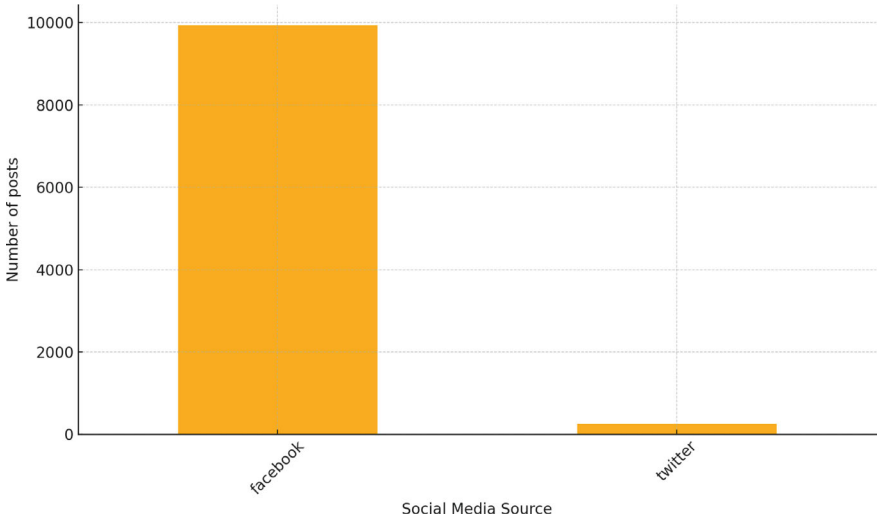


Fig. 3 Distribution of Facebook versus Twitter data

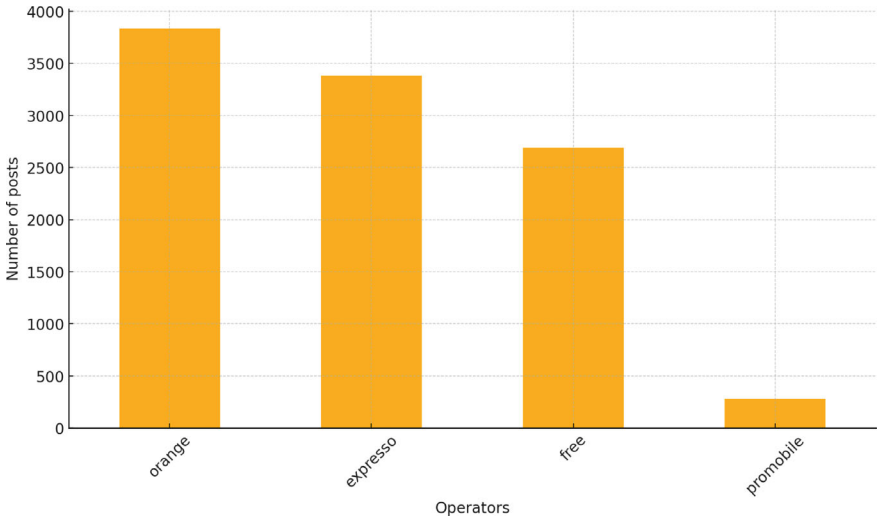


Fig. 4 Data distribution across Senegal's 04 leading telecom operators

in Wolof compared to those in French, as shown in Fig. 6. For this reason, it is essential to take both languages into account when pre-processing data. Since we are mostly dealing with text data, it is quite handy to take a look at the list of all the possible words that were present in our posts (Twitter and Facebook comments). A useful concept called `word cloud` can be used for this task of representing the presence of various words in our list of comments. The size of the words represents the

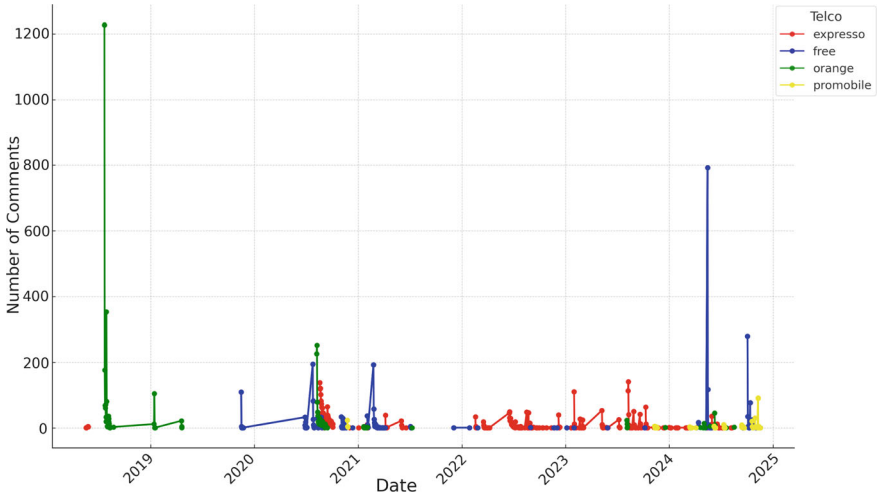


Fig. 5 Data distribution across Senegal's 04 leading telecom operators

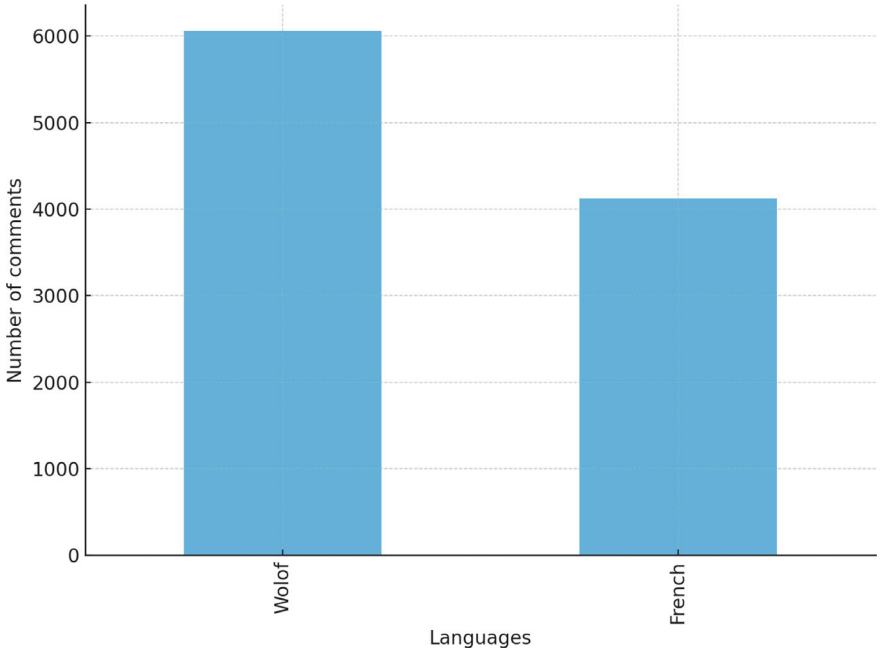


Fig. 6 Proportion of comments in French compared to those in Wolof

frequency of occurrence and we can quickly have an overview of the most used words based on this information. Given that stop words are very frequent but not significant, keeping them will tend to "spam" the word clouds. To remove them, we used the

Natural Language Toolkit (NLTK) [18], which is a popular suite of libraries and programs for symbolic and statistical natural language processing (NLP). It natively integrates French stop words, but not those in Wolof. To mitigate this constraint, we manually translated some French stop words into Wolof and then generated word clouds, identified residual stop words and added them to those in NLTK to clean up the final word clouds as much as possible. We finally generated word clouds on the comments from each operator and can observe an interesting phenomenon from here in terms of the words used for Orange compared to its competitors. In Fig. 7, we see very hostile terms like `voleur` (thief), `arnaque` (scam), `boycotte` (boycott) or `beugouniou` (a Wolof term meaning "we don't want" or "we don't want it"). Indeed, Orange is generally criticized by the population for the high cost of its packages and the speed with which users perceive the consumption of their plans. As a result, they feel "robbed" and often call for a boycott.

In Free's word cloud in Fig. 8, terms like `front`, `contre` (against), `cherté` (expensive), and `coût` (cost) appear, illustrating Free's position as the operator who fights against the high cost of packages. Free is seen as a viable alternative to Orange, as is Expresso, but both operators are often criticized for the lower quality of their networks. For this reason, we see terms like `réseau` (network), `connexion` (connection), `problème` (problem) in their word clouds, especially in that of Expresso presented in Fig. 9. We note a similar trend for promobile illustrated in Fig. 10, a virtual operator based on the Orange network, singled out for its potentially expensive packages and still very poor network coverage. The trends observed in this disparity partly explain the dominance of Orange, despite the hostile terms often noted in user comments. Despite calls for boycotts, users find it hard to switch to a competitor, as the latter often offers inferior network quality and coverage.

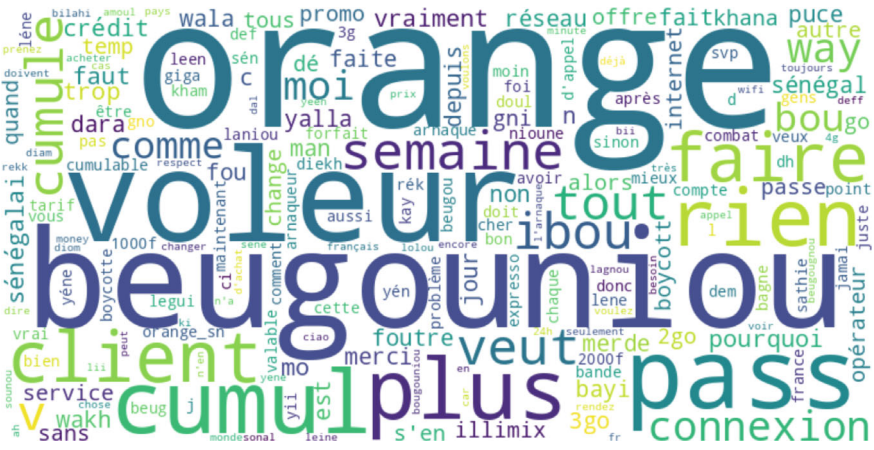


Fig. 7 Word cloud of the most frequent words used in the comments from the Orange operator Senegal

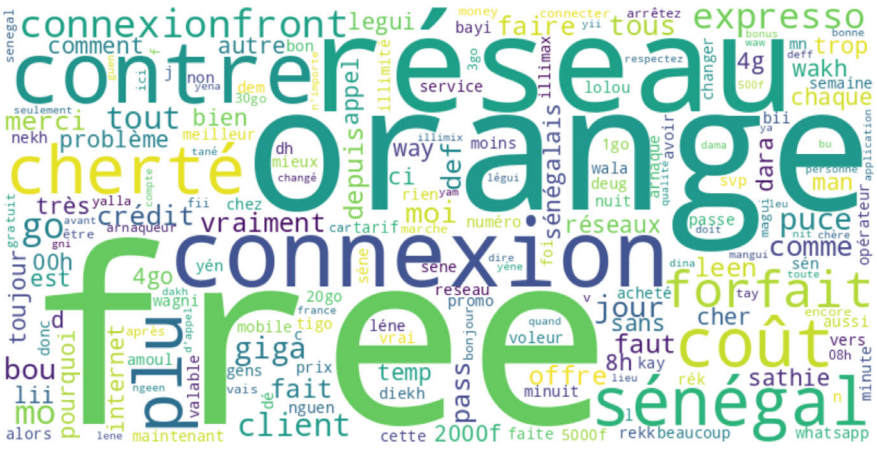


Fig. 8 Word cloud of the most frequent words used in the comments from the free operator Senegal

Fig. 9 Word cloud of the most frequent words used in the comments from the expresso operator Senegal

5 Sentiment Analysis

Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral. To perform this task, we first used the GPT4o model [19], which is OpenAI's state-of-the-art Large Language Model. It is a multimodal model with text, visual and audio input and output capabilities, building on the previous iteration of OpenAI's GPT-4 with Vision model, GPT-4 Turbo. However, the capabilities of this model on low-resource languages like Wolof are limited, although outperforming open-source alternatives as studied

Fig. 10 Word cloud of the most frequent words used in the comments from the promobile operator Senegal

in [20]. The preliminary tests we carried out on our Wolof data highlighted the limitations of this model, which tended to systematically classify the sentiment of Wolof texts as neutral. To remedy this, we used the Google Translate API[9] to translate all Wolof texts into French before sentient extraction. Google's translation model offers greater robustness to variations in Wolof writing than specialized models like the one presented in [21]. Despite this switch to French, which is better supported by GPT4o and multilingual models in general, we observed the same behavior with an over classification to the neutral sentiment. To mitigate this aspect, we used XLM-T presented in [22], which is a multilingual model based on XLM-Roberta [23] and pretrained on nearly 200 million Tweets across some 30 languages (including French). The authors show that a domain-specific model (in this case, social media) is more effective than its general counterpart when it comes to refining task-specific multilingual Language Models. The sentiments obtained on the Orange data are illustrated in Fig. 11, and show a strong negative connotation, as does the overview obtained on the word clouds. In fact, the price of the packages is generally raised with network problems from time to time, even if in general the network is more stable. We observe a similar trend for Free's data in Fig. 12 and Expresso's in Fig. 13, but with a higher proportion of positive comments. Subscribers to these two operators are on the whole satisfied with their plans, but more often decry network quality and coverage. A wave of indignation is however increasingly noted over recent package prices suggesting a subtle increase. Promobile shows the highest proportion of positive comments relative to its total number of reviews as illustrated in Fig. 14. This may be explained by its very recent arrival on the market, and its attempt to maintain attractive prices to attract customers. It is, however, a virtual operator backed by the Orange network, and is nevertheless associated with it in the eyes

[9] https://cloud.google.com/translate.

