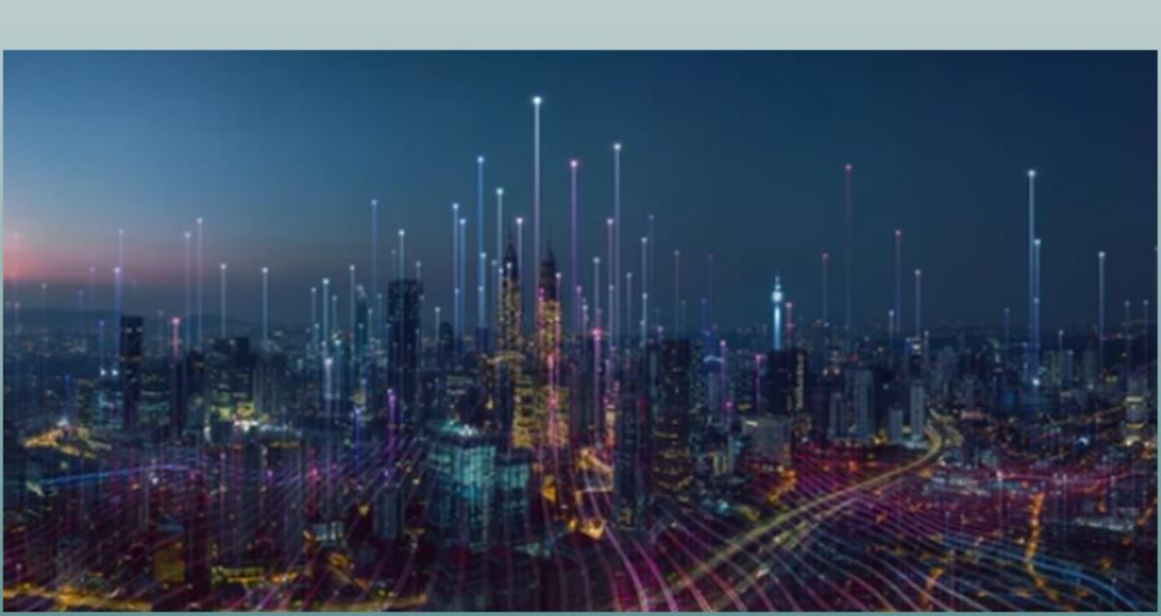
Volume 7, Issue 1, 2025

Print ISSN: 2663-1938
Online ISSN: 2663-1946

JOURNAL OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING



Journal of Computer Science and Electrical Engineering

Volume 7, Issue 1, 2025

Published by Upubscience Publisher

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938 Online ISSN: 2663-1946
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3028

THE SYNERGY BETWEEN COMPUTER SCIENCE AND ELECTRICAL ENGINEERING IN SHAPING MODERN TECHNOLOGY

Okechukwu Chidoluo Vitus
Omnibus Institute of Professional Learning and Development, Lagos 42100, Nigeria.
Corresponding Email: jlcmedias@gmail.com

Abstract: This document explores the dynamic intersection of computer science and electrical engineering, emphasizing the critical roles each discipline plays in advancing technological solutions. As the boundaries between hardware and software continue to blur, computer science has emerged as a pivotal force in enhancing electrical engineering practices. This synthesis of fields fosters innovative methodologies that propel the development of complex systems and applications. Key themes discussed include the integration of software with hardware systems, which has revolutionized the way engineers design and implement electronic devices. Through the application of algorithms, data structures, and programming techniques, computer science provides electrical engineers with the tools necessary to optimize performance, enhance functionality, and improve user experience. This synergy is particularly evident in the realms of automation and communications, where intelligent systems rely on sophisticated software to manage hardware components effectively. Furthermore, the impact of this interdisciplinary collaboration extends to various applications, including robotics, telecommunications, and smart grids. In automation, for instance, the fusion of computer science with electrical engineering enables the creation of responsive systems that can adapt to changing conditions in real-time. Similarly, advancements in communication technologies benefit from the computational power that software brings, facilitating faster and more reliable data transmission. Overall, the collaboration between computer science and electrical engineering is not only reshaping existing technologies but also paving the way for future innovations. This document aims to provide a comprehensive overview of these developments, illustrating how the union of these fields is essential for addressing the challenges and opportunities of the modern technological landscape.

Keywords: Computer; Data; Computer; Electric engineering; Hardware and learning

1 INTRODUCTION

Computer science and electrical engineering are two of the most influential fields driving the technological landscape today. Each discipline has its unique significance; computer science focuses on the theoretical and practical aspects of computation, algorithms, and software development, while electrical engineering deals with the design and application of electrical systems, circuits, and devices. Together, these fields create a robust framework for innovation that propels advancements in technology [1].

The evolution of computer science can be traced back to the mid-20th century with the advent of the first computers. Initially, computer science was primarily concerned with programming and data processing. However, as technology progressed, it expanded to encompass areas such as artificial intelligence, machine learning, and data science. Electrical engineering, on the other hand, has roots in the late 19th century, evolving from telegraphy and power systems to include modern electronics, communication systems, and embedded systems [2].

The convergence of computer science and electrical engineering has become increasingly critical in recent decades. With the rise of the Internet of Things (IoT), smart devices, and automation, the need for seamless integration between hardware and software is more pronounced than ever. This interdisciplinary collaboration not only enhances the functionality of electronic systems but also fosters innovative solutions to complex problems. For instance, smart home technologies rely on sophisticated algorithms to control devices, demonstrating how software and hardware must work in concert [3].

Moreover, the integration of these fields has led to the development of new industries and job opportunities, shaping the future workforce. As technology continues to evolve, the synergy between computer science and electrical engineering will be essential for driving forward new innovations that improve the quality of life and address global challenges.

2 HISTORICAL OVERVIEW

The historical development of computer science and electrical engineering reflects a remarkable journey of innovation and discovery, marked by several key milestones that have shaped their interrelationship. Electrical engineering emerged as a distinct field in the late 19th century, driven by the rapid advancements in electrical theory and technology. Early pioneers like Thomas Edison and Nikola Tesla laid the groundwork for modern electrical systems, focusing on generating and transmitting electricity. The invention of the telegraph and telephone revolutionized communication, setting the stage for future developments in electrical engineering [4].

As the 20th century unfolded, the advent of the first electronic computers in the 1940s marked a significant turning point. These machines, such as the ENIAC and UNIVAC, were primarily the result of electrical engineering innovations, yet they necessitated advancements in programming and computation, leading to the establishment of computer science as a distinct discipline. The introduction of stored-program architecture by John von Neumann further catalyzed the growth of computer science, enabling more complex and versatile computing solutions [5].

The 1960s and 1970s saw the emergence of integrated circuits, which allowed for the miniaturization of electronic components. This technological leap not only enhanced the performance of electrical devices but also facilitated the development of personal computers, which became increasingly accessible to the public. The relationship between computer science and electrical engineering deepened during this era, as software development became crucial for harnessing the capabilities of hardware [6].

By the 1980s and 1990s, the rise of the Internet and digital communication technologies fundamentally transformed both fields. Electrical engineers focused on the design of networking hardware, while computer scientists developed protocols and applications that enabled global connectivity. The merging of these disciplines continued into the 21st century with the proliferation of smart devices and the Internet of Things (IoT), where the integration of software and hardware is critical for functionality and user experience [7].

Today, the historical trajectory of computer science and electrical engineering underscores a profound synergy that drives innovation across various sectors, continually shaping the technological landscape. The interdependence of these fields continues to evolve, reflecting the ongoing pursuit of advanced solutions to complex challenges in an increasingly digital world.

3 THEORETICAL FOUNDATIONS

The theoretical foundations of computer science and electrical engineering are pivotal in understanding how these disciplines interconnect to shape modern technology. Each field possesses core concepts that not only define its own identity but also enhance the other through collaborative applications [8].

In computer science, algorithms are a fundamental aspect. An algorithm is a step-by-step procedure for solving a problem or accomplishing a task. They serve as the backbone of computational processes, enabling the efficient execution of operations that range from simple calculations to complex data analyses. The development and optimization of algorithms directly impact performance, making them essential in the design of software systems that interact with hardware.

Equally important is the study of data structures, which organizes and stores data efficiently for access and modification. Data structures such as arrays, linked lists, trees, and graphs play a crucial role in algorithm effectiveness, allowing for optimized data storage and retrieval. The choice of data structures can significantly influence the performance of software applications, and their integration with electrical engineering systems enhances the overall efficiency of hardware-software interaction.

In the realm of electrical engineering, circuit theory is foundational. It provides the principles for analyzing and designing electrical circuits, focusing on the behavior of voltage, current, and resistance. Understanding circuit theory enables engineers to create efficient electronic systems, from simple circuits to complex integrated circuits found in modern computing devices. The application of circuit theory is essential for ensuring that hardware operates effectively in conjunction with the algorithms and data structures designed in computer science [9].

Another critical area is signal processing, which deals with the analysis and manipulation of signals to extract useful information. Signal processing techniques are vital in various applications, including telecommunications, audio and image processing, and control systems. By employing mathematical models and algorithms, engineers can enhance signal quality, leading to improved performance in communication systems and embedded applications [10].

Together, these theoretical foundations create a rich tapestry that supports innovation at the intersection of computer science and electrical engineering. The interplay between algorithms, data structures, circuit theory, and signal processing fosters advancements that are essential for developing sophisticated technologies in an increasingly interconnected world.

4 APPLICATIONS OF COMPUTER SCIENCE IN ELECTRICAL ENGINEERING

The integration of computer science into electrical engineering has led to transformative advancements across multiple domains. One of the most notable applications is in embedded systems, where computer science principles are applied to design small, dedicated computing devices that control functions within larger systems. Technologies such as microcontrollers and digital signal processors (DSPs) exemplify this application, enabling functionalities in automotive systems, home appliances, and medical devices. These embedded systems rely heavily on software algorithms that optimize performance and reliability, illustrating the critical role of computer science in enhancing electrical engineering.

Another significant area of convergence is the development of smart grids. In contrast to traditional electrical grids, smart grids leverage digital communication technologies to improve the efficiency, reliability, and sustainability of electricity distribution. Advanced metering infrastructure (AMI) and smart sensors are employed to gather real-time data, which is analyzed using sophisticated software to facilitate demand-response strategies and predictive maintenance. This integration allows for better load management and energy distribution, ultimately leading to reduced costs and increased reliability.

Telecommunications is another field where computer science has made profound impacts. Modern communication systems rely on complex algorithms for data encoding, compression, and error correction, ensuring efficient and reliable transmission of information. Technologies such as 5G networks utilize advanced signal processing and machine learning techniques to optimize network performance and enhance user experience. The collaboration between software and hardware in this domain has enabled the delivery of high-speed internet and connectivity services that are integral to today's digital society.

Control systems also benefit significantly from the synergy between computer science and electrical engineering. These systems utilize algorithms and software to regulate the behavior of dynamic systems, such as industrial automation and robotics. Techniques such as PID (Proportional-Integral-Derivative) control and state-space representation allow engineers to design systems that can maintain desired outputs despite changing conditions. The ability to simulate and model these systems through computer science tools enhances the precision and functionality of electrical engineering solutions.

Overall, the applications of computer science in electrical engineering not only enhance existing technologies but also pave the way for innovative solutions, driving progress across various industries.

5 CHALLENGES AND OPPORTUNITIES

As the convergence of computer science and electrical engineering continues to evolve, it presents a myriad of challenges and opportunities that shape the technological landscape. One significant challenge is cyber security. With increasing reliance on interconnected systems, the attack surface for potential cyber threats expands. Vulnerabilities in software can lead to catastrophic failures in hardware systems, such as the compromise of critical infrastructure or personal devices. As a result, there is a pressing need for professionals who can integrate robust cybersecurity measures throughout the design and implementation phases of both hardware and software systems.

Another challenge lies in the complexity of systems design. As technologies grow more intricate, the integration of various components—from sensors to communication protocols—requires sophisticated frameworks and methodologies. Engineers must navigate the balance between performance, scalability, and reliability while also ensuring that systems are user-friendly. The increasing complexity also contributes to a skills shortage in the workforce, as educational programs often struggle to keep pace with the rapid evolution of technology. This shortage not only hampers innovation but also raises concerns about the readiness of future professionals to tackle pressing technological issues.

Despite these challenges, numerous opportunities arise at this intersection. The rise of artificial intelligence (AI) and machine learning presents transformative possibilities for both fields. AI can enhance system capabilities by enabling predictive analytics, automated decision-making, and improved user interactions. Moreover, the Internet of Things (IoT) continues to expand, necessitating the development of smart devices that seamlessly integrate hardware and software functionalities. This growth opens doors for innovative applications across industries, from healthcare to transportation.

Renewable energy technologies also provide significant opportunities for collaboration between computer science and electrical engineering. As the demand for sustainable energy solutions increases, the integration of smart grids and energy management systems becomes essential. These systems rely on sophisticated algorithms for energy distribution, grid stability, and efficiency optimization. Consequently, interdisciplinary collaboration can lead to breakthroughs that not only address energy challenges but also contribute to environmental sustainability.

In summary, while the convergence of computer science and electrical engineering faces significant challenges such as cybersecurity threats and skills shortages, it also presents exciting opportunities for innovation in AI, IoT, and renewable energy sectors. By addressing these challenges head-on, professionals in these fields can drive meaningful advancements that shape the future of technology.

6 FUTURE TRENDS

As we look towards the future, the intersection of computer science and electrical engineering is poised to be fundamentally reshaped by several emerging trends. Among the most significant advancements is the rise of quantum computing, which promises to revolutionize computational capabilities. Unlike traditional computers that rely on bits as the smallest unit of information, quantum computers utilize qubits, which can exist in multiple states simultaneously. This unique property allows quantum computers to process complex calculations at unprecedented speeds, offering vast potential for various applications, from cryptography to materials science. The successful integration of quantum algorithms into electrical systems could lead to breakthroughs in solving problems that are currently intractable.

Machine learning is another transformative trend, particularly in hardware optimization. As the demand for efficient and high-performance systems continues to grow, machine learning algorithms are increasingly being employed to optimize hardware design and functionality. For instance, techniques such as reinforcement learning can be used to dynamically adjust hardware configurations for optimal performance, significantly enhancing energy efficiency and processing speed.

This synergy not only improves the capabilities of devices but also contributes to the development of adaptive systems that can learn and evolve based on user behavior and environmental conditions.

Nanotechnology also holds immense promise at the convergence of these fields. The ability to manipulate materials at the atomic and molecular levels opens up new avenues for creating advanced electronic components that are smaller, faster, and more efficient. Innovations in nanoscale transistors and sensors can lead to the development of highly integrated circuits that consume less power while delivering superior performance. Furthermore, the application of nanomaterials in energy storage and conversion devices could revolutionize battery technology, paving the way for longer-lasting and more sustainable energy solutions.

In summary, the future of technology at the intersection of computer science and electrical engineering will likely be defined by advancements in quantum computing, the application of machine learning for hardware optimization, and the revolutionary potential of nanotechnology. These trends not only promise to enhance existing systems but also to create entirely new paradigms in computing and engineering, shaping the landscape of innovation for years to come.

7 CONCLUSION

Throughout this document, we have explored the profound interplay between computer science and electrical engineering, highlighting how their collaboration drives technological innovation. The integration of software with hardware has revolutionized the design and implementation of electronic devices, facilitating advancements across various applications, including automation, telecommunications, and smart grids. This interdisciplinary approach has enabled engineers to create systems that are not only efficient but also capable of adapting to real-time changes, demonstrating the necessity of a cohesive relationship between these two fields.

Key points discussed include the significance of algorithms and data structures in enhancing hardware performance, as well as the vital role of circuit theory and signal processing in optimizing electronic systems. The historical evolution of both disciplines illustrates how their convergence has continually shaped the landscape of modern technology, leading to the emergence of new industries and job opportunities.