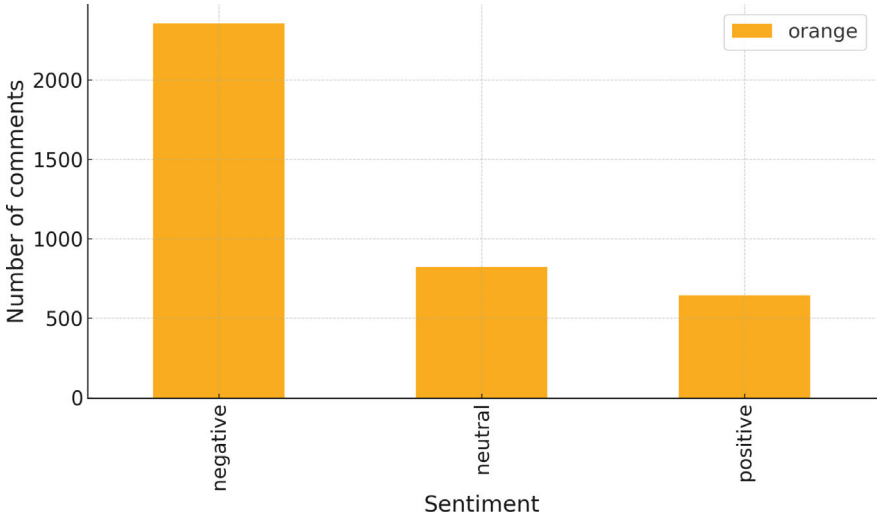


Fig. 11 Distribution of user sentiment toward Internet packages and network of the Orange operator

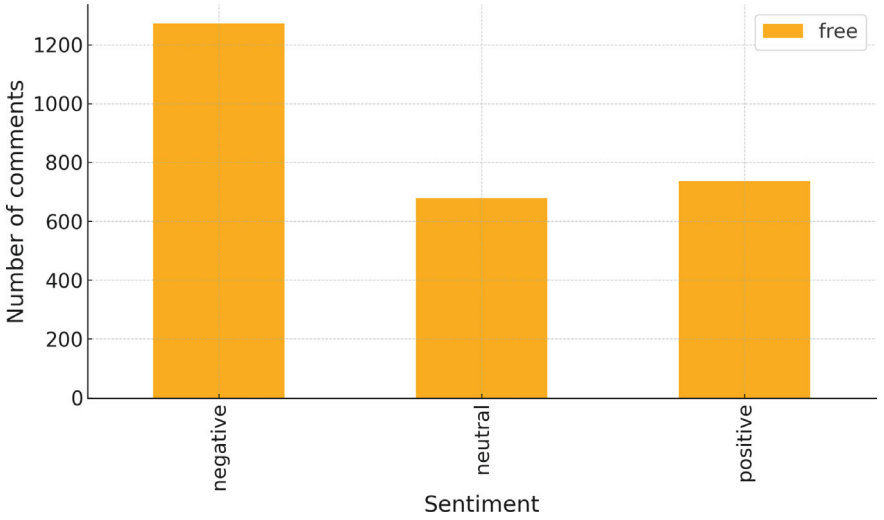


Fig. 12 Distribution of user sentiment toward internet packages and network of the free operator

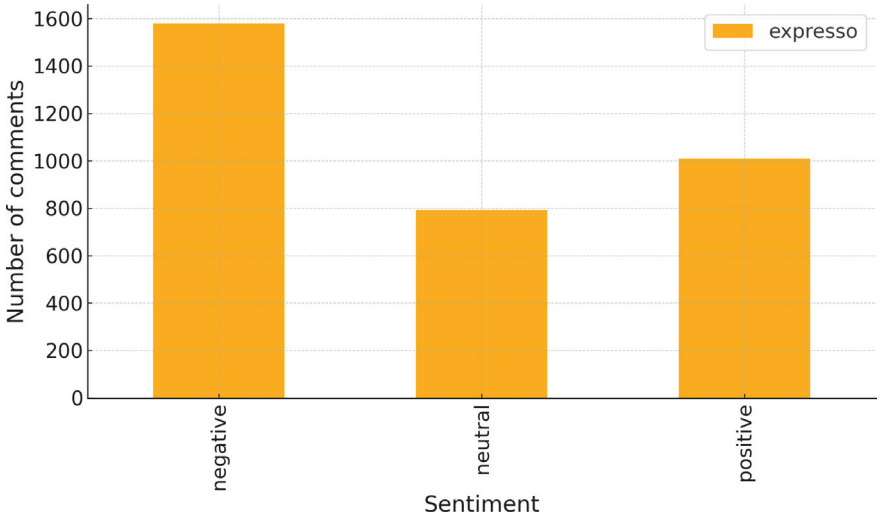


Fig. 13 Distribution of user sentiment toward internet packages and network of the expresso operator

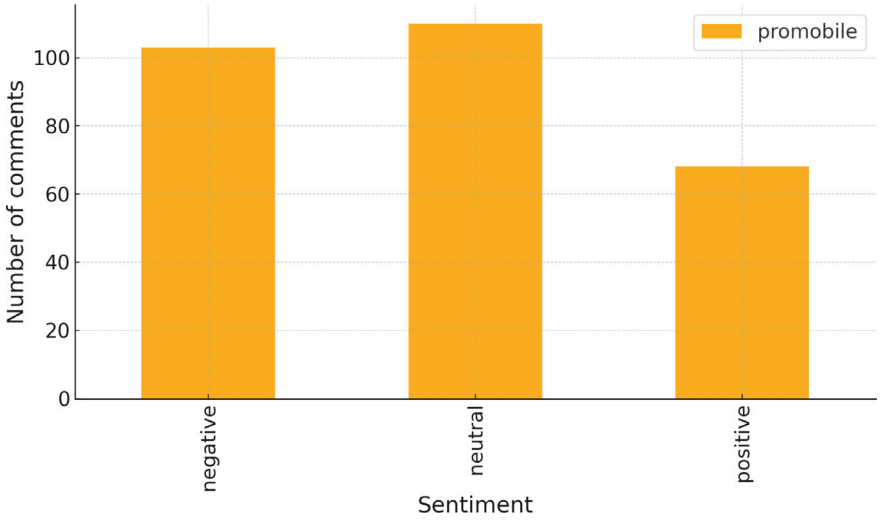


Fig. 14 Distribution of user sentiment toward internet packages and network of the promobile operator

of users which is not much appreciated by them. This analysis points to a general dissatisfaction with operators on the part of users, as the trade-off between network quality and affordability is difficult for operators to satisfy. Better regulation of the local telecoms market and support mechanisms from the government could lead to a significant drop in costs, as well as greater attractiveness. An entire sector of the economy is developing in Senegal around Internet connectivity, such as e-commerce, delivery, "uberization," e-sport, and the development of these activities is still very much affected by the accessibility of a high-quality network.

6 Limitation

Despite the efficiency of our approach to studying data trends, particularly in a low-resource language, we note a few limitations. The over-representation of Facebook data may induce biases, as users may have a different behavior on Twitter. Restrictions on the latter, however, make it difficult to collect a substantial data on this platform to balance the final corpus. Our approach also relies mainly on the translate-test principle, which involves translating data from a source language to a target one, in order to use tools or approaches that work better in the target language. This method has proved highly effective in low-resource environments as studied in [24], but errors in the translation process tend to propagate to the subsequent steps, potentially inducing additional bias.

7 Conclusion

In this paper, we studied the sentiment of Senegalese users toward the cost of accessing Internet services from established operators. We collected a substantial corpus on Twitter and Facebook, the latter being the network with the highest concentration of users in Senegal. After an initial phase of data pre-processing, we studied the trends emerging from the most frequently used words in the comments. We took this analysis a step further by highlighting the sentiments expressed in these posts, underlining a general dissatisfaction with the quality/price ratio of the various operators' offers. The overall study is, however, subject to a number of biases relating to data balance, which will need to be strengthened through further data collection. We also intend to carry out an annotation of the final corpus in order to build up an open sentiment analysis dataset for the Wolof language. This will ultimately provide a more suitable and therefore more powerful classification model to facilitate further studies on a variety of topics.

Acknowledgements This project has been funded by Fondation Botnar.

References

1. Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the seventh international conference on language resources and evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf
2. Muhammad S, Abdulmumin I, Ayele A, Ousidhoum N, Adelani D, Yimam S, Ahmad I, Beloucif M, Mohammad S, Ruder S, Hourrane O, Jorge A, Brazdil P, Ali F, David D, Osei S, Shehu-Bello B, Lawan F, Gwadabe T, Rutunda S, Belay TD, Messelle W, Balcha H, Chala S, Gebremichael H, Opoku B, Arthur S (2023) AfriSenti: A Twitter sentiment analysis benchmark for African languages. In: Bouamor H, Pino J, Bali K (eds) Proceedings of the 2023 conference on empirical methods in natural language processing. Association for Computational Linguistics, Singapore, pp 13968–13981. https://doi.org/10.18653/v1/2023.emnlp-main.862. https://aclanthology.org/2023.emnlp-main.862
3. Yuri MN, Rosli MM (2022) Telcosentiment: sentiment analysis on mobile telecommunication services. J Appl Res Multidiscip Stud 6(3). https://mail.journalppw.com/index.php/jpsp/article/view/5113
4. Amalia Z, Irfan M, Maylawati DS, Wahana A, Zulfikar WB, Ramdhani MA (2022) Sentiment analysis of the use of telecommunication providers on Twitter social media using convolutional neural network. In: 2022 IEEE 8th international conference on computing, engineering and design (ICCED), pp 1–6. https://doi.org/10.1109/ICCED56140.2022.10010357
5. Saleem I, Jamil A, Mehmood MA (2023) Employing sentiment analysis to enhance customer relationships for mobile phone operators working in Pakistan. J Appl Res Multidiscip Stud 4(1). https://doi.org/10.32350/jarms.41.09. https://journals.umt.edu.pk/index.php/jarms/article/view/4336
6. Skoularikis K, Savvas IK, Garani G, Kakarontzas G (2021) A scalable framework for customer sentiment analysis in the telecommunication industry. In: 2021 29th Telecommunications forum (TELFOR), pp 1–4. https://doi.org/10.1109/TELFOR52709.2021.9653423
7. Bisbee J, Munger K (2024) The vibes are off: Did Elon Musk push academics off Twitter? PS Polit Sci Polit, 1–8. https://doi.org/10.1017/S1049096524000416
8. Amara A, Hadj Taieb MA, Ben Aouicha M (2021) Multilingual topic modeling for tracking covid-19 trends based on Facebook data analysis. Appl Intell 51(5):3052–3073
9. Ayo C, Ezenwoke A, Ibukun A (2017) Competitive analysis of social media data in the banking industry. Int J Internet Mark Advert 11:183. https://doi.org/10.1504/IJIMA.2017.10006719
10. Syahriani YAA, Santoso T (2020) Sentiment analysis of Facebook comments on Indonesian presidential candidates using the Naïve Bayes method. J Phys Conf Ser 1641(1):012. https://doi.org/10.1088/1742-6596/1641/1/012012
11. Sandoval-Almazan R, Valle-Cruz D (2018) Facebook impact and sentiment analysis on political campaigns. In: Proceedings of the 19th annual international conference on digital government research: governance in the data age. Association for Computing Machinery, New York, NY. https://doi.org/10.1145/3209281.3209328
12. Baj-Rogowska A (2017) Sentiment analysis of Facebook posts: the Uber case. In: 2017 Eighth international conference on intelligent computing and information systems (ICICIS), pp 391–395. https://doi.org/10.1109/INTELCIS.2017.8260068
13. Afful-Dadzie E, Nabareseh S, Oplatková ZK, Klímek P (2014) Enterprise competitive analysis and consumer sentiments on social media—insights from telecommunication companies. In: Proceedings of 3rd international conference on data management technologies and applications—volume 1: DATA, INSTICC. SciTePress, pp 22–32. https://doi.org/10.5220/0004991300220032
14. Heamida Saif Eldin Mukhtar I, Samani Abd Elmutalib Ahmed AL (2021) The classification model sentiment analysis of the Sudanese dialect used into the internet service in Sudan. In: Hassanien AE, Darwish A, Abd El-Kader SM, Alboaneen DA (eds) Enabling machine learning applications in data science. Springer Singapore, Singapore, pp 369–378