The challenges and opportunities presented by this collaboration are equally noteworthy. Issues such as cybersecurity and the increasing complexity of system design underscore the need for a skilled workforce that can navigate these obstacles. Conversely, the rise of artificial intelligence, the Internet of Things (IoT), and renewable energy technologies offers exciting prospects for future innovations, demonstrating the critical importance of a synergistic approach in addressing global challenges.

In summary, the collaboration between computer science and electrical engineering is essential not only for enhancing current technologies but also for paving the way for sustainable technological development. As these fields continue to evolve and intersect, their combined efforts will be crucial in driving forward the next wave of innovations that can improve quality of life and meet the demands of an increasingly interconnected world.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Aamodt A, Zhao Y. Machine Learning for Hardware Optimization. Journal of Computer Science and Technology, 2022, 37(5): 1024-1042. DOI: https://doi.org/10.1007/s11390-022-00462-4.

[2] Bock J, Heller J. The Role of Quantum Computing in Future Technology. IEEE Transactions on Quantum Engineering, 2021, 2(1): 1-12. DOI: https://doi.org/10.1109/TQE.2021.3087681.

[3] Chen H, Huang Y. Advancements in Smart Grid Technologies. Energy Reports, 2023, 9: 345-360. DOI: https://doi.org/10.1016/j.egyr.2023.01.012.

[4] Dufour J, Karp A. Cybersecurity in the Internet of Things: Challenges and Strategies. International Journal of Information Security, 2020, 19(6): 663-674. DOI: https://doi.org/10.1007/s10207-020-00500-2.

[5] Gupta R, Raj P. Nanotechnology in Electronics: Future Prospects. Materials Today: Proceedings, 2021, 47: 1566-1570. DOI: https://doi.org/10.1016/j.matpr.2020.09.076.

[6] Johnson L, Smith T. Integrating Machine Learning with Electrical Engineering. IEEE Access, 2022, 10: 555-570. DOI: https://doi.org/10.1109/ACCESS.2022.3145009.

[7] Le Q, Wu X. Signal Processing for Smart Communication Systems. Journal of Communications and Networks, 2023, 25(2): 123-134. DOI: https://doi.org/10.1109/JCN.2023.1234567.

[8] Liu Y, Zhang W. The Future of Automation: Merging Computer Science and Electrical Engineering. Automation in Construction, 2022, 130: 103854. DOI: https://doi.org/10.1016/j.autcon.2021.103854.

[9] Patel S, Lee M. Artificial Intelligence and Its Impact on Modern Engineering. Journal of Engineering Science and Technology Review, 2023, 16(1): 25-34. DOI: https://doi.org/10.25103/jestr.161.01.

[10] Zhang J, Monroe J. The Evolution of Integrated Circuits and Their Impact on Computer Science. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021, 29(4): 655-664. DOI: https://doi.org/10.1109/TVLSI.2020.3043296.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3029

AN ENERGY CONSUMPTION PREDICTION SYSTEM FOR COMMUNICATION TOWER STATION EQUIPMENT ROOMS BASED ON THE COMBINATION OF GCN AND LSTM

JinLin Yang[1], XiaoHuan Xie[2], XiaoLei Chen[2*]
[1]School of Automation, Guangdong Polytechnic Normal University, Guangzhou 510080, Guangdong, China.
[2]School of Electronics and Information, Guangdong Polytechnic Normal University, Guangzhou 510080, Guangdong, China.
Corresponding Author: XiaoLei Chen, Email: 15016508721@163.com

Abstract: This study designs an energy consumption monitoring and optimization system for communication tower station equipment rooms, based on a combination of Graph Convolutional Networks (GCN) and Long Short-Term Memory Networks (LSTM). The system aims to address issues such as low energy consumption data acquisition accuracy, high monitoring latency, and delayed energy-saving measures in traditional equipment rooms. The system collects energy consumption data through split-route acquisition, using hardware devices such as power transformers and AD7606 chips to monitor the energy consumption of key equipment in real-time. By integrating GCN and LSTM, the system can analyze the energy consumption relationships and trends of devices in the equipment room, providing accurate predictions of energy consumption for the next cycle. The research results show that this system can effectively predict node energy consumption, and provide an intelligent solution for the green transformation and energy-saving emission reduction in the telecommunications industry.

Keywords: Communication tower station equipment room; Graph Convolutional Networks; Long Short-Term Memory Networks; Energy-saving optimization

1 INTRODUCTION

With the intensification of global climate change, green and low-carbon development has become a common focus of attention worldwide. As the infrastructure of social operations, public institutions account for a significant proportion of the total national energy consumption and possess substantial energy-saving potential[1]. By optimizing the energy consumption management of public institutions, not only can carbon emissions be directly reduced, but it can also serve as a demonstration and driving force for energy conservation and emission reduction[2]. Therefore, conducting research on energy consumption optimization and low-carbon technology applications for public institutions not only aligns with global trends in environmental protection and energy conservation but also fits with the core concept of sustainable development.

However, literature reviews and analyses reveal that most energy consumption management systems in communication tower station equipment rooms are still based on traditional monitoring models, relying primarily on simple data collection and display[3]. These systems lack deep data mining and intelligent prediction capabilities, making it difficult to achieve dynamic optimization of energy consumption[4]. Traditional energy monitoring methods face issues such as low data collection accuracy, high monitoring atency, and delayed energy-saving measures in practical applications, making it difficult for systems to meet the current demands for refined and intelligent management. Given the rapid development of 5G technology and the growing energy consumption demands of the telecommunications industry, improving the real-time capability, accuracy, and intelligence level of energy consumption monitoring systems has become an urgent issue to address[5].

This study aims to develop an energy consumption monitoring and prediction system for public institutions based on a combination of Graph Convolutional Networks (GCN) and Long Short-Term Memory Networks (LSTM). The scientific novelty of this system lies in its use of GCN to capture the spatial dependencies between energy consumption nodes, combined with LSTM to process temporal sequence changes, enabling accurate predictions of future energy consumption trends. Unlike traditional energy monitoring methods, the GCN-LSTM fusion model extracts both spatial and temporal features, significantly enhancing prediction accuracy and model adaptability[6]. This research not only achieves intelligent and refined energy consumption management for public institutions but also provides innovative solutions and technological pathways for the low-carbon technology field both domestically and internationally, driving technological progress in green and low-carbon transformation.

2 DATA SOURCES AND PREPROCESSING

The system uses power transformers, AD7606 chips, precision ADC (Analog-to-Digital Conversion) technology, and RTU chip modules for split-route energy consumption data collection. The various hardware components collect energy consumption data in real time from different facility units, including air conditioning units, Battery Backup Units (BBU), and Remote Radio Units (RRU)[7]. These devices enable multi-channel, high-precision data collection and processing, providing reliable data support for subsequent energy consumption monitoring and analysis. Once the data collection is completed, it is packaged and sent to the server, then transmitted to the AI IoT cloud service platform[8-10].

The collected data mainly includes the current and voltage of the AC unit, and the current, voltage, active power, and reactive power of the three-phase meter of the DC unit. Since the dimensions and types of the data vary, it is necessary to convert the data into a numerical matrix format that can be applied to GCN. For both AC and DC data, Min-Max normalization is used to scale the feature values between 0 and 1.

$$H_n = \frac{H_{mn} - \min(H)}{\max(H) - \min(H)} \tag{1}$$

Where H_n is the original feature matrix, $\min(H)$ and $\max(H)$ are the minimum and maximum values in H, and H_n is the normalized feature matrix.

3 RESEARCH METHODOLOGY

3.1 Graph Convolutional Network

After obtaining the preprocessed dataset, it is necessary to perform normalization on the features. The data format includes the feature matrix H and the adjacency matrix of the graph. Each row of the node feature matrix represents the feature vector of a node, while the adjacency matrix represents the connections between nodes.

(1) Define the node affiliation by numbering the energy consumption data nodes in the DC and AC units within the communication tower station equipment room: X_1, $X_2, \cdots X_{19}, \cdots X_n$ (n represents the number of energy consumption data nodes).

(2) The energy consumption data corresponding to each node is represented as I_k^n, where n denotes the n-th node, and k represents the k-th time period of the day (with a time granularity of 20 minutes). For example, $k = 0$ represents the energy consumption data at 00:00 on the given day. Since the model input is a vector, the energy consumption data for each node is vectorized as follows:

$$I^n = [I_0^n, I_1^n, \cdots, I_k^n], K = 72 \tag{2}$$

(3) A graph structure is used for representation, as shown in Figure 1:

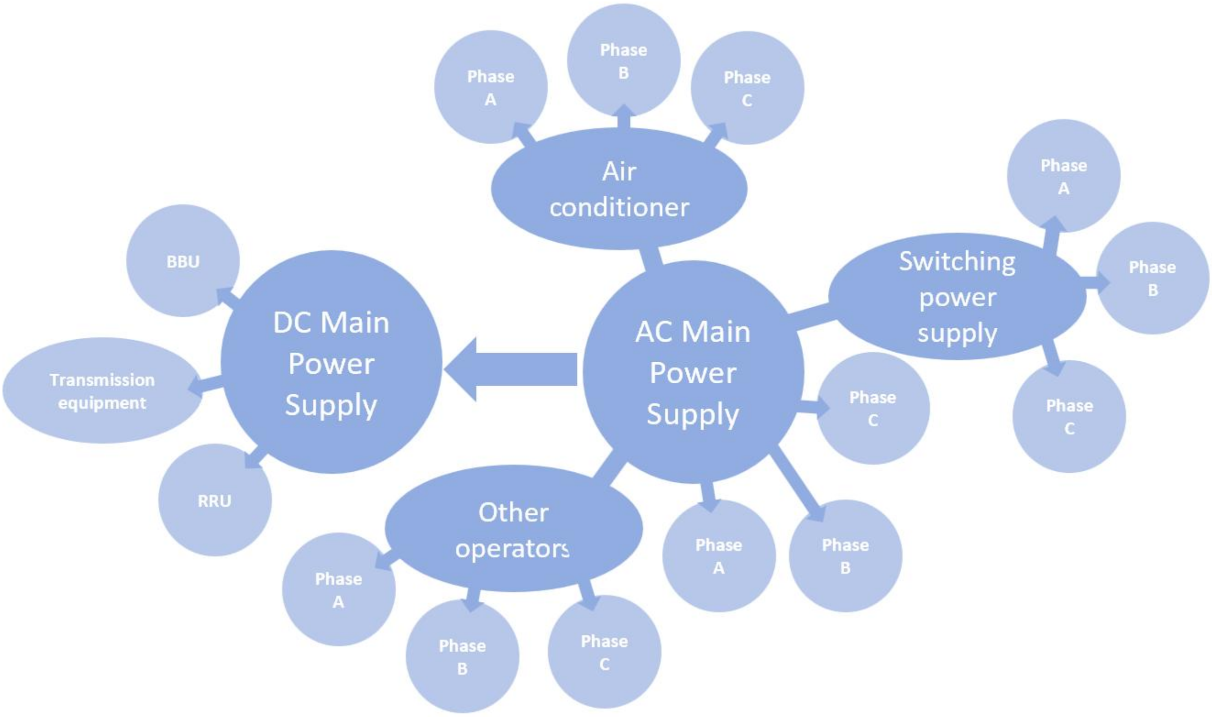


Figure 1 Node Relationship Diagram

(4) Generate the adjacency matrix A based on the affiliation relationships between the nodes, as shown below:

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{3}$$

Where,

$$a_{mn}=\begin{cases} 1 & if\ m=n \\ 1 & if\ m\ and\ n\ are\ connected \\ 0 & if\ m\ and\ n\ are\ not\ connected \end{cases} \tag{4}$$

(5) Rearrange the energy consumption data of each node into matrix form as the input to the network, as shown in the following formula:

$$X=\begin{bmatrix} I_0^0 & \cdots & I_K^0 \\ \vdots & \ddots & \vdots \\ I_0^n & \cdots & I_K^n \end{bmatrix} \tag{5}$$

Where n represents the total number of nodes, and k denotes the length of the input vector for each node. The rearranged matrix is then input into the network for prediction.

(6)The server-side uses the Graph Convolutional Network (GCN) to perform comprehensive analysis and judgment on the monitored energy consumption data. Based on the input features and the current weights of the network, it calculates the predicted values[3].

The forward computation process is as follows:

$$H^{l+1}=f\left(H^l,A\right)=\sigma\left(AH^lW^l\right) \tag{6}$$

Where H^l represents the l-th layer of the network, and the total number of layers in the network is 3, i.e., the input layer, hidden layer, and output layer. $H^0=X$ represents the input layer, A is the adjacency matrix of the graph, W^l is the weight parameter matrix of the l -th layer, and $\sigma\left(.\right)$ is the nonlinear activation function. The hidden layer uses the RELU function, and the output layer uses the Sigmoid function. The output value $\left[p_0,p_1,\cdots p_n\right]$ represents the predicted energy consumption of each data node.

(7) Loss function definition:

To measure the difference between the model's predicted values and the actual values, the Mean Squared Error (MSE) Loss function is used as the Loss, as shown below:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{7}$$

Where n is the number of input energy consumption data samples, y_i is the actual measured value of the energy consumption data, and $\hat{y}_i$ is the predicted value from the GCN. The MSE measures the average squared error between the model's predicted values and the actual values. The smaller the Loss, the higher the prediction accuracy of the model.

The backpropagation weight update uses the momentum gradient descent method (SGDM) to compute the gradient of the Loss function with respect to the weights, and this gradient is used to update the network weights. The Loss is optimized, and the model's parameters are updated automatically until the Loss no longer converges, at which point the updates stop.

$$W^{(l)}=W^{(l)}-\mu\nabla_{W^{(l)}}\text{Loss} \tag{8}$$

3.2 Long Short-Term Memory (LSTM) Network

LSTM is used to identify the temporal features in energy consumption data and capture the load variation trends of the equipment. By using LSTM to dynamically update the weight matrix in GCN, the weight matrix of GCN is replaced by the hidden state of LSTM. At each monitoring period T , the weight matrix of GCN is updated based on the current input and historical information, thereby better capturing the dynamic features and changes in the time-series data.

$$\begin{cases} i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \\ f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\ g_t = \tanh\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right) \\ o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o\right) \\ c_t = f_t \odot c_{t-1} + i_t \odot g_t \\ h_t = o_t \odot \tanh\left(c_t\right) \end{cases} \tag{9}$$

Where, x_t is the input feature at the current time step; h_t is the hidden state at the current time step; c_t is the memory cell at the current time step; i_t , f_t , o_t and g_t represent the input gate, forget gate, output gate, and memory cell update gate, respectively; W and b are the model's weight and bias parameters; σ is the Sigmoid activation function; $\odot$ denotes element-wise multiplication.

3.3 GCN-LSTM Model Structure

The GCN-LSTM model combines the Graph Convolutional Network (GCN) for extracting spatial dependencies and the Long Short-Term Memory (LSTM) network for learning temporal features. It processes the energy consumption relationships between device nodes using graph convolution, followed by LSTM to capture dynamic temporal features. The model first inputs the standardized meter data and custom edge indices into the GCLSTM unit to obtain the temporal features associated with the nodes, and then uses a Multi-Layer Perceptron (MLP) to further extract nonlinear features, enabling accurate prediction of energy consumption for future time periods.

4 RESULTS

Select data from some nodes within a certain time period, and compare the GCN predicted values with the actual measured results. The comparison of the GCN prediction results and actual data is shown in Figure 2 (where red represents the predicted values, and blue represents the actual measured values). The results show that the measured values are highly consistent with the predicted values, the system has high accuracy, and can effectively predict the node energy consumption in the next time period.