15. Mbaye D, Diallo M (2023) BEGI: revitalize the Senegalese Wolof language with a robust spelling corrector. https://arxiv.org/abs/2305.08518, 2305.08518
16. Sitaram S, Chandu KR, Rallabandi SK, Black AW (2020) A survey of code-switched speech and language processing. https://arxiv.org/abs/1904.00784, 1904.00784
17. Kargaran AH, Imani A, Yvon F, Schuetze H (2023) GlotLID: language identification for low-resource languages. In: Bouamor H, Pino J, Bali K (eds) Findings of the association for computational linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp 6155–6218. https://doi.org/10.18653/v1/2023.findings-emnlp.410. https://aclanthology.org/2023.findings-emnlp.410
18. Loper E, Bird S (2002) NLTK: the natural language toolkit. https://arxiv.org/abs/cs/0205028, cs/0205028
19. OpenAI, Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow A, Welihinda A, Hayes A, Radford A, Madry A, Baker-Whitcomb A, Beutel A, Borzunov A, Carney A, Chow A, Kirillov A, Nichol A, Paino A, Renzin A, Passos AT, Kirillov A, Christakis A, Conneau A, Kamali A, Jabri A, Moyer A, Tam A, Crookes A, Tootoochian A, Tootoonchian A, Kumar A, Vallone A, Karpathy A, Braunstein A, Cann A, Codispoti A, Galu A, Kondrich A, Tulloch A, Mishchenko A, Baek A, Jiang A, Pelisse A, Woodford A, Gosalia A, Dhar A, Pantuliano A, Nayak A, Oliver A, Zoph B, Ghorbani B, Leimberger B, Rossen B, Sokolowsky B, Wang B, Zweig B, Hoover B, Samic B, McGrew B, Spero B, Giertler B, Cheng B, Lightcap B, Walkin B, Quinn B, Guarraci B, Hsu B, Kellogg B, Eastman B, Lugaresi C, Wainwright C, Bassin C, Hudson C, Chu C, Nelson C, Li C, Shern CJ, Conger C, Barette C, Voss C, Ding C, Lu C, Zhang C, Beaumont C, Hallacy C, Koch C, Gibson C, Kim C, Choi C, McLeavey C, Hesse C, Fischer C, Winter C, Czarnecki C, Jarvis C, Wei C, Koumouzelis C, Sherburn D, Kappler D, Levin D, Levy D, Carr D, Farhi D, Mely D, Robinson D, Sasaki D, Jin D, Valladares D, Tsipras D, Li D, Nguyen DP, Findlay D, Oiwoh E, Wong E, Asdar E, Proehl E, Yang E, Antonow E, Kramer E, Peterson E, Sigler E, Wallace E, Brevdo E, Mays E, Khorasani F, Such FP, Raso F, Zhang F, von Lohmann F, Sulit F, Goh G, Oden G, Salmon G, Starace G, Brockman G, Salman H, Bao H, Hu H, Wong H, Wang H, Schmidt H, Whitney H, Jun H, Kirchner H, de Oliveira Pinto HP, Ren H, Chang H, Chung HW, Kivlichan I, O'Connell I, O'Connell I, Osband I, Silber I, Sohl I, Okuyucu I, Lan I, Kostrikov I, Sutskever I, Kanitscheider I, Gulrajani I, Coxon J, Menick J, Pachocki J, Aung J, Betker J, Crooks J, Lennon J, Kiros J, Leike J, Park J, Kwon J, Phang J, Teplitz J, Wei J, Wolfe J, Chen J, Harris J, Varavva J, Lee JG, Shieh J, Lin J, Yu J, Weng J, Tang J, Yu J, Jang J, Candela JQ, Beutler J, Landers J, Parish J, Heidecke J, Schulman J, Lachman J, McKay J, Uesato J, Ward J, Kim JW, Huizinga J, Sitkin J, Kraaijeveld J, Gross J, Kaplan J, Snyder J, Achiam J, Jiao J, Lee J, Zhuang J, Harriman J, Fricke K, Hayashi K, Singhal K, Shi K, Karthik K, Wood K, Rimbach K, Hsu K, Nguyen K, Gu-Lemberg K, Button K, Liu K, Howe K, Muthukumar K, Luther K, Ahmad L, Kai L, Itow L, Workman L, Pathak L, Chen L, Jing L, Guy L, Fedus L, Zhou L, Mamitsuka L, Weng L, McCallum L, Held L, Ouyang L, Feuvrier L, Zhang L, Kondraciuk L, Kaiser L, Hewitt L, Metz L, Doshi L, Aflak M, Simens M, Boyd M, Thompson M, Dukhan M, Chen M, Gray M, Hudnall M, Zhang M, Aljubeh M, Litwin M, Zeng M, Johnson M, Shetty M, Gupta M, Shah M, Yatbaz M, Yang MJ, Zhong M, Glaese M, Chen M, Janner M, Lampe M, Petrov M, Wu M, Wang M, Fradin M, Pokrass M, Castro M, de Castro MOT, Pavlov M, Brundage M, Wang M, Khan M, Murati M, Bavarian M, Lin M, Yesildal M, Soto N, Gimelshein N, Cone N, Staudacher N, Summers N, LaFontaine N, Chowdhury N, Ryder N, Stathas N, Turley N, Tezak N, Felix N, Kudige N, Keskar N, Deutsch N, Bundick N, Puckett N, Nachum O, Okelola O, Boiko O, Murk O, Jaffe O, Watkins O, Godement O, Campbell-Moore O, Chao P, McMillan P, Belov P, Su P, Bak P, Bakkum P, Deng P, Dolan P, Hoeschele P, Welinder P, Tillet P, Pronin P, Tillet P, Dhariwal P, Yuan Q, Dias R, Lim R, Arora R, Troll R, Lin R, Lopes RG, Puri R, Miyara R, Leike R, Gaubert R, Zamani R, Wang R, Donnelly R, Honsby R, Smith R, Sahai R, Ramchandani R, Huet R, Carmichael R, Zellers R, Chen R, Chen R, Nigmatullin R, Cheu R, Jain S, Altman S, Schoenholz S, Toizer S, Miserendino S, Agarwal S, Culver S, Ethersmith S, Gray S, Grove S, Metzger S, Hermani S, Jain S, Zhao S, Wu S, Jomoto S, Wu S, Shuaiqi, Xia, Phene S, Papay S, Narayanan S, Coffey S, Lee S, Hall S, Balaji S, Broda T, Stramer T, Xu T, Gogineni

T, Christianson T, Sanders T, Patwardhan T, Cunninghman T, Degry T, Dimson T, Raoux T, Shadwell T, Zheng T, Underwood T, Markov T, Sherbakov T, Rubin T, Stasi T, Kaftan T, Heywood T, Peterson T, Walters T, Eloundou T, Qi V, Moeller V, Monaco V, Kuo V, Fomenko V, Chang W, Zheng W, Zhou W, Manassra W, Sheu W, Zaremba W, Patil Y, Qian Y, Kim Y, Cheng Y, Zhang Y, He Y, Zhang Y, Jin Y, Dai Y, Malkov Y (2024) GPT-4o system card. https://arxiv.org/abs/2410.21276, 2410.21276

20. Adelani DI, Ojo J, Azime IA, Zhuang JY, Alabi JO, He X, Ochieng M, Hooker S, Bukula A, Lee ESA, Chukwuneke C, Buzaaba H, Sibanda B, Kalipe G, Mukiibi J, Kabongo S, Yuehgoh F, Setaka M, Ndolela L, Odu N, Mabuya R, Muhammad SH, Osei S, Samb S, Guge TK, Stenetorp P (2024) Irokobench: a new benchmark for African languages in the age of large language models. https://arxiv.org/abs/2406.03368, 2406.03368
21. Mbaye D, Diallo M, Diop TI (2024) Low-resourced machine translation for Senegalese Wolof language. In: Yang XS, Sherratt RS, Dey N, Joshi A (eds) Proceedings of eighth international congress on information and communication technology. Springer Nature Singapore, Singapore, pp 243–255
22. Barbieri F, Anke LE, Camacho-Collados J (2022) XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. https://arxiv.org/abs/2104.12250, 2104.12250
23. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2020) Unsupervised cross-lingual representation learning at scale. https://arxiv.org/abs/1911.02116, 1911.02116
24. Chen Y, Shah V, Ritter A (2024) Translation and fusion improves zero-shot cross-lingual information extraction. https://arxiv.org/abs/2305.13582, 2305.13582

A Finite-State Morphological Analyzer for Ge'ez Verbs

Tebatso Gorgina Moape, Elleni Aschalew Zeleke, Ernest Mnkandla, and Sirgiw Gelaw Eggigu

Abstract This study addresses the challenge of processing the complex morphology of Ge'ez, an ancient southeast Semitic liturgical language. The research develops a comprehensive finite-state morphological analyzer and generator for all Ge'ez verb categories using bidirectional finite-state technology. The complexity of Ge'ez's non-concatenative morphology, where consonantal roots receive vowel patterns through interdigitation and the absence of native speakers, presents unique challenges for computational processing. The study implements a rule-based analyzer using Foma's finite-state framework and adopts the washära classification system, which recognizes eight head verbs. The morphological analyzer integrates finite-state transducers with lexc-based lexicon development, incorporating roots, affixes, vowel intercalation rules, and morphological alternations. For evaluation, a gold-standard dataset comprising 1365 verbs was compiled from the Ge'ez Bible and prayer book, with manual annotation by Ge'ez experts. The analyzer achieved an accuracy of 97.29% and a precision of 80.24% when evaluated against the gold-standard dataset, demonstrating significant improvement over previous approaches. Compared to earlier studies that focused on single verb categories or achieved limited accuracy, this analyzer successfully processes all verb types, including irregular verbs, and provides analysis and generation capabilities. This tool establishes a foundation for developing advanced NLP applications in Ge'ez, including machine translation and lexicography.

Keywords Ge'ez morphology · Finite-state transducers · Morphological analysis · Semitic language

T. G. Moape (✉) · E. A. Zeleke · E. Mnkandla
University of South Africa, Florida Park, South Africa
e-mail: Moapetg@unisa.ac.za

S. G. Eggigu
Addis Ababa University, Addis Ababa, Ethiopia

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_15

1 Introduction

In recent years, there has been growing interest in Natural Language Processing (NLP) and its application to various languages. Morphological analysis, the study of word structure and formation, is a fundamental component of many NLP applications, such as machine translation, text-to-speech synthesis, information retrieval, question-answering systems, word sense disambiguator, and sentiment analysis [1–4]. This task is crucial for morphologically rich and highly inflected languages, where a single root can generate many word forms.

Ge'ez is one such language with complex morphology. It is a part of the southeast Semitic language family [5, 6] and stands as one of the world's ancient languages with profound historical significance. Until the twelfth century, it served as Ethiopia's official language before being gradually replaced by Amharic and other local languages [7]. Today, while having no native speakers, Ge'ez continues to function as the liturgical language of the Ethiopian Orthodox Tewahido Church, the Eritrean Orthodox Tewahido Church, and the Ethiopian Catholic Church.

Morphologically, Ge'ez exhibits non-concatenative patterns, specifically root-pattern morphology, where consonantal roots receive vowel patterns through interdigitation [8, 9]. The language's verbs represent its most inflectional and productive element, with a single verb capable of generating hundreds of word forms through prefixation and suffixation. Verb classification follows different formats, washära, gonji, and walda with the washära school recognizing eight head verbs [10]. These head verbs serve as patterns for troops, following similar inflection and derivation patterns. Regarding transliteration, System for Ethiopic Representation in ASCII (SERA) is commonly used alongside the International Phonetic Alphabet (IPA) [11]. This study uses SERA to transliterate Ge'ez letters in developing the morphological analyzer and IPA to document the research conducted.

Previous studies have explored morphological analysis for Ge'ez, with some studies focusing on rule-based or data-driven approaches. These efforts have used methods such as memory-based learning [12] and two-level morphology [13]. These studies have contributed to the ongoing academic research work by focusing on classifying Ge'ez verbs and morphological analysis. However, these previous studies on Ge'ez morphological analysis have limitations. Some of them focused on only one verb category [13], while others had limited accuracy [12] or only performed analysis, not including the generation of words from roots and their grammatical information. Moreover, the existing approaches for Ge'ez have not fully addressed the challenges posed by the non-concatenative morphology of the language, where vowels are interdigitated into consonant roots or irregular verb forms.

A gap in the Ge'ez language processing is the absence of a comprehensive and accurate morphological analyzer that can handle all categories of Ge'ez verbs, including the complex morphological alternations that occur during verb formation [14]. This absence of a morphological analyzer hinders the development of more advanced NLP applications for Ge'ez. This study aims to address this gap by developing a finite-state-based morphological analyzer and generator for all categories of

Ge'ez verbs using a bidirectional finite-state technology (FST). FST is a computational framework based on finite-state machines that provides efficient methods for processing strings and modeling natural languages [15]. This technology has been widely used for various NLP tasks, including machine translation, spell-checking [16], speech recognition [17], and morphological analysis [17] for various languages.

The framework consists of mathematical models that can be in one of a finite number of states at any given time called Finite-state automata (FSA), Finite-state transducers (FST), which are extended FSAs that can perform input–output mappings between strings and regular expressions used for pattern matching and string manipulation that can be compiled into finite-state machines. FST is widely used in various NLP tasks, including morphological analysis of different languages, including Semitic languages as used in [18], Hebrew [19], and Amharic [20]. This technology offers the ability to handle concatenative and non-concatenative morphology. It provides high-speed and compact methods of handling lexicons and morphological rules [21]. In addition, the bidirectional feature of FST enables the use of the morphological analyzer as a morphological generator with reverse processing.

This research addresses the limitations of previous work by using the finite-state approach with the tool Foma, a finite-state compiler and library to capture both the concatenative and non-concatenative aspects of Ge'ez morphology. The main research questions were geared toward finding an appropriate Ge'ez verb classification, representing the non-concatenative morphology along with the morphotactics and orthographic rules of Ge'ez verbs using finite-state technology, creating the FSTs, developing a lexicon, and gold-standard evaluation dataset.

This paper is organized as follows: section two presents the five-step process designed to develop and evaluate the morphological analyzer for Ge'ez verbs. This section outlines the complete process of selecting the appropriate Ge'ez classification system, representing the non-concatenative morphology of Ge'ez, integration of morphotactics, and the orthographic rules in the creation of FSTs, lexicon development, implementation, evaluation, and results. Section three discusses the results while section four concludes the paper.

2 Related Works

Studies on the development of Ge'ez morphology are limited due to resource scarcity, the absence of native speakers, and the language's antiquity. However, finite-state technology continues to be used for morphological analysis in other languages. This section highlights studies that have employed FST to develop morphological analyzers for various languages. Research in [22] developed a morphological analyzer for Highland Puebla Nahuatl, a Uto-Aztecan language spoken in the state of Puebla in Mexico. As used in this study, the research used lex formalism for modeling the morphotactics and morphophonological alternations and obtained a precision of 95% on a manually created dataset. The approach parallels the current study, utilizing

transducers that map between surface and lexical forms, which enables both analysis and word generation capabilities.

In [23], a morphological analyzer for Q'eqchi' using Helsinki finite-state was developed. The resulting transducer consisted of 2610 states and 9558 transitions and covered between 75 and 85% of tokens in a Q'eqchi' corpus which lays the groundwork for future work in improved automatic corpus annotation in Q'eqchi'. A precision of 94.12% was achieved for the developed morphological analyzer in [24] using Stuttgart FST for Turkish. Compared to this study, the research extended its scope beyond morphology by incorporating both phonological and morphological rules to extract relevant linguistic information. The finite-state morphological analyzer for Turkish, was presented as a comprehensive tool covering Turkish inflection, derivation, and partial composition.

A Maithili morphological analyzer that handles the inflectional property of language was developed by [25]. While their study achieved 97% average accuracy across multiple parts of speech, including adverbs, adjectives, pronouns, and verbs, the current research focuses exclusively on verb morphology. Their finite-state transducer demonstrated the capability to generate inflected forms across diverse grammatical categories. Researchers in [26] developed a deterministic finite-state morphological analyzer for Urdu nominal system by focusing on inflections of noun forms and studying number, gender, person, and case representations. The developed system analyzed and generated possible forms of standardized Urdu registers, adding the necessary features and values while concatenating nouns based on their specific patterns. The system achieved an accuracy score of 92.70%. Similar to this study, [27] only focused on verbs for Tigrigna and developed a morphological analyzer FST. The study's morphological analysis incorporated an additional technique called memory-based learning hybridizing their method. The evaluation was conducted using optimized parameter settings for regular verbs and linguistic rules of the Tigrigna language allomorph and phonology for the irregular verbs. The system achieved an accuracy of 93.24.

These related works demonstrate the diverse applications of FST across different languages with varying morphological complexities. While many achieve high accuracy rates, the studies differ in scope and approach. Certain works focus on specific parts of speech, while others attempt comprehensive coverage. The commonality among successful implementations is the effective handling of language-specific morphological features through lexicon and FSAs development using FSTs.

3 Materials and Methods

The methodology employed in this study follows a systematic five-step process designed to develop and evaluate a morphological analyzer for Ge'ez verbs. This process consists of verb classification, morphological representation, implementation, lexicon creation, and evaluation phases. Within this systematic approach, step 1 focuses on determining the most appropriate Ge'ez verb classification system

for computational morphology. The challenge of representing non-concatenative morphology using FST is addressed in step 2. Step 3 involves the implementation phase, creating FSTs that represent morphotactics and orthographic rules. The lexicon is developed in step 4. Step 5 concentrates on the analyzer's evaluation using a gold-standard dataset, employing accuracy, and precision metrics.