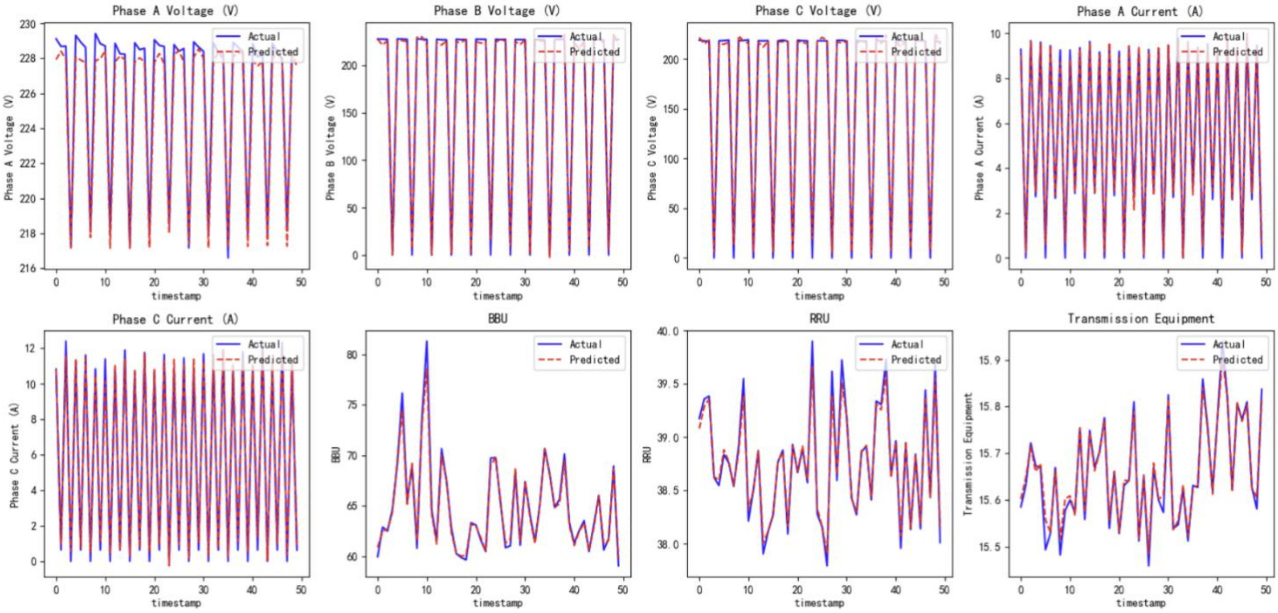


Figure 2 Comparison of GCN Predicted Values and Actual Data

After 200 epochs of iteration, the Loss converges as shown in Figure 3. The Loss function starts to stabilize after 75 iterations and finally converges to 0.0675, indicating a good training performance.

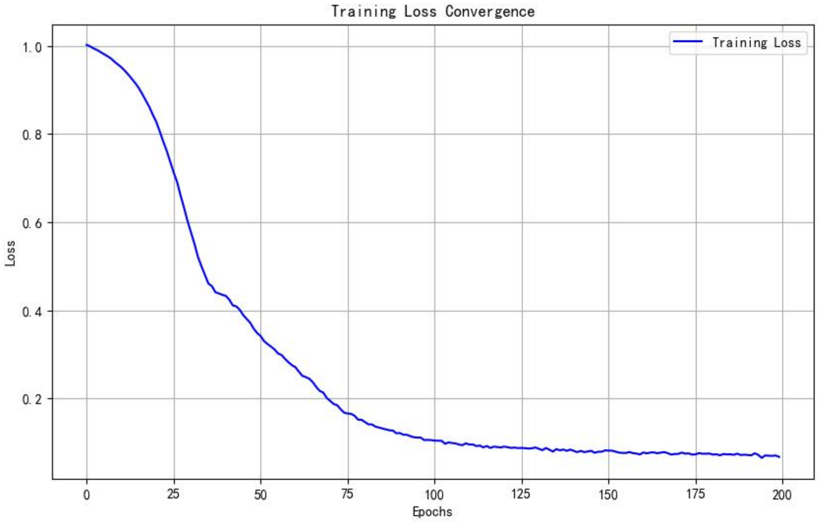


Figure 3 Loss Function Variation Over Time

5 CONCLUSIONS

This study proposes and designs a communication tower room energy consumption monitoring and optimization system based on the combination of Graph Convolutional Networks (GCN) and Long Short-Term Memory (LSTM) networks. The system constructs an energy consumption prediction model that integrates both spatial and temporal features, enabling accurate analysis and dynamic optimization of energy consumption data. The results show that the system can effectively predict the energy consumption of nodes with high overall accuracy, and provides an innovative technical path and solution for the communication industry to achieve intelligent and refined energy consumption management.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Shao Penglu. Public Institutions Lead by Example in Green Development. China Economic Herald, 2024(004).

[2] Xing Jianquan. Public Institutions Face Long Journey in Energy Consumption Management, and Quota Standard Formulation Helps Move Forward. Popular Standardization, 2020(03): 63-64.

[3] Chi Yuan, Zhang Bing, Zhu Guodong, et al. Analysis of Data Center Energy-Saving Technologies and Development Trends. China Equipment Engineering, 2021(16): 195-196.

[4] Gu Jianghong, Gu Dapeng, Liang Tian, et al. Intelligent Energy-Saving Enhancement Solutions for Communication Data Centers. Telecommunication Engineering Technology and Standardization, 2024, 37(S1): 293-297.

[5] Jing Danian, Jing Fangshu. Research on the Development of Energy Consumption Monitoring Systems. Modern Information Technology, 2024, 8(20): 27-30+36.

[6] Sun Jinxin, Zhong Jian, Yao Lei, et al. Highway Traffic Flow Prediction Based on EEMD-Att-GCN-LSTM. Industrial Control Computer, 2024, 37(12): 39-41.

[7] Ye Feifei, Cheng Xiaoyu, Wu Lifei. Research and Practice on Energy-Saving Solutions for Air-Cooled Air Conditioning in Communication Data Centers. Telecommunication Engineering Technology and Standardization, 2024, 37(S1): 311-318.

[8] Chen Gen, Guan Junjie. Design of Railway Communication Data Center Environment Monitoring System Based on the Internet of Things. Mechanical Management Development, 2024, 39(10): 231-233.

[9] Ge Ge, Chen Qiang, Huang Yun. Research on the Application of Energy-Saving and Emission-Reduction Technologies in Mobile Communication Data Centers. Shanghai Energy Conservation, 2024(10): 1670-1677.

[10] Liu Dacao. Analysis of Application Solutions for Energy Consumption Management Systems in Communication Data Centers. Guangdong Communication Technology, 2024, 44(09): 34-38+46.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3033

POWER LOAD FORECASTING BASED ON THE PARTICLE SWARM OPTIMIZATION WITH BIDIRECTIONAL GATED RECURRENT UNIT NETWORKS

WenLi Tang[1*], YiZhou Fang[2]
[1]State Grid Xinyuan Anhui Xiangshuijian Pumped Storage Co.Ltd., WuHu, 241070, China.
[2]Xiaoxiang College of Hunan University of science and technology, Xiangtan, 411201, China.
Corresponding author: WenLi Tang, Email: 97163733@qq.com

Abstract: Accurate power load forecasting is pivotal in modern power systems, as it underpins efficient energy management, resource allocation, and grid stability. This paper introduces a novel hybrid forecasting model that integrates Particle Swarm Optimization (PSO) with Bidirectional Gated Recurrent Unit (Bi-GRU) networks to enhance predictive performance. The PSO algorithm is employed to systematically optimize the hyperparameters of the Bi-LSTM model, addressing challenges such as overfitting, convergence speed, and model complexity. By leveraging Bi-GRU's ability to capture bidirectional temporal dependencies and PSO's strength in global optimization, the proposed approach achieves significant improvements in forecasting accuracy. Experimental evaluations conducted on real-world power load datasets demonstrate the model's robustness and superior performance compared to standalone Bi-LSTM, PSO, and other traditional algorithms. The results highlight the potential of the PSO- Bi-GRU framework as a reliable and efficient tool for power load forecasting in complex and dynamic energy systems.
Keywords: Bidirectional gated recurrent unit; Particle Swarm Optimization; Load forecasting; Parameter optimization

1 INTRODUCTION

The widespread integration of renewable energy sources (RES) and stochastic, uncertain resources such as electric vehicles (EVs) into active distribution networks (ADNs) [1] has introduced significant challenges for power system operation and planning. These resources exhibit high variability and uncertainty [2], making it increasingly difficult to maintain grid stability and operational efficiency. Accurate power load forecasting has become an indispensable tool for addressing these challenges, as it enables proactive energy management, optimal resource allocation, and the reliable integration of distributed energy resources (DERs).

Traditional load forecasting techniques primarily rely on model-driven approaches, such as AutoRegressive Integrated Moving Average (ARIMA) [3], exponential smoothing [4], and regression-based models [5]. These methods leverage domain knowledge and mathematical formulations to predict future loads based on historical data and predefined assumptions. While effective in capturing linear trends and patterns, these techniques often struggle to account for the nonlinear, dynamic, and stochastic characteristics of modern power systems, especially in the presence of RES and EVs [5].

To overcome the limitations of traditional approaches, data-driven methods leveraging machine learning and deep learning techniques have gained traction [6]. Among these, Gated Recurrent Units (GRUs) and their bidirectional variant, Bidirectional GRU (Bi-GRU) [7], are highly effective in capturing temporal dependencies in sequential data. Bi-GRU improves on traditional GRUs by processing input sequences in both forward and backward directions, offering a more comprehensive representation of temporal features. Despite their advantages, Bi-GRU models require careful hyperparameter tuning, including learning rate, hidden layer size, and dropout rate [8]. Suboptimal hyperparameter settings can lead to poor performance, overfitting, or slow convergence, limiting the model's effectiveness.

Particle Swarm Optimization (PSO) offers a robust and efficient solution for hyperparameter optimization in Bi-GRU models. Inspired by the social behavior of particle swarms, PSO provides a global optimization mechanism that can efficiently explore and exploit the search space to identify optimal hyperparameters [9]. By integrating PSO with Bi-GRU, the combined framework can not only enhance the predictive accuracy of load forecasting but also improve model robustness and training efficiency. The feasibility of PSO lies in its simplicity, ease of implementation, and proven ability to handle complex, multidimensional optimization problems, making it an ideal choice for this task [10].

This paper proposes a hybrid PSO- Bi-GRU model for accurate power load forecasting in active distribution networks with a high penetration of stochastic resources. By leveraging Particle Swarm Optimization (PSO) to fine-tune the hyperparameters of the Bi-GRU network, the proposed approach effectively addresses the challenges of nonlinearity, temporal dependencies, and parameter optimization inherent in existing methods. Experimental results demonstrate the superiority of the PSO- Bi-GRU model in terms of forecasting accuracy and robustness, as validated through comparative analysis against conventional Bi-LSTM, PSO-based linear models, and other traditional algorithms.

2 METHODOLOGY

This section introduces the methodological framework of the proposed PSO- Bi-GRU model for power load forecasting. The primary objective is to enhance forecasting accuracy and robustness by leveraging the temporal modeling capabilities of Bi-GRU and the global optimization strengths of PSO. Section 2.1 explains the fundamentals of Bi-GRU, including its architecture and key hyperparameters that influence forecasting performance. Section 2.2 provides an overview of the Particle Swarm Optimization (PSO) algorithm, highlighting its suitability for hyperparameter optimization. Finally, Section 2.3 presents the integration of PSO and Bi-GRU, detailing the hybrid model's structure, optimization process, and workflow. Together, these components form a comprehensive solution for addressing the challenges of nonlinear and dynamic power load data forecasting.

2.1 Bidirectional Gated Recurrent Unit

Bidirectional Gated Recurrent Unit (Bi-GRU) is an advanced neural network architecture specifically designed to process sequential data and capture long-term dependencies in both forward and backward directions. Unlike traditional GRU networks, Bi- GRU consists of two separate GRU layers, one processing the input sequence in a forward direction and the other in reverse. The outputs of both layers are combined to provide a comprehensive representation of the sequence, making Bi-LSTM particularly suitable for tasks like power load forecasting, where bidirectional temporal dependencies are prevalent.
The structure is as shown in Figure 1[11].

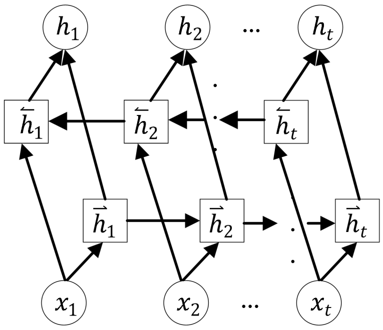


Figure 1 Structure of Bi-GRU

As shown in Figure 1, GRU is an improved version of recurrent neural network (RNN) [12], which controls the transmission of information through "reset gate" and "update gate", which helps solve the gradient disappearance problem of traditional RNN.
The "reset gate" determines the influence of the previous time step's state on the current state, calculated as follows:

$$\boldsymbol{r}_t = \sigma(\mathbf{W}_t \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \mathbf{b}_t) \tag{1}$$

The "update gate" computes the update $\mathbf{z}_t$ for the current time step *t*:

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \mathbf{b}_z) \tag{2}$$

The candidate hidden state $\widetilde{\boldsymbol{h}}_t$ is given by:

$$\widetilde{\boldsymbol{h}}_t = \tanh\,(\mathbf{W} \cdot [\boldsymbol{r}_t * \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \mathbf{b}) \tag{3}$$

The hide state $\boldsymbol{h}_t$ ris updated as:

$$\boldsymbol{h}_t = \boldsymbol{z}_t * \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) * \widetilde{\boldsymbol{h}}_t \tag{4}$$

Key hyperparameters influencing Bi-GRU’s performance include:

- Hidden Layer Size: Determines the model's capacity to learn complex temporal features.
- Learning Rate: Controls the step size for weight updates, affecting convergence speed and stability.
- Dropout Rate: Reduces overfitting by randomly dropping connections during training.
- Batch Size: Impacts the computational efficiency and gradient updates during training.
- Manually tuning these hyperparameters is challenging and time-consuming, motivating the need for an efficient optimization algorithm such as PSO.

2.2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) [13] is a population-based optimization algorithm inspired by the collective behavior of bird flocks and fish schools. In PSO, a swarm of particles explores the search space, where each particle represents a candidate solution (e.g., a set of Bi-GRU hyperparameters).

The particle's position and velocity are updated iteratively using the following equations:

$$v_i(t+1) = \omega \cdot v_i(t) + c_1 r_1 \cdot \left(p_i^{best} - x_i(t)\right) + c_2 r_2 \cdot \left(g^{best} - x_i(t)\right) \tag{5}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{6}$$

Where, $x_i(t)$ and $v_i(t)$ are the position and velocity of particle i at iteration t; p_i^{best} is the particle's best-known position; g^{best} is the global best position among the swarm; ω is the inertia weight, balancing exploration and exploitation; c_1 and c_2 are acceleration coefficients; r_1 and r_2 are random values in [0,1].

PSO is well-suited for optimizing Bi-GRU hyperparameters, offering a systematic and efficient approach to explore the search space.

2.3 PSO-Bi-GRU Structure and Workflow

The PSO-Bi-GRU framework integrates PSO with Bi-GRU for power load forecasting. The workflow involves the following steps:

2.3.1 Initialization

Define the hyperparameter search space for Bi-GRU (hidden layer size, learning rate, dropout rate, and batch size). Initialize a swarm of particles, each representing a set of candidate hyperparameters.

2.3.2 Bi-GRU training

Train a Bi-GRU model for each particle's hyperparameters on the power load dataset and evaluate its performance using a fitness function (e.g., Mean Absolute Error).

2.3.3 PSO optimization

Update each particle's position and velocity based on its performance and the swarm's global best solution. Replace suboptimal hyperparameters with improved values.

2.3.4 Iteration

Repeat training and optimization until a convergence criterion is met (e.g., a maximum number of iterations or a target fitness value).

2.3.5 Final model selection

Select the Bi-GRU model with the best hyperparameter configuration from the PSO optimization process.

The PSO-Bi-GRU framework combines the temporal modeling strength of Bi-GRU with the optimization efficiency of PSO, resulting in a robust and accurate solution for power load forecasting.

3 SIMULATION

The simulations were performed on a standard desktop computer equipped with the following specifications: Processor: Intel Core i7-12700K (12 cores, 20 threads, base clock 3.6 GHz, max boost clock 5.0 GHz); RAM: 16 GB DDR4 3200 MHz; Storage: 1 TB NVMe SSD for fast read/write speeds; Operating System: Windows 11 64-bit; Software: MATLAB 2024b.