Each step builds upon the previous ones, ensuring a comprehensive approach to developing a robust morphological analyzer. This methodology combines theoretical linguistic knowledge with computational implementation while incorporating expert validation throughout the process. The following sub-sections detail each step of the methodology.

3.1 Ge'ex Verb Classification

The first step involves determining which Ge'ez verb classification is appropriate for Ge'ez verb computational morphology. Research in [28] describes how Ge'ez verbs are classified according to prominent Ge'ez schools and states that the Ge'ez verbs are categorized from six to eight main/head verbs. For this study, the washära Ge'ez school's verb classification is deemed the most appropriate for Ge'ez verb computational morphology. This is due to its consistent patterns between head verbs and their associated troops. Compared to other classification systems gonji and walda, washära maintains uniform conjugation patterns that facilitate computational representation. Table 1 illustrates the three Ge'ez verb classifications.

In the Ge'ez verbal system, verbs are organized into heads and troops, where head verbs serve as representative patterns for their respective troops. While different classifications, washära, gonji, and walda, share some common head verbs, they differ in their total count and troop associations. The washära school, supported by numerous Ge'ez scholars, identifies eight distinct head verbs. Table 2 shows Ge'ez verb classification according to the washära classification of Ge'ez.

The washära system's key advantage is in its consistency, troop verbs strictly follow their head verbs' conjugation patterns. This contrasts with gonji and walda

Table 1 Washära, gonji, and walda, Ge'ez verbs classification systems

Washära	Walda	Gonji
ቀተለ-qätälä	ቀተለ-qätälä	ቀተለ-qätälä
ቀደሰ-qäddäsä	ቀደሰ-qäddäsä	ቀደሰ-qäddäsä
ገበረ-gäbrä	ማህረከ-mahräkä	–
አእመረ-äəmärä	ተንበለ-tänbälä	ማህረከ-mahräkä
ባረከ-baräkä	ባረከ-baräkä	ባረከ-baräkä
ሤመ-śemä	ሤሠየ-śesäyä	ጌገየ-gegäyä
ብህለ-bəhlä	ክህለ-kəhle	ጠበጠበ—ṭäbṭäbä
ቆመ-qomä	ጦመረ-ṭomärä	ኖለወ-noläwä

Table 2 Washära Ge'ez verbs classification

Washära	Adhana	Kifle
ቀተለ-qätälä	ቀተለ-qätälä	ቀተለ-qätälä
ቀደሰ-qäddäsä	ቀደሰ-qäddäsä	ቀደሰ-qäddäsä
ገበረ-gäbrä	ገበረ-gäbrä	–
አእመረ-äəmärä	ተንበለ-tänbälä	ማህረከ-mahräkä
ባረከ-baräkä	ባረከ-baräkä	ባረከ-baräkä
ሤመ-sémä	ኤለ-élä	ደገነ-degänä
ብህለ-bəhlä	ክህለ-kəhlä	ደንገጸ-dängäṣe

classifications, which at times exhibit variations in perfective forms or differences in radical numbers and assimilation. For computational implementation, the research defines vocalic patterns for all eight head verbs, which are then systematically applied to their corresponding troops. This forms the foundation for developing a comprehensive morphological analyzer using CV-templates, intercalation, alternation rules, and affixations within a finite-state framework.

3.2 Representing the Non-concatenative Morphology of Ge'ez Verbs Using Finite-State

With the verb classification system established, the next step focuses on representing Ge'ez's non-concatenative morphology using finite-state technology. Ge'ez, like other Semitic languages, is characterized by root-pattern morphology, where a root consisting of consonants is combined with a vocalic pattern (vowels) to form a stem. To achieve an efficient representation, FSTs in conjunction with regular expressions, flag diacritics, cv-patterns, lexicon and rule components, compositional operators, and bidirectional processing are used to effectively model the non-concatenative morphology of Ge'ez verbs by handling the complex interdigitation of vowels into consonants and morphophonological alternations.

A rule-based approach using FSTs, along with flag diacritics to handle the complexities of root-pattern morphology, is used to efficiently represent non-concatenative morphology. FSTs are used to implement the relationship between a word's lexical form and surface form together with the syntactical information about the word. In this case, the FSTs map the root with its feature tags to the surface verb form and vice-versa. For instance, an FST can be designed to map the lexical and surface forms of a word like '**ቀተለ**-qätälä,' as illustrated in Fig. 1, Table 3 provides a description of the lexical and surface forms of 'qätälä.

As depicted in the table, the surface level, "qätälä" appears as the complete verb form. This maps to four distinct components at the lexical level: [Verb1] which indicates that the word belongs to the first verb type category, [qtl] representing the consonantal root of the verb, [VPER] marking the perfective aspect of the verb

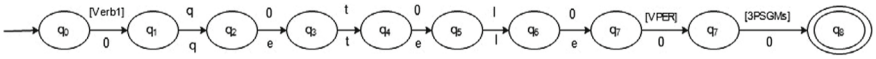


Fig. 1 FST for the word ቀ·ተ·ለ/qetele

Table 3 Lexical and surface form of-qetele

Surface level	Lexical level	Description
-qätälä	[Verb1]	Verb type 1
	[qtl]	root
	[VPER]	Perfective
	[3PSGMs]	3rd person singular male subject

and [3PSGM] which denotes the third person singular masculine subject agreement. Each lexical component serves a specific grammatical function. The description column provides clarification of these functions. This structured representation illustrates how the surface form encodes multiple morphological features, demonstrating the complex relationship between surface realization and underlying grammatical information in Ge'ez verb morphology.

Creating a finite-state transducer for a natural language involves first describing the language using regular expressions and then compiling these expressions into FSTs using finite-state tools such as Foma. Foma is a finite-state compiler that generates finite-state automata (FSA) and FSTs from regular expressions [29]. FSAs define languages by accepting strings that are part of the language while rejecting others, while FSTs, on the other hand, map input to output, allowing them to represent relationships between surface and lexical forms of words.

Regular expression operators supported by Foma are fully described in [30], including expressions such as define VARIABLE regular expression: The define command can be used to define regular expression function define function (prototype) regular expression. Moreover, the define command can be used to define a regular language, Concatenation X Y: The language or relation X concatenated with Y, Union X|Y: The union of languages or relations X and Y, Intersection X and Y: The intersection of languages X and Y, Optionality (X): Defines the language or relation that contains zero or one iteration of X, Kleene Star X*: Zero or more iterations of X, for example [X*] represent a language or a relation with empty set or zero or more iteration of X. so X* includes [], [X], [XX] and Kleene Plus X + : One or more iterations of X, for example [X +] represents a language or a relation with one or more as where X + includes [X], [XX], [XXX], etc.

Foma also includes flag diacritics, which are feature-setting operations that enforce constraints between different parts of words. These are particularly useful in handling non-concatenative morphology, such as the insertion of vowels within consonants as found in Ge'ez. Flag diacritics can set constraints on the co-occurrence of morphemes and are used for feature unification. This process follows specific CV-patterns, where templates and vocalic patterns are inserted into consonant roots to form verbal stems.

In order to design a morphological analyzer using Foma, two main components are needed**,** lexical and rule components. The lexicon is a transducer that consists of the language's roots and the appropriate feature tags expressing the morphotactics. For instance, the lexicon file may contain the following mapping: Cat + noun + plural—Catˆs.

The lexical transducer accepts valid words or lexemes along with feature tags and generates an intermediate output. As illustrated in the mapping above, "cat" is a noun, and the plural form cat is "catˆs." Lexical files are created using the lexc formalism. The rule component's role, on the other hand, is to manage the alternation rules that govern morphophonological changes during morpheme combination. These rules manage both vowel insertion and consonant modifications. Their main objective is to perform the necessary modification on the output of the lexical transducer based on the morphophonological rule of the language. For instance, the lexicon may contain a mapping indicating the concatenation of s to nouns to produce the plural form of a noun. However, some words such as "watch + s" result in "watches" rather than "watchs." The alternation rule transducer does this alternation.

Thus, the rule component is an intermediate between the lexical transducer and the output. The combination of lexicon transducers and rule transducers is achieved through the "composition.o." operator, e.g., Lexicon.o. Rule- > FST (morphological analyzer). Through Foma's composition operator, the lexical transducers are combined with rule transducers to create a comprehensive morphological analyzer capable of bidirectional processing, analyzing both surface forms to extract roots and features, and generating surface forms from underlying representations.

3.3 Creating FSTs that Represent the Morphotactics and the Orthographic Rules of the Ge'ez Verbs—Implementation

Having established the non-concatenative representation, the subsequent step requires merging morphotactic patterns and orthographic rules with the existing finite-state model. The integration of morphotactics and orthographic rules for Ge'ez verbs is accomplished through FSTs using both lexical and rule-based components described in the previous step. A lexical script file (illustrated in Fig. 2) using the lex formalism is created, which operates by declaring labeled lexicons, listing the contents of those lexicons, and specifying the rules that govern how the lexical entries are concatenated to produce the output.

The lexical file combines the declaration of multi-character symbols, the inventory of morphemes (including roots), and the concatenation rules. Together, these elements form the foundation for generating the intermediate output in the morphological analysis process.

To implement the Ge'ez lexicon, it is essential to model Ge'ez verb formation within the lexc formalism. Ge'ez verb formation begins with adding prefixes, when

```
Multichar_Symbols
@U.VERBTYPE.RECIPROCAL@ @U.VERBTYPE.V3P2@ @P.aeInsertion.ae@
+NEG +Verb1 +VIND +RECIP +3PSGMs +3PPLMs +1PSGo +1PPLo

Lexicon Root
PreVerbPrefix;

LEXICON PreVerbPrefix
+NEG:Ai^    VerbPrefix;     // adding prefix Ai to the verb
            VerbPrefix;

LEXICON VerbPrefix
@U.VERBTYPE.V3P2@                V3P2;  //Flag dialects
@U.VERBTYPE.RECIPROCAL@          VReciprocal;

LEXICON V3P2
:y^  VReciprocal;      //adding prefix y to the VReciprocal

LEXICON VReciprocal
+RECIP:t^       Verb;  // adding prefix t to the verb

LEXICON Verb
+Verb1:    VerbType1;

LEXICON VerbType1
qtl             VerbRoot;

LEXICON VerbRoot
+VIND@U.VERBTYPE.V3P2@:@P.aeInsertion.ae@@U.VERBTYPE.V3P2@
VIND3a; //adding prefix using flag dialect @U.VERBTYPE.V3P2@
//@P.aeInsertion.ae@ is used for intercatation of vowels into the
root and is defined in the rule component

LEXICON VIND3a
+3PSGMs:^              VerbSurface;     //no suffix
+3PPLMs:^u             3PPLMs;          // adding suffix u

LEXICON 3PPLMs
#;
:@P.Udeletion.U@            3PPLMo;

LEXICON 3PPLMo
+1PSGo:^uni VerbSurface;          //adding suffix uni

LEXICON VerbSurface
#;
```

Fig. 2 Lexical script file

necessary, to the consonant roots. These prefixes include negation, subject, and derived-verb prefixes, applied in sequence. A Ge'ez verb can either have none of these prefixes or may contain one, two, or all of them, following the order of negation, subject, and derived-verb prefixes. For example, the perfective verb type ቀ·ተ·ለ—qätälä (he killed) does not have a prefix. After adding prefixes to the root, the next step

involves the concatenation of suffixes. Ge'ez language suffixes encompass subject and object suffixes. Similar to prefixes, a Ge'ez verb may or may not require suffixes. The output of the Ge'ez lexical transducer can take on various forms, including: Prefix1 ˆPrefix2 ˆCCC, Prefix1 ˆCCC, Prefix2 ˆCCC ˆSuffix1 ˆSuffix2, CCC + VPER + 3PSM. As a result, the Ge'ez lexical transducer generates an intermediate output comprising Ge'ez consonant roots with affixes and their corresponding feature tags. The Ge'ez lexicon script file consists of the following:

- declaration of the multi-character symbol (using "Multichar Symbols") representing feature tags to mark grammatical information. Table 4 illustrates the multi-character symbols.

A complete list of the multi-character symbols used in the Ge'ez lexical file is found in the lexical file of each verb type. Flag diacritics are used to add the prefixes and rules used to perform vowel intercalation and morphophonological alternation.

- list of prefixes together with rules for adding the prefixes to the roots
- list of roots
- list of suffixes together with rules for concatenating the suffixes to the roots.

Affixes (prefixes and suffixes) and roots are listed, separated by morpheme boundaries (ˆ), with rules for concatenation. Affixes include negation, subject, object, and derived-verb markers. For example, prefixes like negation, subject, and derived-verb prefixes can be listed with rules that add them to the consonant root. Suffixes are also listed with rules for their attachment. Figure 3 demonstrates a graphical representation of a lexicon for simple roots of ቅ-ት-ል-*qtl*, ን-በ-በ-*nbb,* and ን-በ-ር-*nbr*:

The lexical transducer generates an intermediate output that includes Ge'ez consonant roots with affixes and their corresponding feature tags. The output generated by the lexical transducer is a direct consequence of the affixation of morphemes to the root. Intercalation of vowels into the root consonants, as well as orthographic and morphophonological alterations to the verbal stems, are achieved through the use of the alternation rule component.

Table 4 Multi-character symbol

Multi-character symbol	Description
+ VIND	Indicative verb feature tag
+ 1PSG	1st person singular feature tag
@U.VERBTYPE.CAUSITIVE@	Flag dialect for adding causative prefix
@P.eInsertion.e@	A rue component for the insertion of vowels

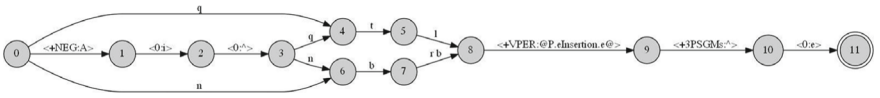


Fig. 3 Lexical FST

For the alternation rule component, a separate Foma script file is created to define alternation rules, which modify the output of the lexical transducer. These rules are compiled into a rule FST. This file includes the following:

- definition of the language alphabet specifying the consonants and vowels used in Ge'ez.
- definition of the flags required for the analysis process.
- labeling of the alternation rules for the rules governing morphophonological alterations.
- labeling of the lexical files needed for reading and labeling lexical script files, such as the one illustrated in Fig. 1.3.
- composition of the lexical FSTs and the rule FST to combine the lexicon FSTs and alternation rule FST.
- compilation the FSTs to produce the final finite-state transducer.