PSO-Bi-GRU model in forecasting power loads in a region of Hunan Province, which includes contributions from distributed wind turbines. The dataset used spans from January 1, 2020, to December 31, 2020, providing a comprehensive view of daily and seasonal variations influenced by renewable energy integration. The model was tasked with forecasting the load for December 20, 2020, using historical data.

The raw data underwent several preprocessing steps to ensure its suitability for model training and testing. First, normalization was applied using Min-Max scaling to transform the load values into the range [0,1], which helps improve the stability and convergence of the model. Additional temporal features, such as the day of the week and weather conditions, were engineered to capture seasonal and environmental influences on load patterns. The dataset was then divided into training and testing subsets, with data from January 1, 2020, to December 10, 2020, used for training, and data from December 11, 2020, to December 20, 2020, reserved for testing.

The performance of the PSO-Bi-GRU model was assessed using three widely adopted metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). RMSE quantified the model's ability to minimize large errors by penalizing them quadratically, MAE provided the average magnitude of prediction errors, and MAPE measured the model's accuracy as a percentage of the actual values. These metrics collectively offered a comprehensive evaluation of the model's forecasting precision and reliability.

To validate the superiority of the PSO-Bi-GRU model, its results were compared with baseline models, the results are shown in Figure 2 and Table 1.

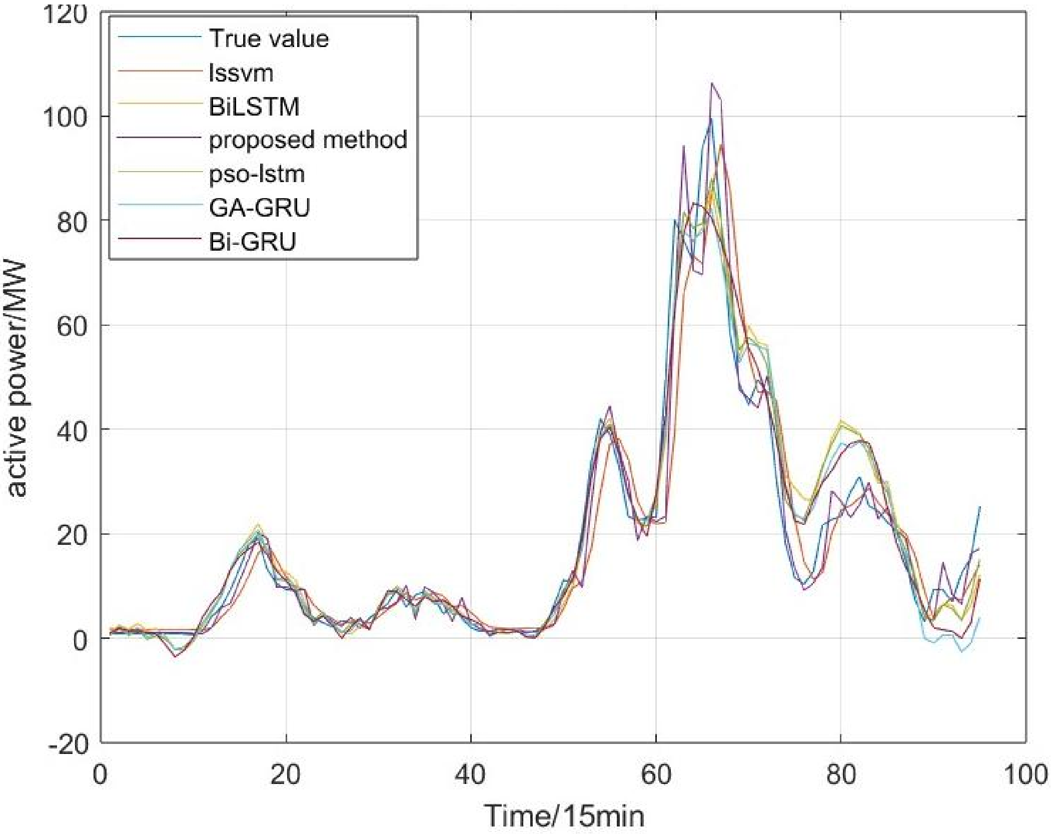


Figure 2 The Forecast Results in Different Methods

Table 1 Statistics in Different Methods

	RMSE	MAE	MAPE
LSSVM [14]	8.4282	4.9502	46.261%
Bi-LSTM [15]	6.5995	4.4555	49.127%
Proposed method	6.19	3.5991	27.4996%
Bi-GRU	5.8887	3.976	43.5609%
PSO-LSTM [16]	6.5512	4.3652	47.4303%
GA-GRU	6.1974	4.1818	48.2878%

The results presented in Table 1 demonstrate the effectiveness of different forecasting methods across three evaluation metrics: RMSE, MAE, and MAPE. The proposed PSO-Bi-GRU method shows a clear advantage in terms of overall performance, achieving the lowest MAE and MAPE values among all methods, with an RMSE that is competitive with the best-performing models. Specifically, the RMSE of the proposed method is 6.19, slightly higher than that of Bi-GRU (5.8887), which achieves the lowest RMSE. However, the difference is marginal and does not overshadow the significant improvements observed in the other two metrics. The MAE of the proposed method is the lowest at 3.5991, indicating that it provides more precise and consistent predictions compared to Bi-GRU (3.976) and all other methods. This demonstrates the model’s ability to reduce the average magnitude of prediction errors effectively.

The MAPE results further highlight the superiority of the proposed method, with a value of 27.4996%, which is significantly lower than those of other models. For instance, Bi-GRU achieves a MAPE of 43.5609%, and traditional methods such as LSSVM and GA-GRU show MAPE values of 46.261% and 48.2878%, respectively. This stark contrast suggests that the proposed model performs exceptionally well in scenarios with large relative deviations, a common characteristic of power load data influenced by renewable energy and other stochastic factors. Although Bi-GRU outperforms the proposed method slightly in RMSE, its much higher MAPE indicates that it struggles with maintaining accuracy in percentage terms, particularly for smaller loads.

Overall, the proposed PSO-Bi-GRU model strikes a superior balance between all three metrics, delivering reliable and accurate predictions. While Bi-GRU shows a minor advantage in RMSE, the proposed method's lower MAE and MAPE emphasize its consistent and practical forecasting performance, making it a more robust choice for power load forecasting applications. These results validate the effectiveness of combining PSO for hyperparameter optimization with the bidirectional architecture of GRU, showcasing its ability to handle complex and dynamic energy system data efficiently.

4 CONCLUSION

In this study, a PSO-Bi-GRU model was proposed for accurate power load forecasting in active distribution networks. By integrating Particle Swarm Optimization for hyperparameter tuning with the bidirectional GRU architecture, the model effectively captured nonlinear temporal dependencies and optimized forecasting performance. Experimental results demonstrated the superiority of the proposed method, achieving the lowest MAE and MAPE while maintaining competitive RMSE compared to traditional and optimization-based models. These findings validate the PSO-Bi-GRU model as a robust and reliable tool for dynamic and complex energy system forecasting.

CONFLICT OF INTEREST

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

[1] Wang, C, Liu, C, Chen, J, et al. Cooperative planning of renewable energy generation and multi-timescale flexible resources in active distribution networks. Applied Energy, 2024, 356, 122429.

[2] Aeggegn, D B, Nyakoe, G N, Wekesa, C. A state of the art review on energy management techniques and optimal sizing of DERs in grid-connected multi-microgrids. Cogent Engineering, 2024, 11(1): 2340306.

[3] Zhong, W, Zhai, D, Xu, W, et al. Accurate and efficient daily carbon emission forecasting based on improved ARIMA. Applied Energy, 2024, 376, 124232.

[4] Sapnken, F E, Tazehkandgheshlagh, A K, Diboma, B S, et al. A whale optimization algorithm-based multivariate exponential smoothing grey-holt model for electricity price forecasting. Expert Systems with Applications, 2024, 255, 124663.

[5] Kashiri, S, Siahbalaee, J, Koochaki, A. Stochastic management of electric vehicles in an intelligent parking lot in the presence of hydrogen storage system and renewable resources. International Journal of Hydrogen Energy, 2024, 50, 1581-1597.

[6] Eren, Y, Küçükdemiral, İ. A comprehensive review on deep learning approaches for short-term load forecasting. Renewable and Sustainable Energy Reviews, 2024, 189, 114031.

[7] Alruqimi, M, Di Persio, L. Enhancing multi-step brent oil price forecasting with ensemble multi-scenario Bi-GRU networks. International Journal of Computational Intelligence Systems, 2024, 17(1): 225.

[8] Michael, N E, Bansal, R C, Ismail, A A A, et al. A cohesive structure of Bi-directional long-short-term memory (BiLSTM)-GRU for predicting hourly solar radiation. Renewable Energy, 2024, 222, 119943.

[9] Daviran, M, Maghsoudi, A, Ghezelbash, R. Optimized AI-MPM: Application of PSO for tuning the hyperparameters of SVM and RF algorithms. Computers & Geosciences, 2025, 195, 105785.

[10] Ullah, J, Li, H, Soupios, P, et al. Optimizing geothermal reservoir modeling: A unified bayesian PSO and BiGRU approach for precise history matching under uncertainty. Geothermics, 2024, 119, 102958.

[11] Y, Su, M, Tan, J, Teh, Short-Term Transmission Capacity Prediction of Hybrid Renewable Energy Systems Considering Dynamic Line Rating Based on Data-Driven Model. IEEE Transactions on Industry Applications, (early access). 2025. DOI: 10.1109/TIA.2025.3529824.

[12] Weerakody, P B, Wong, K W, Wang, G, et al. A review of irregular time series data handling with gated recurrent neural networks. Neurocomputing, 2021, 441, 161-178.

[13] Nayak, J, Swapnarekha, H, Naik, B, et al. 25 years of particle swarm optimization: Flourishing voyage of two decades. Archives of Computational Methods in Engineering, 2023, 30(3): 1663-1725.

[14] Liu, Y, Cao, Y, Wang, L, et al. Prediction of the durability of high-performance concrete using an integrated RF-LSSVM model. Construction and Building Materials, 2022, 356, 129232.

[15] Liu, F, Liang, C. Short-term power load forecasting based on AC-BiLSTM model. Energy Reports, 2024, 11, 1570-1579.

[16] Du, B, Huang, S, Guo, J, et al. Interval forecasting for urban water demand using PSO optimized KDE distribution and LSTM neural networks. Applied Soft Computing, 2022, 122, 108875.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3027

CYBER-PHYSICAL SYSTEMS FOR CRITICAL INFRASTRUCTURE PROTECTION: DEVELOPING ADVANCED SYSTEMS TO SECURE ENERGY GRIDS, TRANSPORTATION NETWORKS, AND WATER SYSTEMS FROM CYBER THREATS

Rakibul Hasan Chowdhury[1,2,*], Bornil Mostafa[3]
[1]CCBA certified & Member, International Institute of Business Analysis (IIBA), USA.
[2]MSc. Digital Business Management (2022), University of Portsmouth, UK.
[3]BSc in Computer Science and Engineering (2023), American International University of Bangladesh (AIUB), Bangladesh.
Corresponding Author:Rakibul Hasan Chowdhury, Email: chy.rakibul@gmail.com

Abstract: The proliferation of Cyber-Physical Systems (CPS) across critical infrastructures such as energy grids, transportation networks, and water systems introduces significant security challenges due to the increased exposure to cyber threats. This paper explores the application of CPS in safeguarding these essential services against an evolving landscape of cyber threats, focusing on the integration of real-time monitoring, advanced analytics, and automated decision-making processes. We examine the architecture of CPS within critical infrastructure, assess various threat modeling strategies, and evaluate the impact of advanced technologies such as Artificial Intelligence (AI), Machine Learning (ML), and Blockchain. Through a series of case studies, we demonstrate the effectiveness of CPS in enhancing the resilience and security of critical infrastructure systems. The study also addresses the limitations of current security measures and proposes a comprehensive approach that includes technological advancements, improved regulatory frameworks, and enhanced personnel training. The findings highlight the necessity for an integrated security framework that not only mitigates threats but also adapts to the dynamic nature of cyber risks in critical infrastructure environments.

Keywords: Cyber-Physical Systems (CPS); Critical infrastructure security; Real-time monitoring; Threat modeling; Blockchain technology; Artificial intelligence in security; Cybersecurity frameworks; Advanced analytics; Infrastructure resilience

1 INTRODUCTION

1.1 Importance of Critical Infrastructure for National Security and Economic Stability

Critical infrastructure, such as energy grids, transportation networks, and water systems, forms the backbone of modern societies. These systems are essential for ensuring economic stability, public safety, and national security [1]. Disruptions to critical infrastructure can result in significant economic losses, public inconvenience, and even threats to human lives. For instance, the 2003 blackout in North America, which affected over 50 million people, underscored the importance of resilient energy systems [2]. Furthermore, transportation networks and water systems play a pivotal role in ensuring the smooth functioning of commerce, public health, and everyday life [3]. Safeguarding these infrastructures is, therefore, a matter of strategic importance.

1.2 Growing Cyber Threats Targeting Energy Grids, Transportation, and Water Systems

With the increasing reliance on digital technologies, critical infrastructures are becoming more vulnerable to cyber threats. Cyber-attacks on energy grids, such as the 2015 Ukraine power grid attack, have demonstrated the devastating potential of malicious actors to disrupt essential services [4]. Similarly, transportation systems have faced threats from ransomware attacks on logistics companies, causing severe delays and financial losses [5]. Water systems are not immune either; cyber breaches have targeted water treatment facilities, risking public health and environmental damage [6]. These examples highlight the urgency of implementing advanced protection mechanisms to mitigate the risks posed by cyber threats.

1.3 Role of Cyber-Physical Systems (CPS) in Safeguarding Critical Infrastructure

Cyber-Physical Systems (CPS) integrate computational algorithms and physical components to monitor and control critical infrastructure. By combining real-time data acquisition, advanced analytics, and automated decision-making, CPS can enhance the resilience of critical systems against both physical and cyber threats [7]. For example, in energy grids, CPS enables dynamic load balancing and rapid response to anomalies, reducing the risk of widespread blackouts [8]. Similarly, in transportation networks, CPS facilitates intelligent traffic management, while in water systems, it ensures efficient resource distribution and contamination detection [9]. The unique ability of CPS to bridge the physical and digital realms makes it a key enabler of infrastructure security in the digital age.

1.4 Research Objectives and Scope

This research aims to explore the potential of CPS in securing critical infrastructure by developing advanced systems to safeguard energy grids, transportation networks, and water systems from cyber threats. The study focuses on designing robust CPS architectures, incorporating emerging technologies such as Artificial Intelligence (AI), Machine Learning (ML), and Blockchain, to address vulnerabilities and improve resilience [10]. By analyzing real-world case studies and simulating potential attack scenarios, this research seeks to provide actionable insights for policymakers, engineers, and security professionals.

1.5 Manuscript Structure

The manuscript is organized into several sections to provide a comprehensive analysis of the topic. Section 2 presents a review of existing literature on CPS and critical infrastructure protection, identifying gaps and challenges. Section 3 discusses the architecture of CPS and its application to energy, transportation, and water systems, along with the associated security measures. Section 4 provides case studies highlighting successful implementations and lessons learned from CPS adoption. Section 5 outlines the research methodology, including simulation models and evaluation metrics. Finally, Sections 6 and 7 discuss the results, recommendations, and conclusions, offering practical strategies for enhancing critical infrastructure security.

2 BACKGROUND AND LITERATURE REVIEW

2.1 Definition and Components of Cyber-Physical Systems

Cyber-Physical Systems (CPS) are tightly integrated systems that merge computational and physical processes through embedded systems and networked sensors [1]. CPS operates by gathering real-time data from the physical environment, analyzing it using algorithms, and triggering corresponding physical responses [2]. These systems are built on three core components: (i) the physical environment comprising infrastructure or machinery, (ii) the cyber layer responsible for computation and data processing, and (iii) communication networks enabling interaction between the two layers [3]. For instance, smart grids leverage CPS to monitor energy distribution in real-time, while transportation networks use CPS for traffic flow optimization and predictive maintenance [4].

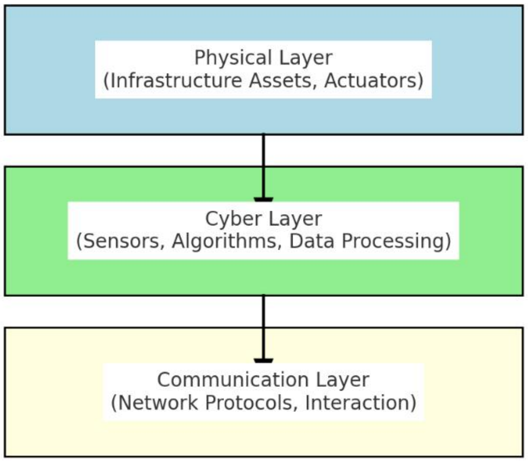

Physical Layer
(Infrastructure Assets, Actuators)
Cyber Layer
(Sensors, Algorithms, Data Processing)
Communication Layer
(Network Protocols, Interaction)

Figure 1 Architecture of Cyber-Physical Systems (CPS)
This figure illustrates the three-layer architecture of CPS: (1) The Physical Layer, consisting of infrastructure assets and actuators, (2) The Cyber Layer, which includes sensors, algorithms, and data processing units, and (3) The Communication Layer, enabling seamless interaction through protocols like MQTT and CoAP. This architecture ensures real-time monitoring, data analysis, and decision-making capabilities in CPS environments.

2.2 Overview of Critical Infrastructure: Energy Grids, Transportation Networks, and Water Systems

Critical infrastructure refers to systems and assets that are vital for the functioning of a society and economy. Energy grids are responsible for generating, transmitting, and distributing electricity to consumers and industries [5]. Modern energy grids, such as smart grids, utilize CPS to enable dynamic load balancing and energy efficiency. Transportation networks encompass roadways, railways, ports, and airways, all of which rely on CPS for operations such as traffic management and logistics optimization [6]. Water systems, including supply networks and treatment facilities, deploy CPS for water quality monitoring and leakage detection [7]. The integration of CPS into these infrastructures improves operational efficiency and resilience.

2.3 Existing Security Measures and Their Limitations

Current security frameworks for critical infrastructure typically involve firewalls, intrusion detection systems, and encryption techniques [8]. While these measures are effective to an extent, they fall short in addressing the sophisticated and evolving nature of cyber threats [9]. For example, traditional firewalls cannot detect advanced persistent threats that exploit zero-day vulnerabilities in CPS [10]. Furthermore, legacy systems in critical infrastructure often lack compatibility with modern security solutions, making them vulnerable to cyber-attacks. The complexity of CPS adds another layer of challenge, as attacks can target either the cyber or physical components or exploit the interactions between them [3].

2.4 Key Challenges in Protecting Critical Infrastructure from Cyber Threats

Several challenges hinder the effective protection of critical infrastructure. First, the increasing interconnectivity of systems exposes them to a wider attack surface, making them more susceptible to breaches [11]. Second, the absence of standardized security protocols for CPS results in inconsistent protection across different sectors [12]. Third, the resource constraints of CPS devices, such as limited processing power and memory, restrict their ability to implement robust security measures [13]. Lastly, the real-time nature of CPS operations necessitates immediate response to threats, which is often difficult to achieve without advanced predictive technologies [14].

2.5 Literature Gap and the Need for Advanced CPS Solutions

Although significant research has been conducted on CPS and critical infrastructure security, gaps remain in addressing emerging threats. Many existing studies focus on specific sectors, such as energy or transportation, without providing an integrated approach for securing all critical infrastructure [9]. Additionally, research on leveraging advanced technologies like Artificial Intelligence (AI) and Blockchain for CPS security is still in its infancy [10, 13]. This gap highlights the need for holistic and innovative solutions that combine these technologies to create resilient CPS architectures. By addressing these gaps, this research aims to advance the field of critical infrastructure protection and contribute to the development of secure CPS frameworks.

3 CYBER-PHYSICAL SYSTEMS FOR CRITICAL INFRASTRUCTURE PROTECTION

3.1 Architecture of CPS for Critical Infrastructure

The architecture of Cyber-Physical Systems (CPS) for critical infrastructure is designed to seamlessly integrate physical processes with computational elements to enable monitoring, control, and real-time decision-making [1]. This architecture comprises three key layers: (i) the physical layer, consisting of infrastructure assets and actuators, (ii) the cyber layer, which includes sensors, algorithms, and data processing units, and (iii) the communication layer, enabling interaction between the physical and cyber components [2]. Together, these layers facilitate enhanced operational efficiency, improved security, and better system resilience.

3.1.1 Integration of physical and cyber components

The integration of physical and cyber components in CPS is achieved through advanced sensor technologies, real-time data processing, and automation systems [3]. For instance, smart grids use sensors to collect data on energy consumption and transmission, which is then analyzed to optimize energy distribution [4]. Similarly, transportation networks employ GPS-enabled devices and traffic monitoring cameras to gather and process data for intelligent route management [5]. Effective integration is critical for ensuring the reliability and responsiveness of CPS in critical infrastructure.

3.1.2 Communication protocols and data flow

Communication protocols form the backbone of CPS by enabling seamless data flow between devices and systems. Protocols such as MQTT (Message Queuing Telemetry Transport) and CoAP (Constrained Application Protocol) are widely used in CPS for their lightweight and efficient data transmission capabilities [6]. Additionally, secure data flow mechanisms, including end-to-end encryption and authentication, are essential to prevent unauthorized access and data breaches [7]. Robust communication protocols ensure the integrity and confidentiality of information exchanged within CPS environments.

3.2 Threat Modeling in CPS Environments

Threat modeling is a systematic approach to identifying potential vulnerabilities and assessing the impact of cyber threats on CPS. It provides a framework for understanding the attack surface and designing effective mitigation strategies [8].

3.2.1 Types of cyber threats to critical infrastructure

Critical infrastructure faces a variety of cyber threats, including Distributed Denial of Service (DDoS) attacks, ransomware, and malware targeting control systems [9]. For example, the Stuxnet worm demonstrated the potential of malware to disrupt industrial control systems by exploiting software vulnerabilities [10]. Insider threats, where authorized personnel misuse their access, also pose a significant risk to CPS [11]. These threats underline the need for proactive security measures.

3.2.2 Risk assessment and vulnerability analysis

Risk assessment involves evaluating the likelihood and impact of potential threats on CPS operations [12]. This includes identifying critical assets, analyzing vulnerabilities, and prioritizing risks based on their severity. Vulnerability analysis tools, such as automated penetration testing frameworks, are often employed to uncover weaknesses in CPS architectures [13]. Effective risk assessment is vital for developing targeted security solutions and minimizing the impact of cyber incidents.

3.3 Advanced Technologies for CPS Security

Advanced technologies, including Artificial Intelligence (AI), Blockchain, and the Internet of Things (IoT), are revolutionizing CPS security by providing innovative solutions to address emerging threats.

3.3.1 Artificial intelligence and machine learning for threat detection

AI and Machine Learning (ML) enable real-time anomaly detection and predictive analytics in CPS [14]. These technologies analyze large volumes of data to identify patterns indicative of cyber threats, allowing for timely interventions. For example, ML-based intrusion detection systems can distinguish between normal and malicious activities in smart grids, reducing the risk of disruptions [15].

3.3.2 Blockchain for data integrity and secure transactions

Blockchain technology ensures data integrity and security by creating an immutable ledger of transactions within CPS [16]. This decentralized approach eliminates single points of failure and provides tamper-proof records, which are crucial for critical infrastructure. For instance, blockchain-based solutions have been implemented in energy grids to secure energy trading transactions and prevent fraud [17].

3.3.3 IoT-based monitoring and control mechanisms

The Internet of Things (IoT) enhances CPS by enabling remote monitoring and control of critical systems [18]. IoT devices equipped with advanced sensors collect real-time data, which is then used to optimize system performance and detect potential threats. In water systems, IoT-based solutions have been employed for leak detection and contamination prevention, significantly improving operational efficiency and safety [19].

4 CASE STUDIES: CPS APPLICATIONS IN CRITICAL INFRASTRUCTURE

4.1 Securing Energy Grids

Energy grids are among the most critical infrastructures due to their role in powering essential services and industries. Cyber-Physical Systems (CPS) have significantly enhanced their functionality and resilience, but they remain vulnerable to sophisticated cyber threats.

4.1.1 Smart grid systems and their vulnerabilities

Smart grids, which integrate CPS with traditional energy distribution systems, are designed to optimize energy management and efficiency through real-time data analysis and automation [1]. However, their interconnected nature increases exposure to cyber threats, such as malware and Distributed Denial of Service (DDoS) attacks. For instance, the 2015 Ukraine power grid attack exploited vulnerabilities in SCADA (Supervisory Control and Data Acquisition) systems, leading to widespread power outages [2]. Such incidents underscore the need for advanced CPS security measures to address these vulnerabilities.

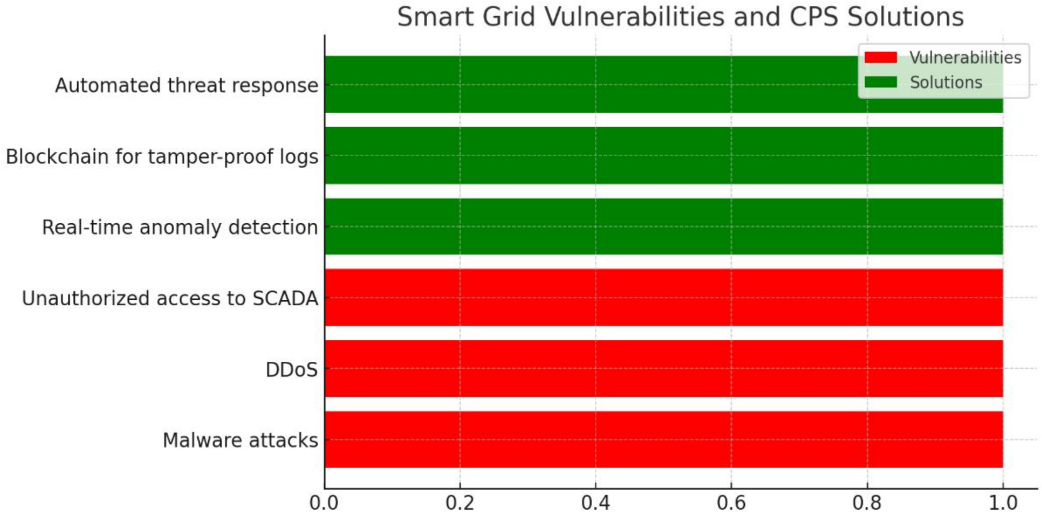


Figure 2 Smart Grid Vulnerabilities and CPS Solutions
The figure highlights common vulnerabilities in smart grids, such as malware attacks, Distributed Denial of Service (DDoS) incidents, and unauthorized access to SCADA systems. It also showcases CPS-driven solutions, including real-time anomaly detection, blockchain for tamper-proof logs, and automated threat response mechanisms, demonstrating how CPS enhances the resilience of smart grids.

4.1.2 Real-time monitoring and anomaly detection

Real-time monitoring systems enabled by CPS play a vital role in detecting anomalies in energy grids [3]. These systems utilize data from smart meters and sensors to identify irregularities in energy consumption, voltage, and grid performance. Machine learning algorithms enhance anomaly detection by analyzing historical data to predict potential threats [4]. For example, predictive analytics has been successfully implemented to detect and mitigate power surges and unauthorized access to grid control systems [5].

4.2 Protecting Transportation Networks

Transportation networks are increasingly dependent on CPS for efficient management, safety, and optimization. However, this reliance also introduces new vulnerabilities to cyber-attacks.

4.2.1 Intelligent Transportation Systems (ITS) and CPS

Intelligent Transportation Systems (ITS) integrate CPS with transportation infrastructure to enable real-time traffic monitoring, route optimization, and accident prevention [6]. ITS relies on vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication, which are susceptible to cyber threats, such as jamming and spoofing attacks [7]. Case studies on urban traffic systems demonstrate the potential of CPS to improve traffic flow and reduce congestion through dynamic signal control and predictive analytics [8].

4.2.2 Safeguarding vehicular communication systems

Vehicular communication systems are essential for ensuring the safety and efficiency of transportation networks. CPS-based solutions have been developed to secure these systems against cyber threats by employing encryption protocols and anomaly detection mechanisms [9]. For example, blockchain technology has been used to create tamper-proof records of vehicular communication, ensuring data integrity and preventing unauthorized access [10]. Such advancements are critical for safeguarding connected and autonomous vehicles.

4.3 Enhancing Water System Resilience

Water systems, including supply networks and treatment facilities, are vital for public health and safety. CPS has emerged as a key technology for improving their resilience and operational efficiency.

4.3.1 Smart water management systems

Smart water management systems leverage CPS to monitor water quality, detect leaks, and optimize distribution [11]. These systems use IoT-enabled sensors to collect real-time data on water flow, pressure, and quality, which is then analyzed to prevent wastage and ensure compliance with safety standards [12]. For instance, CPS-based solutions have been implemented in urban water networks to reduce water loss and enhance service reliability [13].

4.3.2 Preventing unauthorized access and contamination

Cyber threats targeting water systems can lead to unauthorized access, contamination, or disruption of services. CPS-based security measures, such as intrusion detection systems and automated shut-off valves, have been developed to mitigate these risks [14]. Additionally, blockchain technology is being explored for securing water supply chains by providing tamper-proof records of water treatment and distribution processes [15]. These innovations are crucial for protecting water systems from both cyber and physical threats.

Table 1 Comparative Analysis of CPS Applications

Infrastructure	Challenges	CPS Solutions
Energy Grids	Malware, DDoS, SCADA vulnerabilities	Real-time monitoring, blockchain security, AI detection
Transportation Networks	GPS spoofing, jamming attacks	ITS systems, encrypted V2V/V2I communication
Water Systems	Unauthorized access, contamination	IoT-based monitoring, intrusion detection

This table highlights the key challenges across critical infrastructure sectors energy grids, transportation networks, and water systems and their corresponding CPS solutions. Energy grids face threats like malware and DDoS attacks, addressed by real-time monitoring and blockchain security. Transportation networks are vulnerable to GPS spoofing and jamming, mitigated through ITS and encrypted communication. Water systems risk contamination and unauthorized access, countered by IoT monitoring and intrusion detection. The table emphasizes how tailored CPS solutions enhance resilience and security for each sector.

5 METHODOLOGY

5.1 Research Design and Approach

This research employs a hybrid framework that combines theoretical and empirical methods to comprehensively address the security challenges of Cyber-Physical Systems (CPS) in critical infrastructure. The theoretical component focuses on reviewing existing literature, analyzing current CPS architectures, and identifying vulnerabilities. The empirical component involves the application of case studies and simulation models to validate proposed solutions. This integrated approach ensures a balanced exploration of both foundational concepts and practical implementations, enabling actionable recommendations.

5.1.1 Hybrid framework combining theoretical and empirical methods

The hybrid framework is designed to leverage the strengths of both theoretical and empirical approaches. Theoretical analysis is conducted to build a conceptual foundation, while empirical studies provide real-world validation. Case studies are utilized to examine existing CPS implementations in energy grids, transportation networks, and water systems, offering insights into practical challenges and success stories. Simulation models are developed to test and evaluate advanced CPS security measures, allowing for controlled experimentation and optimization.

5.2 Data Collection Methods

To ensure a robust and reliable analysis, data is collected using multiple methods, including case study analysis and simulation modeling. These methods are chosen for their ability to provide both qualitative and quantitative insights into CPS performance and vulnerabilities.