Once all the components are in place, the script compiles the finite-state transducers to generate the final Ge'ez verb morphological analyzer. The composition operator (.o.) in Foma is used to combine the lexical FSTs with the rule FSTs. This merges the lexicon with the morphotactic and orthographic rules. For each Ge'ez verb type, the respective lexical transducer is combined with applicable rule FSTs. The output is a series of intermediate FSTs. These intermediate FSTs are then compiled together, and further merged with rule FSTs that apply to all verb types. This results in the final FST that serves as the Ge'ez morphological analyzer.

3.4 Ge'ex Verb Lexicon—Data Collection and Cleaning

With the foundation of washära classification and finite-state representation established, along with integrated morphotactic and orthographic rule FSTs, the next critical step involves developing evaluation datasets to assess the morphological analyzer's performance. The datasets were created through a combination of manual extraction from religious texts, data pre-processing, expert linguistic annotation, and root extraction, ensuring that the datasets were both comprehensive and accurate.

To create the datasets, Ge'ez language scholars, affiliated with the Ethiopian Orthodox Tewahido Church, extracted verbs from two primary sources: the Ge'ez New Testament Bible (specifically the books of Matthew, Luke, Mark, and John) and the Ge'ez prayer book "ውዳሴ ማርያም"—"wudase Maryam." A total of 1519 verbs were initially collected from these sources. The initial dataset was cleaned by removing the letters "ወ" (wä) and "ዘ" (zä) to ensure consistency. Repeated words were removed to ensure only unique verb forms were included. In addition, words that belonged to other parts of speech (POS) were also removed from the dataset. After cleaning, 1365 unique verbs were identified for the evaluation dataset. Table 5 shows the total number of words extracted for each type of verb and the number of roots in each verb type.

Table 5 Total number of words extracted for each type of verb and the number of roots in each verb type

Verb type	Number of verbs	Total number of roots	Roots in test data	Percentage (%)
ቀተለ—qätälä (Verb 1)	595	215	178	43.59
ቀደሰ—qäddäsä (Verb 2)	146	62	44	10.70
ገበረ—gäbrä, (Verb 3)	308	76	62	22.56
አእመረ—äəmärä (Verb 4)	56	20	15	4.10
ባረከ—baräkä (Verb 5)	14	8	4	1.03
ሤመ—śəmä (Verb 6)	24	11	10	1.76
በህለ—bəhlä (Verb 7)	76	19	15	5.57
ቆመ—qomä, (Verb 8)	100	22	17	7.33
ቤለ—belä (Verb 9)	19	1	1	1.39
Irregular verbs (Verb 10)	27	5	3	1.98
Total	1356	439	349	100

For the distribution of the verb types, the cleaned dataset included a variety of Ge'ez verb types, with 43.59% from the ቀተለ—qätälä verb type, 10.70% from ቀደሰ—qäddäsä, 22.56% from ገበረ—gäbrä, and smaller percentages from other verb categories including አእመረ—äəmärä, ባረከ—baräkä, ሤመ—śemä, በህለ—bəhlä, ቆመ—qomä, as well as ቤለ—belä and irregular verbs. In terms of the gold-standard annotation, Ge'ez language experts manually annotated the collected verbs with their correct morphological analysis where each word was given a single, context-specific analysis. This annotation included the root of the verb, verb type, tense, mood, person, and gender. The annotators also included structural information related to affixes and their meanings. For example, a gloss explaining the role of a prefix or suffix. For instance, the analysis of the Ge'ez verb "ይነብቡኒ"—"yənbəbuni" included the root, verb type, mood, subject, and object markers.

From the gold-standard data, 349 unique Ge'ez root verbs were extracted. An additional 90 Ge'ez roots were organized and added by Ge'ez experts to expand the lexicon. The lexicon created contains a total of 439 Ge'ez roots. The annotated dataset served as the gold standard against which the accuracy of the Ge'ez morphological analyzer was evaluated.

3.5 Evaluation

The finite-state morphological analyzer for Ge'ez verbs was evaluated using the gold-standard dataset. The accuracy of the Ge'ez morphological analyzer was assessed by comparing its output to the gold-standard data. In the initial stages of the evaluation, there were varying results, some words were correctly analyzed, others produced multiple outputs, some were incorrect, and some yielded no analysis. The unexpected outputs were attributed to inaccuracies in verb transliteration, unavailability of roots in the Ge'ez lexicon, and incorrect placement of roots within their respective head verbs. To address these issues, corrections were made in consultation with Ge'ez experts. These corrections included ensuring that the roots were correctly positioned within their head verbs and incorporating all roots found in the evaluation dataset into the Ge'ez lexicon. These measures significantly improved the accuracy of the Ge'ez morphological analyzer.

After implementing these corrective measures, the Ge'ez morphological analyzer was re-evaluated. The performance of the morphological analyzer was measured using the following metrics accuracy (A) on (1) and precision (P) on (2).

$$A = (\mathrm{tp} + \mathrm{tn})/(\mathrm{tp} + \mathrm{tn} + \mathrm{fp} + \mathrm{fn}) \tag{1}$$

True positives (tp) and true negatives (tn) represent cases where the analyzer correctly identified and analyzed a verb, matching the gold-standard annotation. True negatives are not relevant in this specific context since the evaluation dataset consists exclusively of verbs. False positives (fp) occur when the analyzer produces incorrect or additional analyses that don't match the gold standard, while false negatives (fn) represent cases where the analyzer failed to produce any analysis for a verb. The overall accuracy of the morphological analyzer was measured by the total number of correctly analyzed verbs divided by the total number of verbs analyzed.

$$P = \mathrm{tp}/(\mathrm{tp} + \mathrm{fp}) \tag{2}$$

Precision measures the proportion of accurately analyzed verbs over the proportion of the analysis returned by the morphological analyzer. The overall precision of the morphological analyzer was measured by the total number of correctly analyzed verbs divided by the total number of analysis outputs by the morphological analyzer. Table 6 provides the evaluation results of the morphological analyzer.

Out of the 1328 correctly analyzed verbs (unique verbs), the analyzer generated an additional 327 possible analyses. The precision of the Ge'ez morphological analyzer was calculated at 80.24%. This figure considers that the analyzer provided all possible analyses for each verb, irrespective of context.

Table 6 Evaluation results

Verb type	Number of verbs	Accuracy (%)	Precision (%)	Not analyzed (%)
ቀተለ—qätälä (Verb 1)	595	96.97	80.03	3.03
ቀደሰ—qäddäsä (Verb 2)	146	97.26	83.53	2.74
ገበረ—gäbrä, (Verb 3)	308	97.40	81.52	2.60
አእመረ—äəmärä (Verb 4)	56	96.43	85.71	3.57
ባረከ—baräkä (Verb 5)	14	100.00	58.33	0.00
ሤመ—sɨ́mä (Verb 6)	24	100.00	80.00	0.00
በህለ—bəhlä (Verb 7)	76	96.05	81.11	3.95
ቆመ—qomä, (Verb 8)	100	98.00	81.67	2.00
ቤለ—belä (Verb 9)	19	100.00	48.72	0.00
Irregular verbs (Verb 10)	27	100.00	90.00	0.00
Total	1356	97.29	80.24	2.71

4 Discussion

The development and evaluation of the Ge'ez morphological analyzer revealed several key findings and challenges. A significant issue emerged regarding morphophonological alternations caused by specific letters (አ, ዕ, ሀ, ኀ, ሐ, ሠ, and ደ). While rules were implemented to handle these alternations, inconsistencies were observed among verbs sharing the same head verbs, particularly in the analysis of ቤለ (belä) verbs, which showed dual categorization with ኀበህለ (bəhlä) verbs due to overlapping inflections.

The analyzer's performance significantly improved through iterative refinement. Initial testing produced 72% accuracy, but through expert review of manual annotations, correction of transliteration errors, and thorough verification of the Ge'ez lexicon, the accuracy increased to 97.2%. The precision rate of 80.24% reflects the analyzer's comprehensive approach to generating all possible analyses, contrasting with the context-specific single analyses in the gold-standard dataset.

A small percentage (2.71%) of the test set received no analysis, primarily due to the presence of the aforementioned challenging letters. Of the 37 unanalyzed words, 34 contained these letters in their roots, indicating challenges in handling certain morphophonological alternations. Comparative analysis with previous research demonstrates significant improvement. The system substantially outperforms [13]' rule-based approach, which achieved 73.98% accuracy but was limited to the ቀተለ (qätälä) verb category. Similarly, it surpasses [12]' data-driven approach, which reached maximum accuracy of 60.3%. The current analyzer's comprehensive coverage of all verb types, including irregular verbs, and its high accuracy of 97.29% represent a substantial improvement in Ge'ez morphological analysis.

5 Conclusion

This research has successfully developed a comprehensive finite-state morphological analyzer for Ge'ez verbs with an accuracy of 97.29% and a precision of 80.24%. The study adopted the washära classification system, effectively handling Ge'ez's complex verb morphology. The study's contribution includes the demonstration of the effectiveness of FST in handling non-concatenative morphology in Semitic languages, the development of a morphological analyzer that successfully processes all verb types, including irregular verbs, improving on the limitations of previous approaches, the bidirectional functionality of the analyzer that enables analysis and generation of verb forms and the creation of a gold-standard dataset with 1365 manually annotated verbs that is a valuable resource for future research in Ge'ez computational linguistics given the scarcity of linguistic resources for the language. Future work will focus on the development of specialized handling mechanisms for problematic letters ጵ-ə, ዕ-'ə, ሀ-h, ኀ-ḫ, ሐ-ḥ, ሠ-w, and ፀ-y that currently account for most unanalyzed cases and expanding the analyzer's coverage to other parts of speech.

References

1. Akelew H (2023) Morpheme based Bi-directional machine translation the case of ge'ez to Tigrigna. ST. Mary's University
2. Abebe G (2021) Ge'ez to English machine translation using RNN
3. Alemu AA, Fante KA (2021) A corpus-based word sense disambiguation for geez language. Ethiop J Sci Sustain Dev 8(1):94–104
4. Gebeyehu S, Wolde W, Shibeshi ZS (2023) Information extraction model from Ge'ez texts. Indones J Electr Eng Comput Sci 30(2):787–795
5. Petrácek K (1960) Edward ullendorf," the semitic languages of Ethiopia (Book Review). Archív Orientální 28(1):161–164
6. Hetzron R, Kaye AS, Zuckermann GA (2018) Semitic languages. In: The world's major languages. Routledge, pp 568–576
7. Appleyard D (2015) Ethiopian Semitic and Cushitic. Ancient contact features in Ge 'ez and Amharic. In: Semitic languages in contact. Brill, pp 16–32
8. Weninger S (1993) Ge'ez, vol 1. Lincom Europa
9. Butts AM (2019) GəʕəZ (Classical Ethiopic). The Semitic languages. Routledge, pp 117–144
10. Gidey GG, Teklehaymanot HK, Atsbha GM (2024) Morphological synthesizer for Ge'ez language: addressing morphological complexity and resource limitations. In: Proceedings of the fifth workshop on resources for African indigenous languages@ LREC-COLING 2024
11. Firdyiwek Y, Yaqob D (1997) The system for Ethiopic representation in ASCII. citeseer.ist.psu.edu/56365.html
12. Abate Y, Assabie Y (2014) Morphological analysis of ge'ez verbs using memory based learning. Addis Ababa University, Addis Ababa
13. Desta BW (2010) Design and implementation of automatic morphological analyzer for Ge'ez verbs. Addis Ababa University
14. Asmare HS, Yibre AM (2023) Ge'ez syntax error detection using deep learning approaches. In: Pan African conference on artificial intelligence. Springer
15. Yan Y et al (2023) Survey on applications of algebraic state space theory of logical systems to finite state machines. Sci China Inf Sci 66(1):111201

16. Cubel, E., et al. *Finite-state models for computer assisted translation.* in *ECAI*. 2004.
17. Hori T, Nakamura A (2022) Speech recognition algorithms using weighted finite-state transducers. Springer Nature
18. Beesley KR, Karttunen L (2003) Finite-state morphology: Xerox tools and techniques. CSLI, Stanford, pp 359–375
19. Yona S, Wintner S (2005) A finite-state morphological grammar of Hebrew. In: Proceedings of the ACL workshop on computational approaches to Semitic languages
20. Amsalu S, Gibbon D (2005) Finite state morphology of Amharic. In: 5th Recent advances in natural language processing, pp 47–51
21. Baxi J, Bhatt B (2024) Recent advancements in computational morphology: a comprehensive survey. arXiv preprint arXiv:2406.05424
22. Pugh R, Tyers F (2023) A finite-state morphological analyser for highland Puebla Nahuatl. In: Proceedings of the workshop on natural language processing for indigenous languages of the Americas (AmericasNLP)
23. Christopherson CS (2023) A finite-state morphological analyzer for Q'eqchi'using Helsinki finite-state technology (HFST) and the Giellatekno infrastructure
24. Kayabaş A et al (2019) TRMOR: a finite-state-based morphological analyzer for Turkish. Turk J Electr Eng Comput Sci 27(5):3837–3851
25. Rahi R et al (2020) A finite state transducer based morphological analyzer of maithili language. arXiv preprint arXiv:2003.00234
26. Alblwi A et al (2023) A deterministic finite-state morphological analyzer for Urdu nominal system. Eng Technol Appl Sci Res 13(3):11026–11031
27. Gebremeskel HG, Chong F, Heyan H (2023) Unlock Tigrigna NLP: design and development of morphological analyzer for Tigrigna verbs using hybrid approach. Available at SSRN 4627238
28. Kassahun A (2012) Ye'kolo-temari: children's perspectives on education, mobility, social life and livelihood in the Ethiopian orthodox church (EOC) traditional school in Dangila. Norges teknisk-naturvitenskapelige universitet, Fakultet for ….
29. Hulden M (2009) Foma: a finite-state compiler and library. In: Proceedings of the demonstrations session at EACL 2009
30. Hulden M (2011) Morphological analysis tutorial: a self-contained tutorial for building morphological analyzers

Development of a Virtual Reality Training Program: Integrating FDS Simulation and Performance Optimization with Unreal Engine on Heterogeneous Hardware

D. Kim

Abstract To develop a more realistic virtual reality firefighting training platform, this study utilizes fire spread data, including smoke and heat, obtained from numerical simulations using the Fire Dynamics Simulator (FDS). FDS employs MPI and OpenMP for large-scale fire simulations by dividing computational domains into subdomains, with OpenACC applied to support heterogeneous hardware architectures. Performance tests using CSIRO Grassland Fires demonstrated a 1.89 × speedup with combined CPU-GPU computation and a 21 × speedup with 1 GPU and 16 CPUs, validating enhanced fire analysis capabilities. Additionally, existing VR engines were improved by integrating WFDS data into Unreal Engine for realistic smoke and heat visualization using FGA files. The program dynamically visualizes flame and smoke movements based on wind speed and direction, achieving a 100% match between WFDS data and Unreal Engine output, with combustion stages rendered in real-time through mass-based material updates.