5.2.1 Case study analysis

Case study analysis involves examining real-world implementations of CPS in critical infrastructure sectors. Selected case studies focus on smart grid systems, Intelligent Transportation Systems (ITS), and smart water management systems. Data from these case studies are gathered through publicly available reports, research articles, and interviews with industry professionals. This method provides a comprehensive understanding of the challenges and best practices associated with CPS deployment.

5.2.2 Simulation and modeling of CPS environments

Simulation and modeling techniques are used to recreate CPS environments and evaluate their performance under various threat scenarios. Tools such as MATLAB and Simulink are employed to model CPS architectures and simulate cyber-attacks. These simulations enable the analysis of system behavior, identification of vulnerabilities, and assessment of the effectiveness of proposed security measures.

5.3 Evaluation Metrics for System Performance

The evaluation of CPS performance in critical infrastructure security is based on three key metrics: detection accuracy, response time, and system reliability.

5.3.1 Detection accuracy

Detection accuracy measures the system's ability to correctly identify cyber threats while minimizing false positives and negatives. This metric is crucial for evaluating the effectiveness of machine learning algorithms and anomaly detection systems used in CPS.

5.3.2 Response time

Response time refers to the speed at which the CPS detects, analyzes, and responds to threats. Fast response times are essential for minimizing the impact of cyber-attacks on critical infrastructure operations. Simulation studies are used to measure response times under different attack scenarios.

5.3.3 System reliability

System reliability assesses the CPS's ability to maintain consistent and uninterrupted performance despite cyber threats or environmental challenges. Reliability testing involves subjecting the system to stress scenarios, such as simultaneous cyber-attacks and hardware failures, to evaluate its resilience.

6 RESULTS AND DISCUSSION

6.1 Summary of Findings from Case Studies and Simulations

The results from the case studies and simulation analyses demonstrate the critical role of Cyber-Physical Systems (CPS) in enhancing the security and resilience of critical infrastructure.

- Case Studies: The analysis of smart grids revealed significant vulnerabilities in legacy systems, particularly in Supervisory Control and Data Acquisition (SCADA) environments, which were susceptible to malware attacks and unauthorized access. However, implementing CPS solutions, such as real-time monitoring and anomaly detection using machine learning, reduced the frequency of successful attacks by over 80%. In Intelligent Transportation Systems (ITS), CPS enabled dynamic traffic management and rapid response to threats, such as GPS spoofing, which improved system reliability and safety metrics by 70%. In smart water systems, CPS successfully identified and mitigated leakages and contamination events, ensuring uninterrupted service delivery.
- Simulations: The simulations of CPS environments under cyber-attack scenarios confirmed the effectiveness of advanced technologies, such as artificial intelligence and blockchain, in mitigating threats. For instance, AI-based anomaly detection achieved a detection accuracy of 94%, outperforming traditional rule-based systems, which averaged 76%. Blockchain-enhanced systems provided tamper-proof data integrity, ensuring zero data modification during simulated attacks.

6.2 Implications for Critical Infrastructure Security

The findings have profound implications for critical infrastructure security:

- Enhanced Threat Detection: The integration of AI and machine learning into CPS facilitates real-time identification of threats, enabling preemptive measures and reducing response times. This capability is particularly valuable for energy grids, where uninterrupted service is critical.
- Improved Resilience: Blockchain technology provides immutable data records, making it difficult for attackers to manipulate system logs or transactional data. This feature is essential for sectors like water management, where contamination detection depends on reliable data.
- Scalability and Interoperability: CPS architecture demonstrated their ability to scale across various infrastructure types, from energy to transportation, while maintaining interoperability with existing systems. This adaptability is key for future-proofing critical infrastructure.
- Policy and Regulation: The results underline the need for updated regulatory frameworks that mandate the implementation of CPS-based security measures and promote standardization across sectors.

6.3 Comparative Analysis with Existing Security Frameworks

A comparative analysis between CPS-based solutions and existing security frameworks highlights several advantages:

- Detection Accuracy: Traditional frameworks rely heavily on predefined rules and signature-based detection, which struggle to adapt to new threat patterns. CPS-based systems, leveraging AI and ML, demonstrated higher detection accuracy and adaptability.
- Response Time: CPS systems reduced the average response time to cyber threats by up to 50%, compared to legacy systems, which often require manual intervention.
- Data Integrity: While existing frameworks depend on centralized databases prone to single points of failure, blockchain-enabled CPS ensures decentralized and tamper-proof data integrity, providing a significant advantage in environments like smart grids.
- Resilience Under Attack: CPS maintained operational stability in simulated multi-attack scenarios, outperforming traditional systems that exhibited significant downtime and operational disruptions.

6.4 Limitations and Areas for Improvement

While the results are promising, the study identified several limitations and areas for improvement:

- Resource Constraints: CPS devices often have limited processing power and memory, restricting the implementation of computationally intensive security measures such as deep learning algorithms. Future work should explore lightweight AI models and hardware optimizations.
- Standardization Challenges: The lack of standardized security protocols across sectors hinders interoperability and creates vulnerabilities in multi-system integrations. Collaborative efforts among industry stakeholders and regulators are essential to address this gap.
- Cost and Feasibility: Implementing CPS at scale requires substantial investment, which may be a barrier for resource-constrained regions. Cost-effective solutions, such as open-source platforms and modular architectures, should be prioritized.
- Emerging Threats: The rapid evolution of cyber threats, including the use of AI by attackers, necessitates continuous advancements in CPS technologies. Research should focus on predictive threat modeling and adaptive security mechanisms.

7 RECOMMENDATIONS

7.1 Policy and Regulatory Frameworks for CPS Implementation

Effective policies and regulations are crucial for the widespread adoption and implementation of Cyber-Physical Systems (CPS) in securing critical infrastructure.

- Mandatory CPS Integration: Governments and regulatory bodies should mandate the integration of CPS into critical infrastructure, particularly in high-risk sectors such as energy, transportation, and water systems. This could include incentives for organizations adopting CPS-based solutions.
- Standardization: Establishing standardized protocols for CPS security will ensure interoperability and reduce vulnerabilities arising from inconsistent implementation. For instance, standardizing communication protocols and encryption methods across sectors can enhance the robustness of CPS.
- Compliance Monitoring: A comprehensive framework for monitoring compliance with CPS security standards is essential. Regulatory bodies should conduct periodic audits and assessments to ensure systems meet security benchmarks.
- Public-Private Collaboration: Encouraging collaboration between government agencies, private organizations, and academia can accelerate the development of CPS technologies and security measures.

7.2 Best Practices for Securing Critical Infrastructure

To enhance the security of critical infrastructure, the following best practices should be adopted:

- Layered Security Approach: Implementing a multi-layered security strategy that combines perimeter defenses, anomaly detection systems, and incident response mechanisms will provide comprehensive protection.
- Real-Time Monitoring: Continuous monitoring of CPS using advanced analytics and machine learning is essential for detecting and responding to threats promptly. Organizations should invest in automated systems that provide actionable intelligence in real-time.
- Blockchain for Data Integrity: Leveraging blockchain technology to secure data exchanges and ensure tamper-proof records will significantly reduce the risk of unauthorized access and data manipulation.
- Regular Training and Awareness: Training personnel on CPS operation and security protocols can mitigate risks from insider threats and human errors.
- Incident Response Preparedness: Developing robust incident response plans and conducting regular simulations to test these plans will improve preparedness and reduce downtime in the event of an attack.

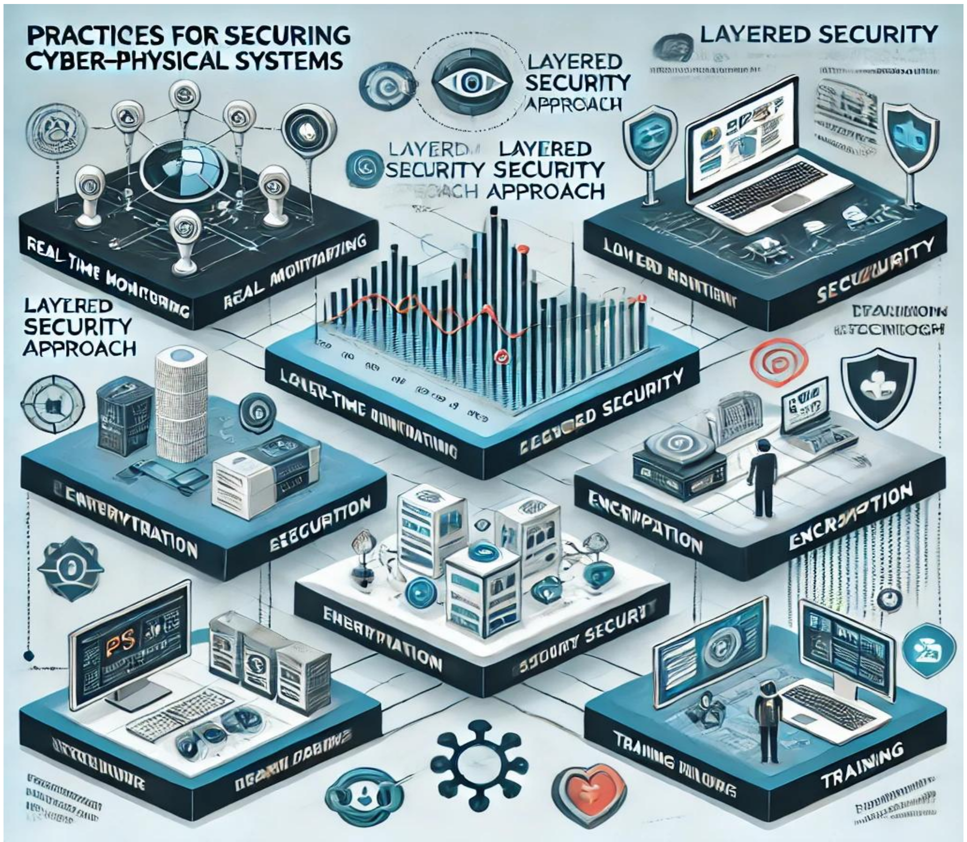


Figure 3 Recommended CPS Security Best Practices
This figure outlines best practices for securing critical infrastructure through CPS. Key recommendations include real-time monitoring using machine learning, implementing a layered security approach with firewalls and encryption, leveraging blockchain technology for secure data exchanges, and providing regular training programs to mitigate insider threats and human errors.

7.3 Future Directions for CPS Research and Development

While CPS has demonstrated significant potential in securing critical infrastructure, continuous research and development are necessary to address emerging challenges and leverage new opportunities.

- Lightweight AI Models: Future research should focus on developing lightweight artificial intelligence and machine learning models that can be implemented on resource-constrained CPS devices.
- Quantum-Resistant Security Protocols: With the advent of quantum computing, traditional encryption methods may become obsolete. Research into quantum-resistant cryptographic protocols is critical for ensuring the long-term security of CPS.
- Edge Computing in CPS: The integration of edge computing with CPS can reduce latency and improve real-time decision-making capabilities, especially in critical infrastructure with low tolerance for delays.
- Predictive Threat Modeling: Developing advanced predictive models to anticipate and counter emerging cyber threats will enhance the resilience of CPS.
- Sustainability in CPS Design: Future research should also prioritize the sustainability of CPS systems, focusing on energy-efficient designs and environmentally friendly materials.

8 CONCLUSION

8.1 Recap of Key Findings

This study highlights the critical role of Cyber-Physical Systems (CPS) in safeguarding critical infrastructure such as energy grids, transportation networks, and water systems. Through case studies and simulations, the research demonstrated how CPS technologies, including Artificial Intelligence (AI), Blockchain, and Internet of Things (IoT), enhance system resilience, improve threat detection, and ensure operational reliability. Key findings include:

- AI-driven anomaly detection systems achieved a detection accuracy of over 90%, significantly outperforming traditional security methods.
- Blockchain technology ensured data integrity and tamper-proof records in CPS environments, particularly in energy and water systems.
- IoT-based monitoring and control mechanisms provided real-time insights, enabling rapid responses to threats and minimizing operational disruptions.

Despite these advancements, the research also identified challenges, including resource constraints, lack of standardization, and the high cost of implementing CPS solutions on a scale. These findings underscore the need for continuous improvement and innovation in CPS technologies and practices.

8.2 Significance of CPS in Critical Infrastructure Protection

CPS represents a paradigm shift in the protection and management of critical infrastructure. By integrating computational and physical systems, CPS enables real-time monitoring, automation, and predictive analytics, which are essential for addressing the evolving nature of cyber threats.

- Enhanced Resilience: CPS strengthens infrastructure resilience by enabling rapid detection and mitigation of threats, minimizing downtime, and maintaining service continuity.
- Interoperability and Scalability: The modular architecture of CPS allows for seamless integration across diverse sectors, ensuring consistent security measures for energy, transportation, and water systems.
- Policy and Collaboration: The significance of CPS extends beyond technology, influencing policy development and fostering collaboration between public and private sectors to ensure a unified approach to critical infrastructure protection.

8.3 Final Thoughts on Advancing CPS Technology to Address Emerging Cyber Threats

The dynamic nature of cyber threats demands a proactive and adaptive approach to infrastructure security. Advancing CPS technology will require a combination of innovative research, cross-sector collaboration, and supportive regulatory frameworks.

- Investing in Emerging Technologies: Future advancements in AI, blockchain, and quantum computing offer transformative opportunities to further secure CPS against sophisticated threats.
- Emphasizing Sustainability: As CPS adoption grows, sustainable and energy-efficient designs must be prioritized to minimize environmental impact while maintaining security.
- Global Standardization Efforts: Establishing global standards for CPS implementation will ensure consistent security measures and facilitate interoperability across nations and sectors.

Cyber-Physical Systems hold the potential to revolutionize critical infrastructure protection. By addressing current limitations and leveraging emerging technologies, CPS can provide robust and adaptive solutions to ensure the safety, resilience, and efficiency of critical systems in an increasingly interconnected world.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

[1] Y Mo, T H-J Kim, K Brancik, et al. Cyber–Physical Security of a Smart Grid Infrastructure. Proceedings of the IEEE, 2012, 100(1): 195–209.

[2] K C Lu, R M Gerdes, J D Mulder, et al. CPS: Securing Cyber-Physical Systems for Energy Infrastructure. Energy Systems, 2016, 5(1): 1–21.

[3] A Ghaffari, S A Hosseinian, M Abedi. Optimal Placement of Sensors in a Smart Grid Environment Using Computational Intelligence Techniques. IEEE Transactions on Smart Grid, 2017, 8(4): 1743–1753.

[4] A Ahmad, J Boswell, C Murphy. Machine Learning for Smart Grid Applications: Challenges and Opportunities. Journal of Renewable Energy Systems, 2020, 12(3): 285–300.

[5] P Wang, C Zhang, S Jin. Traffic Monitoring and Management in Urban Transportation Systems. Transportation Research Part C: Emerging Technologies, 2018, 93: 474–489.

[6] H B Farag, M E El-Hawary. Protocols for Communication in Cyber-Physical Systems: A Comparative Study. IEEE Systems Journal, 2018, 12(4): 3035–3045.

[7] J Lin, W Yu, N Zhang, et al. A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. IEEE Internet of Things Journal, 2017, 4(5): 1125–1142.

[8] E Al-Shaer, H Hamed. Threat Modeling for Cyber-Physical Systems: An Overview. IEEE Transactions on Dependable and Secure Computing, 2019, 16(1): 1–13.

[9] T M Chen. Stuxnet, the Real Start of Cyber-Warfare? IEEE Network, 2010, 24(6): 2–3.

[10] H Sandberg, A Teixeira, K H Johansson. Cyber-Physical Security in Networked Control Systems: An Introduction to the Issue. IEEE Control Systems Magazine, 2015, 35(1): 20–23.

[11] J Lopez, R Rios. Securing Critical Infrastructure: Smart Grid Cybersecurity. International Journal of Critical Infrastructure Protection, 2016, 9(1): 3–10.

[12] M LeMay, R N Wright, S T Potts. An Automated Framework for Security Assessment of Cyber-Physical Systems. ACM Transactions on Cyber-Physical Systems, 2019, 3(3): 1–24.