Keywords Virtual reality (VR) · 3D visualization · Unreal Engine · Virtual training · Wildland-Urban Interface Fire Dynamics Simulator (WFDS)

D. Kim (✉)
Jeonju University, Jeonjusi, Jeonbuk State, Republic of Korea
e-mail: 72donghyunkim@jj.ac.kr

IIASA(International Institute for Applied Systems Analysis), Laxenbug, Austria

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_16

1 Introduction

1.1 A Subsection Sample

Forest fires continue to pose significant risks to firefighter safety, causing numerous fatalities and injuries globally each year. To address these challenges, organizations have implemented various forest fire education and training programs aimed at improving firefighters' understanding of wildfire behavior and enhancing their situational response capabilities [6]. However, traditional theoretical instruction and laboratory-scale training methods often fail to replicate the dynamic and complex nature of real-world forest fire environments, limiting their effectiveness [1].

Recent advancements in virtual reality (VR) technology have opened new avenues for the development of immersive fire education and training simulators. These simulators, often combined with game-based training programs, aim to provide realistic experiences for firefighters. However, current systems typically rely on simplistic image-based depictions of flames and smoke, which lack the integration of chemical and physical analyses crucial for simulating realistic fire dynamics [3]. This limitation underscores the need for advanced methodologies to create more effective training tools [13, 16].

This study aims to address these shortcomings by developing a 3D VR-based education and training authoring tool that integrates chemical species and physical analyses. Utilizing data from the Wildland-Urban Interface Fire Dynamics Simulator (WFDS), the proposed methodology enables the visualization of forest fire flames and smoke within the Unreal Engine (UE), based on temperature and smoke distribution data. This integration enhances the realism of VR training simulations, providing firefighters with more accurate and immersive learning environments.

The research also explores computational advancements to improve simulation performance. With single-core CPU improvements reaching saturation, multi-core architectures and accelerators, such as GPUs, have emerged as essential solutions for overcoming hardware limitations [4, 15]. OpenACC, a parallelization framework optimized for heterogeneous hardware architectures, is utilized in this study to accelerate the Fire Dynamics Simulator (FDS) [14]. By leveraging parallel processing, the framework significantly reduces computation times, enhancing the efficiency of fire spread analyses.

The Fire Dynamics Simulator (FDS), developed by the National Institute of Standards and Technology (NIST), employs computational fluid dynamics to analyze low-velocity, heat-driven fluid flows from fires. Although FDS is a robust tool for modeling fire dynamics, its computational demands increase substantially for detailed simulations [5, 8, 9]. To address this, FDS incorporates parallelization techniques such as the Message Passing Interface (MPI) and Open Multi-Processing (OpenMP) libraries, which distribute computational tasks across multiple processors [10]. In this study, OpenACC was partially integrated into FDS to further optimize its performance on heterogeneous hardware [7, 12, 16].

The findings of this research demonstrate the feasibility of realistic fire visualization in VR environments. By combining physical analyses with VR technology, the study lays the groundwork for developing immersive and efficient training tools that contribute to safer and more effective firefighter education.

2 Methods

2.1 WFDS Analysis

WFDS was developed to analyze fire spread in open spaces adjacent to forests by applying topography and vegetation fuel by extending FDS developed for structural fire analysis. In this study, the X-axis, Y-axis, and Z-axis were set to 200 m, 160 m, and 50 m, respectively, and the cell spacing was set to 2 m, resulting in a total of 200,000 cells were configured as 100 cells on the X-axis, 80 cells on the Y-axis, and 25 cells on the Z-axis.

For forest fuel, a combustion material, combustion properties were applied to Pinus densiflora with an average height of 17 m and 400 trees/ha, and Cone type was applied for the grid configuration.

2.2 Visualization of VR Contents Forest Fires Spread Simulation

Result values analyzed with WFDS are stored as Plot3Ddata files, velocity vectors, and physical quantities for fire analysis, and then extracted as vector files FGA files with a cycle of N(3 ~ 5) seconds so that they can be imported into 3D Unreal Engine. The FGA file contains data values of vectors and physical quantities in the analysis result. FGA file imported from 3D UE is created as a vector field, and Boundary values such as position, rotation, stile, intensity, and tightness are set. With the generated vector field, the particles are completed by applying the fire and smoke effect used in VR 3D contents. For visualization of forest fire diffusion and forest fuel combustion, first, for visualization of flames and smoke, FGA temperature and smoke (CO, CO_2) concentration data are expressed through particle contrast and lighting effect functions. Second, image change due to forest fuel combustion used the Mesh Decal function as a visualization method with image values created according to the mass reduction rate. Mesh Decal uses the mixed function of semi-transparent Blend Mode and Deffered Decal by updating GBuffer and DBuffer after rendering 3D object surface geometry. In addition, it was configured so that external results such as changes in heat flow of each grid value, wind speed, and wind direction in the UE can be reflected in the impact effect. This is useful for expressing the movement and concentration of smoke in each grid cell differently.

2.3 *Parallel Computing Method*

MPI and OpenMP are standard methods for performing parallel computation in distributed memory system and shared memory system environment. Parallel computation has been supported since FDS version 5.4, and two parallel processing methods are provided, OpenMP and MPI. These two methods can be simultaneously applied to FDS according to the hardware configuration. The FDS User's Guide, it was reported that the maximum speed improvement of OpenMP is approximately doubled, and that when MPI and OpenMP are used together, most of the reduction in analysis time is achieved by MPI (Fig. 1).

2.3.1 MPI

MPI is a standard that describes information exchange in distributed and parallel processing. The Message Passing method is an operation model in which data to be exchanged between processors is exchanged using a Message Passing function. MPI defines the standard of the Message Passing Library, which is a collection of these functions, and several MPI libraries have been developed accordingly. Subroutines that require parallelization in sequence code are constructed in parallel through appropriate parallelization techniques.

2.3.2 OpenMP

OpenMP is an application programming interface that supports shared memory multiprocessing of programs in programming languages such as C, C++, and Fortran on various computer platforms. OpenMP provides users with a simple and flexible interface for developing a variety of parallel applications from desktops to supercomputers, and is easily extensible. And by using OpenMP and MPI, the application program can also be built as a Hybrid OpenMP/MPI model.

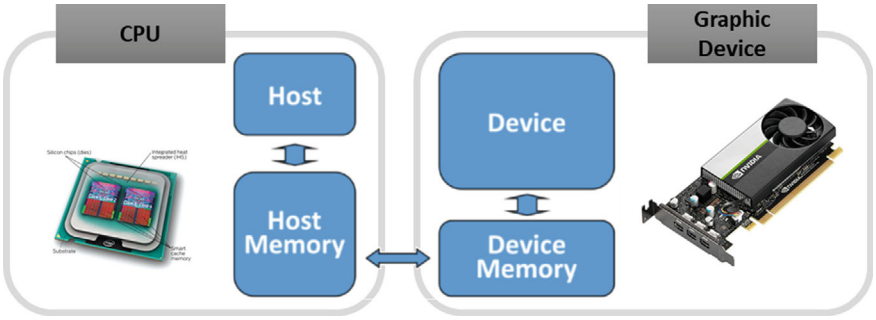


Fig. 1 OpenACC's abstract accelerator model

3 Study Result

3.1 *Visualization of VR Contents Forest Fires Spread Simulation*

As a result of expressing the file extracted from the WFDS analysis result as an FGA file with the Unreal Engine, the vector and physical quantity data values such as heat and smoke of each 200,000 grid cells expressed 100% the same value without error even with time series changes. For the visualization of the flame temperature and smoke values for the grid cell, the color and smoke concentration of the flame were adjusted and expressed through the particle contrast and lighting effect on the grid. The results are shown in Figs. 2 and 3. The accuracy of the visualization image for the flame intensity and smoke concentration was implemented so that it could be adjusted through the particle lighting function.

In addition, for the image change according to combustion of the 3D combustion material object data, the Mesh Deca technique was applied to create a material image, and then a method was used to replace the image by mass reduction rate. Figure 4 is a screen visualized in 3D UE according to flame temperature, smoke, and combustion of trees.

Figure 5 shows the simulation results of forest fire spread in Unreal 3D engine using the forest fire diffusion formula of Alexandridis et al. [2]. There were 2400 grids used for the terrain, and the prediction of wildfire volcanoes was expressed by applying the following conditions to each variable value such as wind, inclination, ambient temperature, and type of fuel material.

- Wind speed: 4.6 m/s.

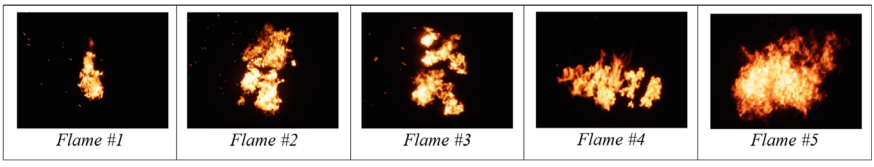


Fig. 2 Forest fire flame visualization image results in 3D UE

Fig. 3 Forest fire smoke visualization image results in 3D UE

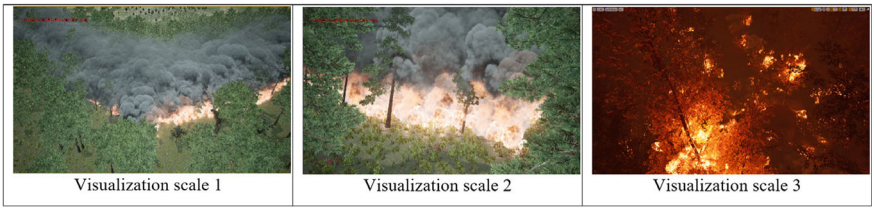


Fig. 4 Forest fire visualization image results to flame temperature, smoke and tree in 3D UE

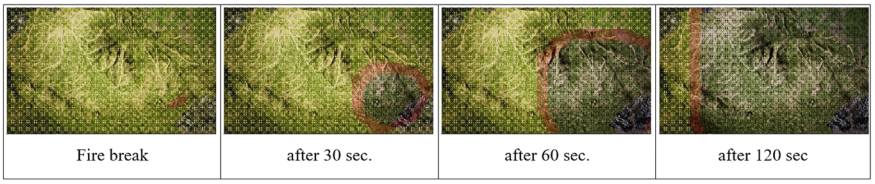


Fig. 5 Result of forest fire spread analysis using 3D UE

- Wind direction: southwest wind.
- Temperature: 32 °C
- FMC (Fuel Moisture Contents): 6.3%
- Cell size: 5 cm^2.

3.2 Computing Performance

The computational performance is evaluated through the verification case of FDS, CSIRO Grassland Fires (Fig. 6). As hardware for evaluation, a personal computer composed of dual Xeon 2678-V3 and GeForce GTX 1070 was used. FDS source code applies OpenACC using PGI Fortran as a compiler in Linux environment. In the FDS source code, "!$ACC PARALLEL LOOP" of OpenACC was partially applied to

"PRESSURE_ITERATION_LOOP," the part that determines convergence during calculation (Table 1).

In the case of MPI, when 16 CPUs are used compared to one CPU, the analysis time is 12 times faster from 109 to 9 h. On the other hand, when one CPU and one GPU are used together by applying OpenACC, the calculation time is 1.89 times faster than the result using one CPU. In the case of using 1 GPU and 16 CPUs (MPP + OpenACC), the analysis time is 5 h, which is about 21 times faster.

Table 2 shows the results of measuring the parallel performance of MPI+OpenACC according to the increase in the size of the computation area. When parallelization was performed with MPI, the performance improvement was evident as the number of CPUs and the size of the problem increased. On the other hand,

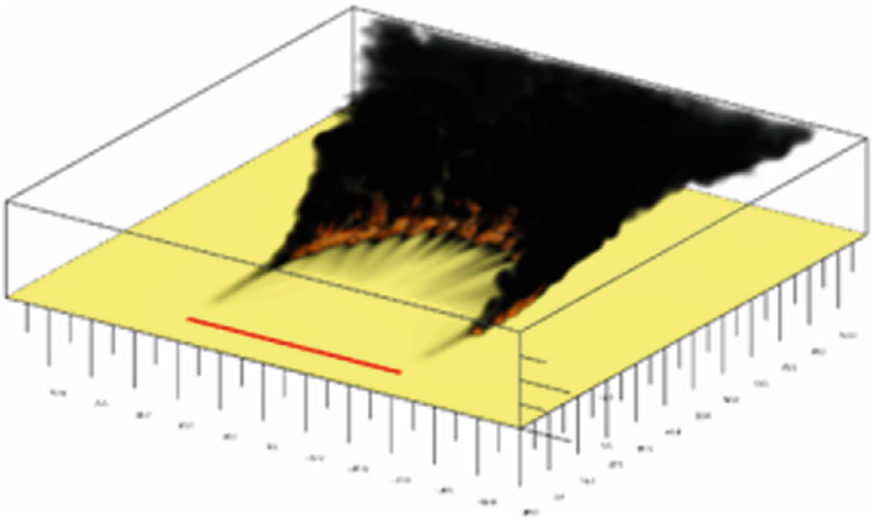


Fig. 6 WFDS result of grassland fires in CSIRO Australia

Table 1 Computing performance

MPI	Analysis time (h)	OpenACC with MPI	Analysis time (h)
1 CPU	109	1CPU+1GPU	57
2 CPU	59	2CPU+1GPU	32
4 CPU	31	4CPU+1GPU	16
8 CPU	17	8CPU+1GPU	9
16 CPU	9	16CPU+1GPU	5

in the case of MPI+OpenACC, the performance of MPI+OpenACC with 2 CPUs or more is lower than that of MPI. It can be seen that the amount of CPU allocated during MPI parallelization is also allocated to the GPU as much as the number of CPUs, which increases the load and decreases performance. In addition, if the size of the calculation area is about 100 million and 8 CPUs are used, the memory performance of the GPU, GTX 1070, is exceeded, and calculation cannot be performed. However, when one or two CPUs are used, it can be seen that the performance of MPI+OpenACC is significantly superior to that of MPI depending on the size of the calculation area.