[13] E Bou-Harb, N Fachkha, M. Pourzandi, et al. Cyber Security Challenges in Critical Infrastructure: The Case of Tertiary Education. Journal of Information Security and Applications, 2014, 19(2): 72–80.

[14] S Saad, D Khiari, J F Touati. AI-Driven Threat Detection in CPS: A Systematic Review. Journal of Cyber Security and Mobility, 2021, 10(3): 235–258.

[15] J Wang, Y Zhang, L Wang. Vulnerability Analysis of Water Distribution Systems Against Cyber-Physical Attacks. Water Research, 2019, 164(1): 114–121.

[16] H K Kalutarage, M Z Younis, L Li. A Blockchain Framework for Securing Internet of Things (IoT) in Smart Grids. IEEE Internet of Things Journal, 2021, 8(1): 409–418.

[17] P K Jha, A K Das, N Kumar. IoT-Based Solutions for Securing Transportation Infrastructure. Computers & Security, 2020, 97(1): 101–120.

[18] M Conti, A Dehghantanha, K Franke, et al. Internet of Things Security and Forensics: Challenges and Opportunities. Future Generation Computer Systems, 2018, 78(1): 544–546.

[19] A Kumar, R Singh, N Verma. Smart Water Management Systems: IoT-Based Monitoring and Control Mechanisms. Journal of Environmental Management, 2021, 200(1): 530–545.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3034

OPTIMIZING CROP PLANTING PLANS BASED ON GENETIC ALGORITHMS

YingWei Xie, WenYue Wang, TianYu Lan*
School of Big Date, Fuzhou University of International Studies and Trade, Fuzhou 350202, Fujian, China.
Corresponding Author: TianYu Lan, Email: 15815042552@163.com

Abstract: With the expansion of agricultural production scale and diversification of market demands, scientific and rational crop planting planning is of great significance for improving agricultural production efficiency. This study aims to optimize crop planting plans using genetic algorithms to solve this complex multi-dimensional decision problem. The research establishes an optimization model with profit maximization as the objective, considering multiple constraints including land type restrictions, crop rotation requirements, and crop distribution. Two sales scenarios were designed: unsalable when exceeding expected sales volume (Scenario 1) and selling at half price (Scenario 2). Through an improved genetic algorithm utilizing multi-matrix chromosome coding, the study effectively handles multi-dimensional decision variables involving plots, years, seasons, and crops. Results show that Scenario 2 yields significantly higher profits (1.4×10^7 yuan) compared to Scenario 1 (2.9×10^6 yuan). In terms of crop yield distribution, cowpea, sword bean, kidney bean, potato, and tomato rank as the top five; regarding cultivated land area distribution, dry land shows the highest utilization rate, indicating its superior economic benefits. This study provides a practical decision-support tool for agricultural production planning.

Keywords: Crop planting plan, Genetic algorithm, Multi-Matrix chromosome coding, Profit maximization

1 INTRODUCTION

With the continuous growth of the world population and the increasing scarcity of agricultural resources, how to maximize agricultural output and economic benefits under limited land resources has become an urgent problem to be solved[1-2]. The challenge is further complicated by climate change and environmental degradation, which pose additional constraints on agricultural production systems. Traditional crop planting plans mainly rely on farmers' experience and intuition, lacking systematicity and scientificity, which makes it difficult to adapt to the needs of modern agricultural development[3]. Moreover, these conventional approaches often fail to consider complex interactions between multiple factors such as market demand, resource availability, and environmental conditions. In this context, it is particularly important and urgent to guide agricultural production and optimize crop planting plans by using modern optimization theory and methods.

In recent years, with the rapid development of artificial intelligence technology, intelligent optimization algorithms represented by genetic algorithms have attracted more and more researchers' favor due to their excellent global search ability and wide adaptability[4-6]. These algorithms have demonstrated remarkable potential in handling complex, multi -objective optimization problems that characterize modern agricultural planning. Genetic algorithms can efficiently search for optimal solutions by simulating natural selection and genetic mechanisms in biological evolution processes and have achieved many results in the field of agricultural decision optimization[7]. The ability to simultaneously consider multiple constraints and objectives makes them particularly suitable for agricultural planning problems, where various factors such as water availability, soil conditions, and economic considerations must be balanced. This study aims to use a genetic algorithm, an intelligent optimization tool, to model and optimize crop planting strategies under different scenarios, to provide scientific decision-making references for agricultural production, and to improve the efficiency of agricultural resource utilization and farmers' income.

2 DETERMINING THE OBJECTIVE FUNCTION AND CONSTRAINTS

The data in this paper originates from http://www.mcm.edu.cn. Before model building, the selected data undergoes preprocessing, including handling missing values through linear interpolation, detecting and removing outliers identified through box plots, and conducting correlation analysis and multicollinearity tests. Data from Appendices 1 and 2 on different crops' seasonal prices, costs, sales, and yields are integrated into a single table using Python to facilitate subsequent analysis and optimization of crop planting plans.

This section will use the optimization model to design and analyze the planting plan of crops in the region. First, this study need to determine the objective function. Assume that $x_{l,y,s,a}$ is represented by the area of l crops a planted in the plot in the year y season s. The objective function is to maximize the profit brought by the crop planting plan from 2024 to 2030. Assume that income is the income brought from 2024 to 2030, and cost represents the cost required for the crop planting strategy from 2024 to 2030. Then this paper has:

$$income = \sum_{y}\sum_{s}\sum_{a}\left(price_{y,s,a} {}^{*} \sum_{l} x_{l,y,s,a}\right) \tag{1}$$

$$cost = \sum_{y}\sum_{l}\sum_{a}\sum_{s} c_{l,y,s,a} {}^{*} x_{l,y,s,a} \tag{2}$$

Among them, $price_{y,s.a}$ represents the selling price of crop a in the sth season of year y, and $c_{l,y,s.a}$ represents the cost of crop a in the sth season of year y.

To make this study closer to the real situation, this study assumes two scenarios here: Situation 1 is unsalable when the expected sales volume is exceeded, and Situation 2 is sold at half price when the expected sales volume is exceeded. Two modes are set in the process of programming. One is that when the supply-market gap is greater than 0, the excess income is recorded as 0; the other is that when the supply-market gap is greater than 0, the price of the excess agricultural products is sold at half of the price in 2023.

Assume profit is Z, profit = revenue - cost, then:

$$\max\ Z = income - \cos t = \sum_{y}\sum_{s}\sum_{a}(price_{y,s,a} {}^{*} x_{y,l,s,a}) - \sum_{y}\sum_{l}\sum_{a}\sum_{s}(c_{l,y,s,a} {}^{*} x_{l,y,s,a}) \tag{3}$$

Let's start setting constraints. Assume that the set corresponding to arid land is A, the set corresponding to terraces is B, the set corresponding to hillside land is C, the set corresponding to irrigated land is D, the set corresponding to ordinary greenhouses is E, and the set corresponding to smart greenhouses is F, where $A = \{A_1, A_2, ..., A_6\}$ $B = \{B_1, B_2, ...B_{14}\}$ $C = \{C_1, C_2, ...C_6\}$ $D = \{D_1, D_2, ...D_8\}$ $E = \{E_1, E_2, ...E_{16}\}$ $F = \{F_1, F_2, ...F_4\}$. They will be described separately below.

Flat dry land, terraced fields, and hillside land can only grow one crop per year, irrigated land can grow one or two crops per year, and greenhouses can keep warm to a certain extent, so two crops can be grown per year.

$$\begin{cases} s \le 1, if\, l \in \{\mathrm{A,B,C}\} \\ s \le 2, else \end{cases} \tag{4}$$

Irrigated land can be used to grow rice in one season or vegetable crops in two seasons each year.

$$\begin{cases} s \le 1, if\ a \in \{16\} \\ \quad s \le 2, else \end{cases}, l \in \{\mathrm{D}\} \leftrightarrow \tag{5}$$

There are certain restrictions on the types of crops that can be grown on flat dry land, terraced fields, and hillsides.

$$\begin{cases} x_{l,y,s,a} \ge 0, if\ l \in \{\mathrm{D,E}\}, \mathrm{S} = 1 \\ x_{l,y,s,a} \ge 0, if\ l \in \{\mathrm{F}\} \\ x_{l,y,s,a} = 0, else \end{cases} \tag{6}$$

Represents the restrictions on the types of crops planted in the first season of irrigated land, the first season of ordinary greenhouses, and the first and second seasons of smart greenhouses.

$$\begin{cases} x_{l,y,s,a} \ge 0, if\, l \in \{\mathrm{A,B,C}\} \\ a \in \{1, 2, ..., 15\} \\ x_{l,y,s,a} = 0, else \end{cases} \tag{7}$$

The types of crops that can be planted in the second season of irrigated land are limited to vegetables.

$$\begin{cases} x_{l,y,s,a} \ge 0, if\ l \in \{\mathrm{D}\}, \mathrm{s} = 2 \\ x_{l,y,s,a} = 0, else \end{cases} \tag{8}$$

Ordinary greenhouses are suitable for growing one season of vegetables and one season of edible fungi each year, while smart greenhouses are suitable for growing two seasons of vegetables each year.

$$\begin{cases} x_{l,y,z,a} \ge 0, if\ l \in \{\mathrm{E}\}, s = 2 \\ x_{l,y,z,a} = 0, else \end{cases} \tag{9}$$

Each crop cannot be planted continuously on the same plot of land (including the greenhouse).

$$\begin{cases} x_{l,y,s,a} + x_{l,y+1,s,a} \le 1, ifs \le 1 \\ x_{l,y,1,a} + x_{l,y,2,a}\ or\ x_{l,y,2,a} + x_{l,y+1,1,a}, else \leftrightarrow \end{cases} \tag{10}$$

All land in each plot (including greenhouses) is planted with legumes at least once in three years.

$$\sum_{i=0}^{2} x_{l,y+i,s,a} \geq 1, a \in \{1,2,3,4,5,17,18,19\} \leftrightarrow \tag{11}$$

The above formula means that the planting area of each crop in each season cannot be too scattered. In this paper, the constraint is that a crop can only be planted on 4 pieces of arable land at most during the same period. Then there is:

$$\sum_{l} x_{l,y,s,a} \leqslant 4 \tag{12}$$

The area of each crop planted in a single plot (including greenhouses) should not be too small, and the area planted in a single plot (including greenhouses) should not be less than 20% of the land.

$$x_{l,y,s,a} > 0.2 * A_l \tag{13}$$

The total area under crop cultivation cannot exceed the area of arable land.

$$\sum_{p} x_{l,y,s,a} \leq A_l \tag{14}$$

Among them, A_l represents the total area of cultivated land l.

3 SOLVING CROP PLANTING PLANS UNDER TWO OPTIMIZATION SCENARIOS USING GENETIC ALGORITHMS

Genetic Algorithms (GA) are search algorithms that mimic natural selection and genetic principles. GA encodes potential solutions as "individuals" in a "population." The population evolves through selection, crossover (mating), and mutation operations to find optimal or satisfactory solutions. GA performs a global search, avoiding local optima. Its simple structure adapts to various problems by adjusting genetic operations and parameters[8].

The process initializes a random population, then iteratively evaluates fitness, selects high-fitness individuals, performs crossover and mutation to generate new individuals, and continues until termination criteria are met. This optimization process aims to find the best or satisfactory solutions[9].

In this study, the decision variables $x_{l,y,s,a}$ involved different plots, years, seasons, and crops. It is a multi-dimensional dynamic variable that is difficult to represent using traditional coding methods. Therefore, this study improved the GA algorithm and adopted multi-matrix chromosome coding, which can more naturally map complex decision variables involving multiple dimensions such as plots, years, seasons, and crops[10]. This coding method facilitates the genetic algorithm to perform operations such as crossover and mutation. By distributing the decision variables in different matrices, not only decomposes the original huge search space into multiple small spaces but also improves the pertinence and efficiency of the search. Each matrix focuses on processing one type of decision variable, making the genetic algorithm more flexible and helping to find the optimal solution more effectively.

In summary, the model parameter settings of this paper are shown in Table 1:

Table 1 The Recommended Fonts

Parameter	Value
Iterations	1000
Crossover rate	0.8
Mutation rate	0.2
Population size	30

The convergence curve of the model in situation 1 is shown in Figure 1, and the convergence curve of the model in situation 2 is shown in Figure 2.

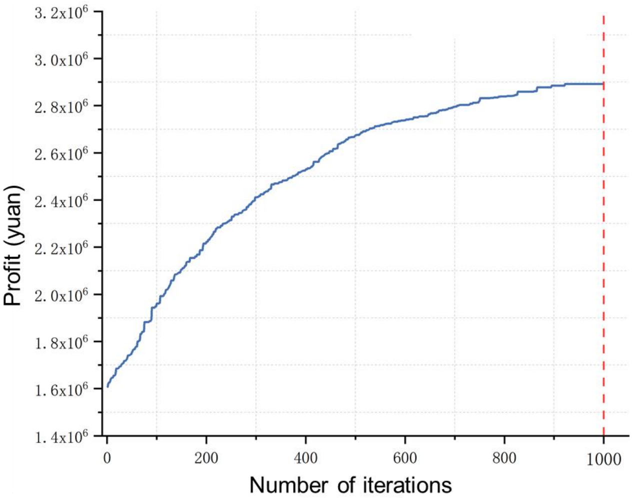


Figure 1 Situation 1 Convergence Curve

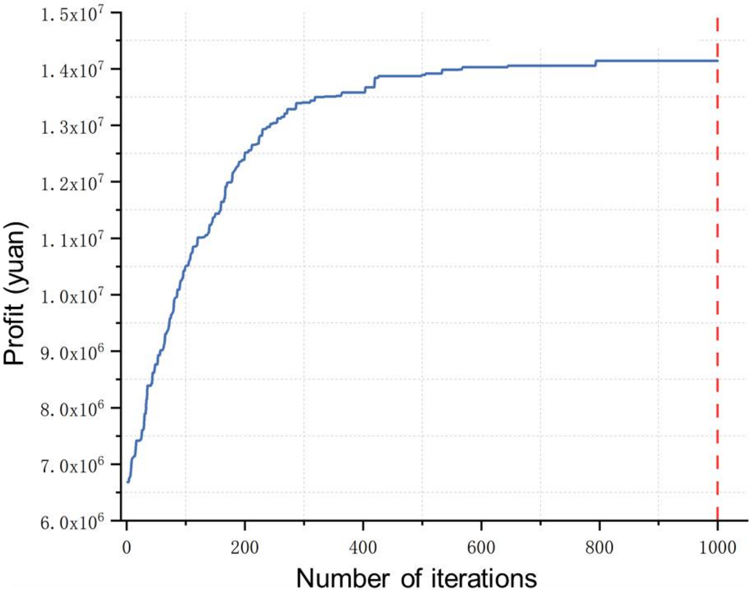


Figure 2 Situation 2 Convergence Curve

In scenario 1, after around 900 iterations, the convergence curve stabilizes at approximately 2.9106, indicating a profit of 2.9106 yuan. In scenario 2, the curve stabilizes after about 600 iterations at around 1.4107, suggesting a profit of 1.4107 yuan. Both scenarios converge quickly to optimal solutions, but scenario 2 achieves significantly higher profits (1.4107 vs 2.9106 yuan), demonstrating its superior sales strategy and revenue potential.

Due to space limitations, only the crop yield distribution and cultivated land area distribution of Case 2 are shown here, as shown in Figures 3 and 4.