Table 2 2D heat conduction speedup with problem size and parallel method

MPI	1024 × 1024 (@1 million)		2048 × 2048 (@4 million)		10,240 × 10,240 (@100 million)	
Parallel Method	MPI	MPI + OpenACC	MPI	MPI + OpenACC	MPI	MPI + OpenACC
1 CPU	1	1.48	1	2.93	1	4.96
2 CPU	1.96	1.41	1.99	2.82	2.63	4.69
4 CPU	3.70	0.85	4.15	1.70	4.06	2.88
8 CPU	7.21	0.59	6.46	1.08	8.02	*
16 CPU	12.12	0.31	16.84	0.74	16.74	*

*Out of GPU memory

4 Conclusions

This study on the visualization program of VR Unreal Engine (UE) for virtual reality forest fire training yielded the following conclusions:

1. The SMV data and UE visualization program showed a 100% agreement in grid values for temperature and smoke, based on WFDS calculation results.
2. A visual comparison of flame and smoke concentration revealed differences in brightness and contrast in displayed images.
3. By adjusting the color and smoke concentration of flames through particle contrast and lighting effects in the 3D UE, the visual effects were refined to closely resemble SMV visualization data.
4. A methodology for accurately visualizing forest fire properties for realistic training experiences was proposed. However, this study did not incorporate real-time weather condition changes into the visualization process.
5. OpenACC, a parallelization technique for heterogeneous hardware architectures, was partially applied to FDS. Computational performance was evaluated using CSIRO Grassland Fires, showing a 1.89 × speedup with combined CPU-GPU processing and a 21 × speedup with 1 GPU and 16 CPUs, reducing analysis time to 5 h. This demonstrates the potential of hardware parallelization to significantly reduce computation times for forest fire spread analysis using personal computers.

Future development of fire training content in virtual environments, including educational and evaluation scenarios, based on these findings, is expected to provide a safer and more effective learning tool for forest firefighters.

Acknowledgements This paper was written with support from the National Fire Agency's research project, a study on the Establishment of the Advanced Virtual Reality Firefighting Training System (20008389).

References

1. Alexander ME, Thomas DA (2006) Prescribed fire case studies, decision aids, and planning guides. Fire Manag Today 66(1):5–20
2. Alexandridis A, Vakalis D, Siettos CI, Bafas GV (2008) A cellular automata model for forest fire spread prediction: the case of the wildfire that swept through Spetses Island in 1990. Appl Math Comput 204:191–201. https://doi.org/10.1016/J.AMC
3. Calore E, Gabbana A, Schifano SF, Tripiccione R, Fyta M (2016) Massively parallel lattice-Boltzmann codes on large GPU clusters. Parallel Comput 58:1–24
4. Hennessy JL, Patterson DA (2019) Computer architecture: a quantitative approach, 6th edn. Morgan Kaufmann
5. Hostikka S, McGrattan K (2001) Large eddy simulation of industrial-scale fires. Int J Comput Fluid Dyn 15(2):127–146
6. Jolly WM, Cochrane MA, Freeborn PH, Holden ZA, Brown TJ, Williamson GJ, Bowman DM (2015) Climate-induced variations in global wildfire danger from 1979 to 2013. Nat Commun 6:7537
7. Kraus M, Poinsot T (2016) Accelerating CFD simulations of combustion processes using OpenACC. Comput Fluids 136:152–162
8. McGrattan K et al (2013) Fire dynamics simulator, technical reference guide. NIST Special Publication 1018–5, National Institute of Standards and Technology
9. McGrattan K et al (2013) Fire dynamics simulator user's guide. NIST Special Publication 1019–5, National Institute of Standards and Technology
10. McGrattan K, Forney G (2004) Fire dynamics simulator (Version 4)–user's guide. NIST Special Publication 1019, National Institute of Standards and Technology
11. OpenACC-Standard.org (n.d.) OpenACC for programmers: concepts and strategies. Retrieved from https://www.openacc.org/. Accessed on 05 Sep 2024
12. Putnam T (1995) Findings from the wildland firefighters human factors workshop. USDA Forest Service
13. Silver D et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489
14. Sutter H (2005) The free lunch is over: a fundamental turn toward concurrency in software. Dr. Dobb's J 30(3)
15. Wienke S, Springer P, Terboven C, Mey D (2012) OpenACC—first experiences with real-world applications. In: European conference on parallel processing. Springer, pp 859–870
16. Williams-Bell FM, Kapralos B, Hogue A, Murphy B, Weckman EJ (2015) Using serious games and virtual simulation for training in the fire service: a review. Fire Technol 51:553–584

Smart IoT Water Curtain System for Protecting Wildland-Urban Interface (WUI) Village from Forest Fires

Donghyun Kim

Abstract This study introduces the Crown Water Spray Equipment for Forest Fires Protection (CWSEFFP), a novel system designed to prevent the spread of large-scale forest fires and protect village communities, cultural heritage sites, and forest recreational facilities. The devastating 2005 Yangyang forest fire in South Korea, which destroyed Naksan Temple and 22 state-designated cultural properties, highlighted the urgent need for enhanced fire prevention systems. CWSEFFP I and II are designed to spray water over areas of 200 m × 80 m and 2000 m × 80 m, respectively, using strategically installed large nozzles. Since 2012, CWSEFFP I has been implemented in 240 locations across South Korea, with plans to deploy CWSEFFP II in 2024. This study details the design, functionality, and technical specifications of CWSEFFP, emphasizing its role in mitigating damage in the Wildland-Urban Interface (WUI) areas. The system exemplifies an aggressive approach to forest fire prevention, addressing the increasing frequency and intensity of fires globally. It not only suppresses active fires but also prevents their spread, thereby offering a robust solution for safeguarding vulnerable communities.

Keywords Forest fires · Water supply system · Wildland-Urban Interface (WUI) · Forest fire protection · Smart IoT fire suppression

D. Kim (✉)
Jeonju University, Jeonjusi, Jeonbuk State, Republic of Korea
e-mail: 72donghyunkim@jj.ac.kr

IIASA (International Institute for Applied Systems Analysis), Laxenbug, Austria

X. Yang et al. (eds.), *Proceedings of Tenth International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 1440,
https://doi.org/10.1007/978-981-96-9709-0_17

1 Introduction

1.1 A Subsection Sample

Forest fires pose great threats not only to forests, but also to villages and facilities adjacent to forests and to houses within forests. Many reports warn that the occurrence and severity of forest fires will increase due to climate change, and cases of occurrence of forest fires supporting the foregoing are appearing around the world. The occurrence of forest fires as such not only affects even areas that have not been invaded by humans, such as Alaska, the frozen soil regions of Siberia, and the Amazon, but also poses threats to many cities where people are living in areas adjacent to forests. In cases of occurrence of forest fires in South Korea, the maximum forest fire spread rate was found to be 8.8 km/h and the distance of forest fire spread by flying sparks was found to be 2 km in an investigation. In the case of a forest fire in Portugal in 2017, the maximum forest fire spread rate was at least 9 km/h and the distance of forest fire spread by flying sparks was at least 3 km. In the United States, a maximum fire spread rate of 13 km/hr and a distance of forest fire spread by flying sparks of 3.2 km were observed in the Wallace Greece forest fire in 2003.

The increasing frequency and intensity of forest fires, driven by the twenty-first century climate crisis, demand proactive measures to protect Wildland-Urban Interface (WUI) areas. Transforming WUI spaces into fire-resistant communities requires strategies to reduce combustible materials and prevent ignition. Approaches like FireSmart and FireWise offer guidance by designating zones for fuel removal and installing firefighting equipment. These methods include providing water supplies and spraying systems to mitigate fire spread, ensuring safer environments for residents in vulnerable areas.

Roof sprinklers installed on houses adjacent to forests, as proposed by programs like FireSmart and FireWise, are limited in their ability to prevent the spread of fires across entire villages or forest-adjacent areas. While effective in protecting individual homes, they fall short in addressing wide-area fire spread. To enhance the water screening capacity of such systems, this study introduces the Crown Water Spray Equipment for Forest Fires Protection (CWSEFFP). Designed for more aggressive suppression of large-scale forest fires, CWSEFFP aims to protect Wildland-Urban Interface (WUI) communities more comprehensively. Previous experiments demonstrated that a minimum water application rate of 100 $m\ell/m^2$ is required to extinguish or prevent the spread of surface fire flames fueled by pine forest litter under standard conditions. However, the necessary watering amount can vary significantly depending on factors such as fire intensity and weather conditions. This study highlights CWSEFFP's potential to address these challenges and ensure broader fire protection in vulnerable areas.

To safeguard Wildland-Urban Interface (WUI) communities from increasingly severe and frequent forest fires, it is essential to implement powerful wide-area water screen systems. These facilities can block radiant heat and prevent the spread of forest

fires caused by flying sparks, thereby protecting entire villages. This urgency underscores the need for aggressive forest fire prevention measures. In South Korea, the 2005 Yangyang forest fire devastated Naksan Temple, a thousand-year-old cultural heritage site, along with 22 state-designated cultural properties. In response, efforts to protect temples and recreational facilities in forested areas led to the development of the Crown Water Spray Equipment for Forest Fires Protection (CWSEFFP). Designed by Kim (2013), the CWSEFFP I system utilizes large, 360-degree rotating nozzles installed above tree height to spray water over an area of approximately 200 m x 80m. Since its initial deployment in 2012, engineering advancements have culminated in CWSEFFP II, capable of covering up to 2000 m $\times$ 80 m with a single pressurized water supply device. This study introduces the CWSEFFP systems, detailing their development, functionality, and potential for large-scale forest fire prevention.

2 Methods

2.1 Case Study CWSEFFP I

The Crown Water Spray Equipment for Forest Fires Protection Version I (CWSEFFP I) is designed to spray water from a height exceeding tree levels, providing protection for cultural heritage sites, forest recreational facilities, and residential properties in small villages located near forested areas. This system is intended both to prevent the spread of forest fires and to extinguish them when they occur. The CWSEFFP I comprises several key components, including a pressurized water supply device, piping systems with water supply pipes and selection valves, towers, large spray nozzles, and a remotely operated control unit for pump activation (Fig. 1). During operation, the pressurized water supply device initiates the flow of water to the spray nozzles, which discharge water with a spray radius of approximately 42.5 ± 7.5 m per nozzle (Fig. 2). The pump and nozzle configurations can be tailored to site-specific conditions, allowing for adjustments to both the water flow rate and spray distance to optimize fire suppression effectiveness.

As illustrated in the facility schematic diagram in Fig. 2, the CWSEFFP I supports three methods of water spraying: (a) deploying water through three crown water spray towers, (b) utilizing a fire hose connected to an outdoor fire hydrant, and (c) employing mobile water spray equipment for flexible application. Figure 3 depicts the operation of two facilities equipped with the CWSEFFP I system in real-world settings. The figure illustrates the system in action, demonstrating its effectiveness in protecting forests and forest-adjacent facilities from forest fires. The CWSEFFP I allows for adjustments to the spray radius and water flow rate by customizing the performance parameters of each nozzle.

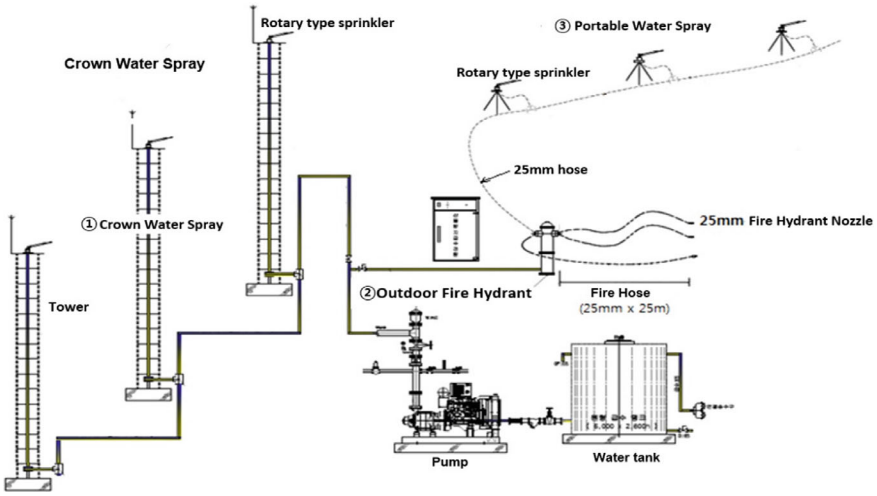


Fig. 1 Facility schematic for CWSEFFP I

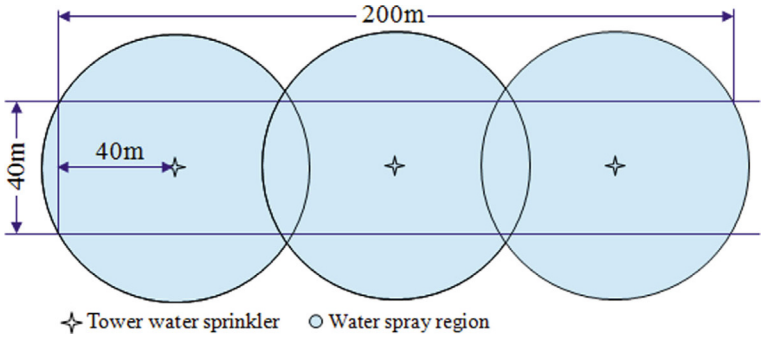


Fig. 2 Water spray boundary from water sprinkler tower system

Fig. 3 CWSEFFP I operation scene that has been installed in S. Korea

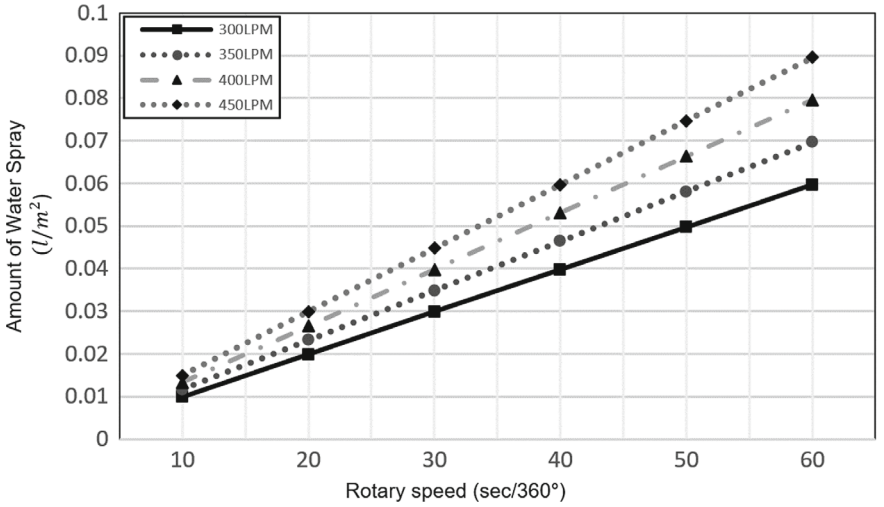


Fig. 4 Water spray per unit area according to nozzle rotation speed

2.2 Nozzle Spray Radius and Water Flow Rate

Figure 4 shows the water distribution per unit area based on the nozzle's rotational speed per minute when operating at flow rates of 300 LPM, 350 LPM, and 400 LPM, with a spray radius of 40 m. The appropriate water application rate from the CWSEFFP I should be determined based on the fire intensity of the surrounding combustible materials. For instance, grasslands with lower fire intensity require a reduced water application rate compared to pine forests, which exhibit higher fire intensity.

2.3 CWSEFFP II Design

The CWSEFFP I utilizes three large spray nozzles mounted on tower structures to deliver water over an area of approximately 200 m × 80 m, effectively protecting forest-adjacent facilities such as recreational sites and significant cultural heritage temples in South Korea. However, its coverage is limited when addressing larger residential areas near forests or extensive forested regions requiring protection from major fire spread. In contrast, the CWSEFFP II employs multiple water spray systems with differential spraying technology, significantly expanding the coverage area of the original CWSEFFP I. By integrating smart IoT control technology, the CWSEFFP II enhances its capabilities to protect forest-adjacent areas, increasing the effective spraying range tenfold from 200 m to 2000 m. This development addresses the global challenge of prolonged forest fires that persist for several weeks and cause

extensive damage to villages and facilities located near forested regions. This study details the enhanced capabilities of CWSEFFP II, highlighting its innovative differential spraying technology and its potential for large-scale forest fire prevention and suppression.

2.3.1 Optimizing the Location of Pumps and Water Sources

To achieve a spray range of 2000 m, the water pressure at the nozzle of the water curtain tower at the furthest end must meet a minimum threshold. To ensure this, the following two conditions are critical when determining the placement of the pressurizing pump and water source:

① *Installation Altitude: The pressurizing pump and water source should be located at the highest possible altitude to maximize the benefits of head pressure compensation.*
② *Centralized Location: The installation site must be strategically positioned to optimize head pressure distribution, ideally at a central point within the 2000-m range.*

Additionally, the performance of the pressurizing pump must be specified to ensure that the water pressure at the nozzle of the CWSEFFP II meets or exceeds the required minimum value. Proper calibration of these parameters is essential to maintain the system's operational efficiency and effectiveness in long-range water spraying for fire suppression.

2.3.2 Smart IoT Tower Control Technology

① *Onboard AI Chip-Enabled CCTV Technology*: The onboard AI chip technology embedded within CCTV cameras enables real-time analysis and decision-making directly at the edge, minimizing latency and reducing reliance on external computational resources. Key steps include.

- Real-Time Data Processing: Video feeds from the CCTV cameras are analyzed using deep learning algorithms pre-trained on datasets for specific anomaly detection, such as fire flames, smoke, or unauthorized access.
- Event Detection and Classification: AI algorithms classify detected events based on pre-defined parameters. For instance, the system distinguishes between smoke caused by a forest fire and benign atmospheric changes.
- Triggering Alerts: Upon detection of anomalies, the system generates automated alerts and transmits them to the central monitoring hub or directly to the pipeline valve control system.

② *Pipeline Valve Control Technology*: The water pipeline valve control system is integrated with the onboard AI-enabled CCTV to ensure immediate response to detected anomalies. This includes.

- Automated Control: Based on input from the AI system, valves are programmed to respond autonomously by adjusting water flow to affected areas. For instance:
- Open-Valve Protocol: Valves are opened to release water for fire suppression or to maintain pressure in critical areas.
- Close-Valve Protocol: Valves are closed to isolate damaged sections of the pipeline or to redirect water resources.
- Remote Manual Override: Operators can manually control valves via a centralized interface when AI automation requires supplementary decision-making.
- Feedback Loops: Sensors embedded in the system provide real-time data on water flow, pressure, and valve status, creating feedback loops that enhance operational accuracy.

③ *System Integration Workflow*: The integration of AI-enabled CCTV technology with water pipeline valve control systems follows a structured workflow designed to enhance efficiency and response time in mitigating wildfire hazards. The process is outlined as follows:

- Detection: AI-enabled CCTV cameras continuously monitor designated areas, detecting anomalies such as smoke, flames, or other signs of wildfire ignition in real time.
- Analysis and Classification: The onboard AI module processes the captured visual data, classifying detected events. Upon identifying a potential wildfire ignition point, the system verifies the event and initiates the activation of a control mechanism, such as opening an alarm check valve to deliver water to the affected area.
- Automatic Response: If no other wildfire-related hazards are detected, the system autonomously triggers the appropriate valve action to deliver water directly to the threatened zone, thereby mitigating the risk of fire spread.
- Central Monitoring and Reporting: All operational activities, including detection, analysis, and response, are recorded and visualized in real time on a centralized monitoring dashboard. This platform enables remote supervision and facilitates coordinated decision-making during critical events.

3 Study Results

The CWSEFFP II system consists of 10 sets of fire towers, with each set containing three nozzles, resulting in a total of 30 nozzles. Each set is configured to spray water at a rate of 0.08 ℓ/m^2 per minute for a 10-min duration, allowing for a total of 0.32 ℓ/m^2 of water to be applied over 40 min as each set operates four times within a spraying range of up to 2 kms. The design of CWSEFFP II enables selective operation of individual fire tower nozzles, with the arrangement of N fire tower nozzles providing coverage across approximately 2000 m of the protected area.

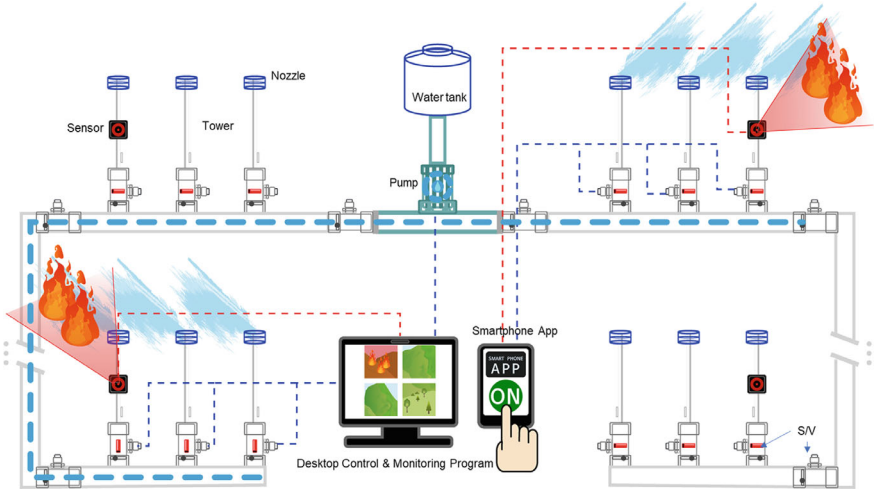


Fig. 5 Schematic diagram of CWEFFP II

To facilitate precise control, automatic opening, and closing valves are installed on the pipelines connected to each nozzle, numbered from 1 to N. These valves can be sequentially controlled or selectively activated based on the location of the detected ignition point of a forest fire. The system integrates onboard AI chip-enabled technology, which analyses CCTV footage in real time to detect forest fires automatically. This functionality allows for immediate and automated control of the valves, ensuring that water is directed efficiently to the nozzles nearest to the fire source.

The control system is programmed to ensure that at least three fire tower nozzles, selected from numbers 1 to N, remain active at any given time, maintaining a minimum spraying range of 200 m. A conceptual diagram of the CWSEFFP II system is presented in Fig. 5. Additionally, the system allows for remote operation and monitoring through a computer or smartphone application. Real-time operational status and functionality can also be verified via integrated CCTV systems, further supported by AI-driven fire detection to enhance overall efficiency and responsiveness (Fig. 6).

4 Conclusions

This study highlights significant advancements in forest fire prevention and suppression through the development of the Crown Water Spray Equipment for Forest Fires Protection (CWSEFFP) systems. These systems, particularly the CWSEFFP II, mark a substantial improvement in protecting Wildland-Urban Interface (WUI) areas and forest-adjacent communities from the increasing threat of forest fires, which are intensifying due to climate change. The key findings of this study are as follows:

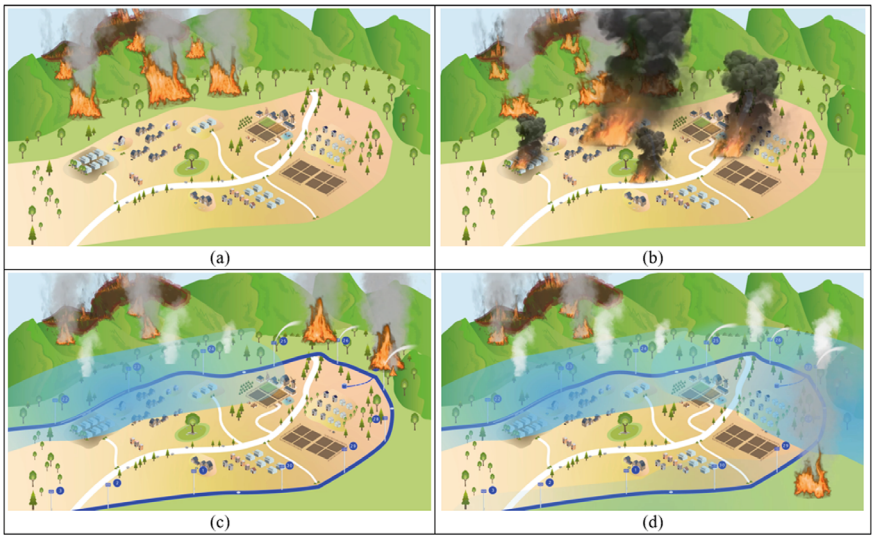


Fig. 6 Conceptual diagram of CWSEFFP II system operation (protection of village units in areas adjacent to forests); **a** forest fire spread to areas adjacent to forests, **b** village destruction by fire after forest fires spread to areas adjacent to forests, **c** protection of areas adjacent to forests after CWSEFFP II operation, **d** spotting fire suppression after CWSEFFP II operation

1. Enhanced Fire Suppression Capabilities: The CWSEFFP II system, equipped with advanced differential spraying technology and AI-driven controls, achieves a spraying range of up to 2 km—ten times greater than the original CWSEFFP I model. This advancement enables the protection of extensive areas, including entire villages and cultural heritage sites, effectively mitigating the spread of large-scale forest fires.
2. Integration of Smart IoT Technologies: The incorporation of onboard AI-enabled CCTV systems facilitates real-time monitoring and rapid detection of fire ignition points. This ensures precise and automated water distribution, enhancing fire suppression efficiency while minimizing resource wastage.
3. Scalability and Customization: The modular design of the CWSEFFP II system supports scalable deployment. Customizable nozzle configurations and adjustable water flow rates enable site-specific optimization, while automatic valves provide flexibility to target high-risk areas efficiently.
4. Global Applicability: The CWSEFFP systems address the global challenge of prolonged and increasingly intense forest fires. Their innovative design offers a practical and effective solution to safeguard vulnerable communities and critical ecosystems in various environmental and geographical contexts.

In conclusion, this study underscores the importance of integrating advanced IoT and AI technologies into forest fire suppression systems. The CWSEFFP II system not only mitigates the immediate impacts of forest fires but also establishes a robust framework for addressing the growing challenges posed by wildfires worldwide.

Further research and widespread deployment of such systems hold significant potential for enhancing global resilience to forest fires and reducing their devastating consequences.

Acknowledgements This research was supported by Grant funded by Cultural Heritage Administration (Project No: RS-2004-00402244)," Development and demonstration of fire safety technology for wooden architectural heritage in response to climate change.

References

1. Atroshenko YuK, Kuznetsov GV, Volkov RS, Strizhak PA (2019) Protective lines for suppressing the combustion front of forest fuels: experimental research. Process Saf Environ Prot 131:73–88
2. Chiu C-W, Li Y-H (2015) Full-scale experimental and numerical analysis of water mist system for sheltered fire sources in wind generator compartment. Process Saf Environ Prot 98:40–49
3. Cui Y, Liu J (2021) Research progress of water mist fire extinguishing technology and its application in battery fires. Process Saf Environ Prot 149:559–574
4. Kim D-H, Kim E-S, Kim J-W (2011) Study on estimating the unit of suppression ability of forest fire suppression resources. In: Proceeding of 2011 Korean Institute of Fire Science & engineering spring conference, pp 144–147
5. Kim D-H, Nam S-H, Keum S-H (2013) Study on guideline of water supply system for forest fire. J Korea Inst Fire Sci Eng 27(3):38–46
6. Grant G, Brenton J, Drysdale D (2000) Fire suppression by water sprays. Prog Energy Combust Sci 26:79–130
7. Gupta M, Rajora R, Sahai S, Shankar R, Ray A, Kale SR (2012) Experimental evaluation of fire suppression characteristics of twin fluid water mist system. Fire Saf J 54:130–142
8. Voytkova IS, Volkova RS, Kopylovb NP, Syshkinab EYu, Tomilinb AV, Strizhak PA (2021) Impact of scattered radiation on thermal radiation shielding by water curtains, Process Saf Environ Prot 154:278–290
9. Korea Forest Service (2011) Specifications for forest fire water supply system in Yangyang National Forest Station. Specification of Construction
10. Korea Forest Service (2011) Specifications for forest fire water supply system in Kangreung National Forest Station. Specification of Construction
11. Kuznetsov GV, Strizhak PA, Volkov RS, Vysokomornaya OV (2016) Integral characteristics of water droplet evaporation in high temperature combustion products of typical flammable liquids using SP and IPI methods. Int J Therm Sci 108:218–234
12. Liu Z, Kim AK (1999) A review of water mist fire suppression systems—fundamental studies. J Fire Prot Eng 10(3):32–50
13. Liu ZG, Kim AK (2001) A review of water mist fire suppression technology: part II—application studies. J Fire Prot Eng 11:16–42
14. National Fire Protection Association (2012) NFPA 1141; standard for fire protection infrastructure for land develop ment in wildland, rural, and suburban areas, 2012 edn
15. National Emergency Management Agency (2012) Regulation for outdoor fire hydrant systems. National Fire Safety Code 109, No. 2012-126