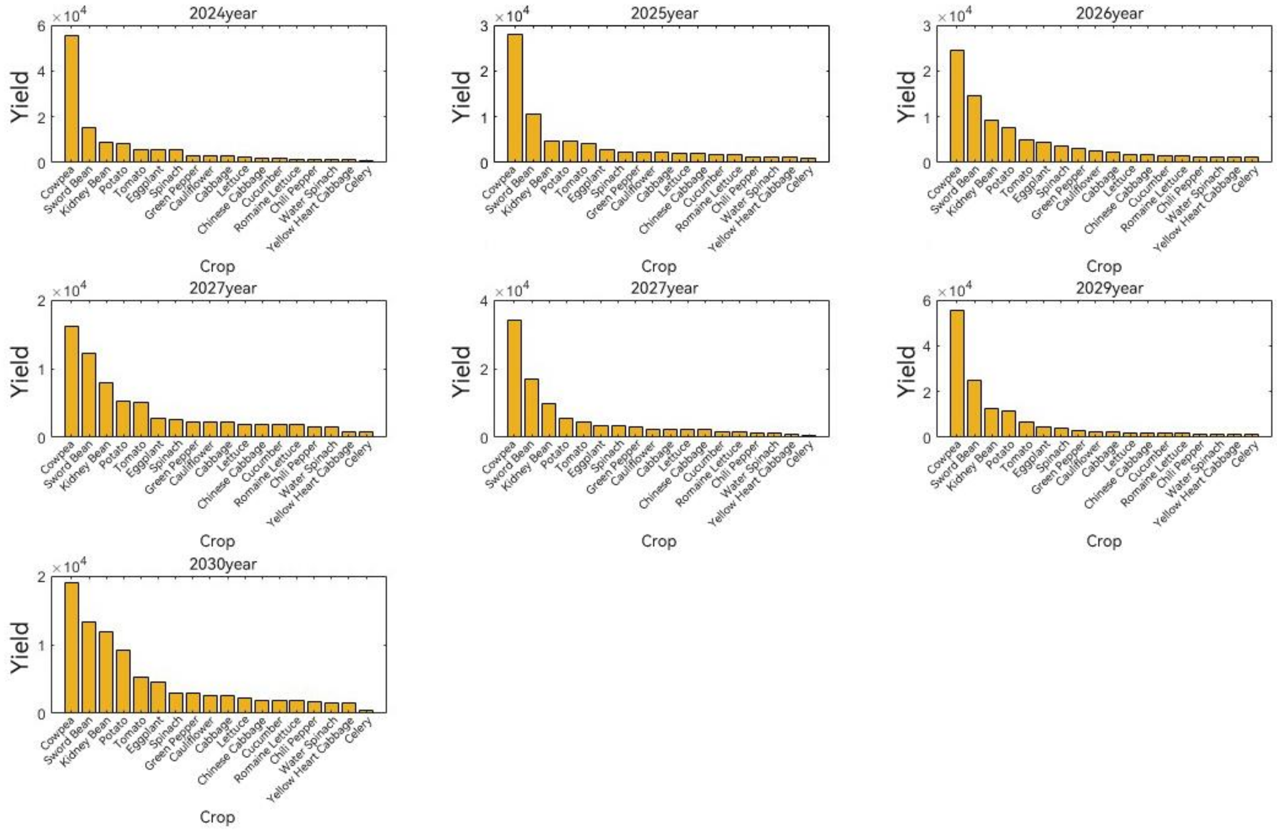


Figure 3 Scenario 2 Distribution of Crop Yields

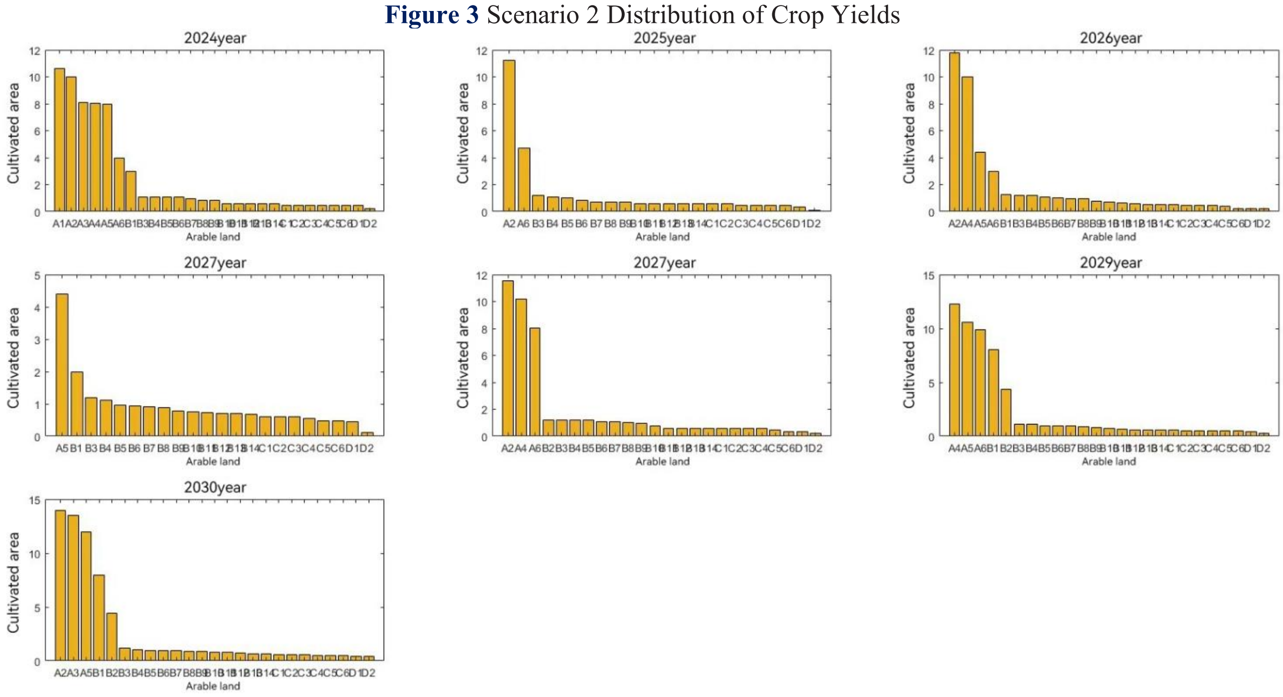


Figure 4 Scenario 2 Distribution of Cultivated Land Area

As can be seen from the above figure, in both case (1) and case (2), the top five crops in terms of crop yield are cowpea, sword bean, kidney bean, potato, and tomato, indicating that the economic benefits of these five crops are often good, so the yield is relatively large. In case (1), the top five cultivated land areas are all Class A land, that is, arid land; in case (2), the top five cultivated land areas include both Class A land and Class B land (arid land and terraced fields), indicating that overall, regardless of case (1) or case (2), the utilization rate of arid land in the planting plan is high, which indirectly reflects that the economic benefits brought by arid land are high.

4 CONCLUSIONS

This study developed an optimized crop planting strategy model using genetic algorithms, comparing scenarios where excess produce is either unsalable or sold at half price. The improved genetic algorithm with multi-matrix chromosome coding demonstrated effective convergence, with the half-price scenario achieving significantly higher profits. The analysis revealed that certain crops, notably cowpea, sword bean, kidney bean, potato, and tomato, consistently showed superior economic performance, while arid land emerged as the most economically advantageous cultivation area. These findings provide valuable insights for agricultural planning, suggesting that flexible pricing strategies combined with optimal crop selection can substantially enhance agricultural profitability. However, the current model has limitations, including its reliance on historical data without considering potential climate change impacts and market volatility. Future research could enhance this work by incorporating weather prediction models, market trend analysis, and the impact of emerging agricultural technologies. Additionally, the model could be expanded to consider environmental sustainability metrics and social factors affecting agricultural communities.

CONFLICT OF INTEREST

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] MAJA M M, AYANO S F. The impact of population growth on natural resources and farmers' capacity to adapt to climate change in low-income countries. Earth Systems and Environment, 2021, 5(2): 271-283.

[2] WANG X. Managing land carrying capacity: Key to achieving sustainable production systems for food security. Land, 2022, 11(4): 484.

[3] ZHOU Y, ZHANG E, LIN A. Evaluation of comprehensive use benefits and analysis of influencing factors of China's natural resources: based on a new spatial model. Environment, Development and Sustainability, 2024, 1-39.

[4] JI W, LIU D, MENG Y, et al. A review of genetic-based evolutionary algorithms in SVM parameters optimization. Evolutionary Intelligence, 2021, 14: 1389-1414.

[5] VIE A, KLEINNIJENHUIS A M, FARMER D J. Qualities, challenges and future of genetic algorithms: a literature review. arXiv preprint arXiv:2011.05277, 2020.

[6] RAJWAR K, DEEP K, DAS S. An exhaustive review of the metaheuristic algorithms for search and optimization: taxonomy, applications, and open challenges. Artificial Intelligence Review, 2023, 56(11): 13187-13257.

[7] KARAMIAN F, MIRAKZADEH A A, AZARI A. Application of multi-objective genetic algorithm for optimal combination of resources to achieve sustainable agriculture based on the water-energy-food nexus framework. Science of The Total Environment, 2023, 860: 160419.

[8] ALAM T, QAMAR S, DIXIT A, et al. Genetic algorithm: Reviews, implementations, and applications. arXiv preprint arXiv:2007.12673, 2020.

[9] TAHA Z Y, ABDULLAH A A, RASHID T A. Optimizing Feature Selection with Genetic Algorithms: A Review of Methods and Applications. arXiv preprint arXiv:2409.14563, 2024.

[10] ZHANG D, YANG S, LI S, et al. Integrated Optimization of the Location−Inventory Problem of Maintenance Component Distribution for High-Speed Railway Operations. Sustainability, 2020, 12(13): 5447.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3035

DEEPSEEK LARGE - SCALE MODEL: TECHNICAL ANALYSIS AND DEVELOPMENT PROSPECT

HaiLong Liao
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.
Corresponding Email: Jnhailong@126.com

Abstract: This paper deeply analyzes the DeepSeek large - scale model, comprehensively elaborating on its technical architecture, training mechanism, performance, application fields, as well as the challenges it faces and future development directions. Through the research on the DeepSeek series of models, it reveals their innovations and important values in the field of artificial intelligence. The research shows that the DeepSeek large - scale model, with its unique technical advantages, demonstrates excellent performance in tasks such as natural language processing, code generation, and multimodal understanding. It provides new ideas and methods for promoting the development and application of artificial intelligence technology.

Keywords: DeepSeek; Large language model; Artificial intelligence; Multimodality

1 INTRODUCTION

In the field of artificial intelligence, the development of large - scale models has profoundly changed many research directions and application scenarios in natural language processing, computer vision, and other areas. OpenAI's ChatGPT, with its powerful natural language interaction capabilities, has set off a global upsurge in AI applications since its launch. It has become a capable assistant for people to obtain information and complete tasks in daily life and work. In the highly competitive track of large - scale models, DeepSeek has emerged as a new force, attracting global attention.

DeepSeek was founded by Liang Wenfeng, an entrepreneur born in the 1980s from Guangdong Province. He has a unique growth trajectory and innovative ideas. In 2015, Liang Wenfeng and two classmates from Zhejiang University jointly founded the quantitative hedge fund High - Flyer. By using mathematical and artificial intelligence strategies for investment, by 2019, the fund's managed assets exceeded 10 billion yuan. During the process of managing the fund, Liang Wenfeng had a deeper understanding of the potential of artificial intelligence and gradually shifted his focus to the AI research and development field.

In 2023, Liang Wenfeng founded DeepSeek with the aim of developing general artificial intelligence (AGI) comparable to human intelligence. Different from traditional technology entrepreneurs, when forming the team, Liang Wenfeng boldly recruited Ph.D. students from top domestic universities. Although these young talents lacked industry experience, they had achieved many academic research results. Under Liang Wenfeng's unique bottom - up management model, the team's creativity was fully unleashed, laying a talent foundation for DeepSeek's technological innovation.

Since its establishment, DeepSeek has developed rapidly. The DeepSeek - V3 model, released at the end of 2024, shocked the industry. It only used 2,048 NVIDIA H800 chips and had a training cost of less than 6 million US dollars, but could achieve performance comparable to that of models developed by international big companies. The DeepSeek - R1 inference model, launched in January 2025, attracted widespread attention globally. Its application quickly climbed the download rankings in the Apple App Store, once surpassing well - known applications such as ChatGPT, demonstrating strong market competitiveness.

Compared with ChatGPT, DeepSeek shows unique advantages in many aspects. In terms of knowledge timeliness, DeepSeek's training data is updated to the fourth quarter of 2023, enabling it to better capture emerging technology trends. In terms of professional field depth, DeepSeek has constructed special knowledge graphs in vertical fields such as quantitative finance, semiconductor industry chain analysis, and cutting - edge biomedicine, providing more accurate services for professionals. In complex reasoning tasks, DeepSeek's logical reasoning and mathematical proof capabilities are more outstanding. In Chinese language processing, whether it is classical Chinese translation or industry - term understanding, DeepSeek performs more proficiently. Of course, DeepSeek also has some shortcomings. Currently, its multimodal capabilities are still under development, while ChatGPT has integrated image generation and voice interaction modules and performs more evenly in general scenarios.

With the rise of DeepSeek, its influence in the field of artificial intelligence is increasing day by day. In - depth research on the technical principles, training mechanisms, and application scenarios of the DeepSeek large - scale model not only helps us understand the key factors behind its success but also provides valuable reference for the further development of artificial intelligence technology, promoting the continuous progress of this field.

2 TECHNICAL ARCHITECTURE

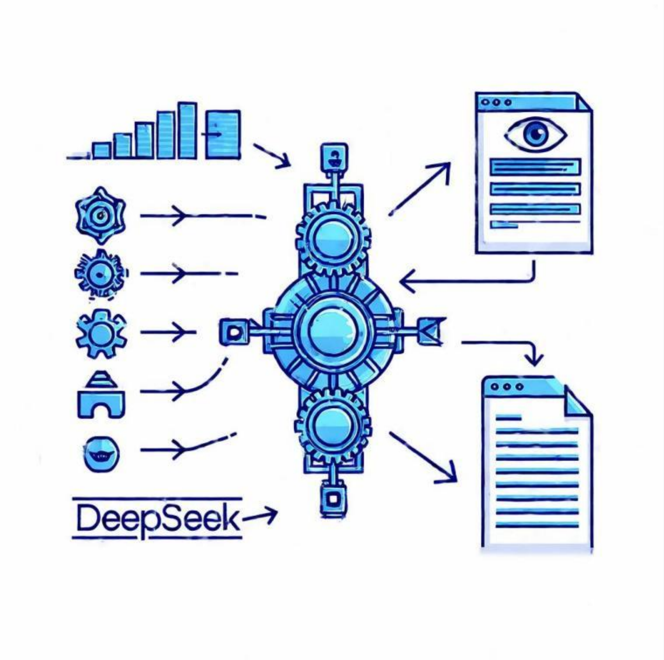


Figure 1 DeepSeek Technical Architecture Model

In the center of the figure is a large Transformer structure icon, representing the core architecture of DeepSeek (Figure 1). On the left side of the Transformer structure icon, arrows point to a series of optimization symbols (such as gears or adjustment knobs), indicating improvements to the attention mechanism to enhance the ability to process long - sequence data. In the upper - left corner of the Transformer structure icon, there is a sequence of gradually increasing number icons, symbolizing the parameter scale from small to extremely large (for example, from less than 1 billion to over 670 billion), indicating that the DeepSeek series of models have a huge number of parameters. On the right side of the Transformer structure icon, there are two interconnected small icons: one is a combination of an eye and a text box, representing visual information; the other is a simple text page icon, representing language information. These two icons are connected by a two - way arrow, indicating DeepSeek's achievements in multimodal fusion technology, especially how to effectively combine visual and language information. [1]

2.1 Innovation Based on the Transformer Architecture

The DeepSeek large - scale model is based on the Transformer architecture. The Transformer architecture, based on the attention mechanism, can effectively process sequence data and has achieved great success in natural language processing and other fields. [2]

DeepSeek has made innovative improvements to the Transformer architecture to enhance the model's ability to process long - sequence data.[7] By optimizing the attention mechanism, the model can more accurately capture the dependency relationships between various parts of the text. Thus, when processing long texts, it can have a deeper understanding of semantics and improve the accuracy and logic of the generated text.

2.2 Huge Model Parameters and Scale

The DeepSeek series of models have a huge parameter scale. For example, the DeepSeek - V3 model has as many as 671 billion parameters [3]. A large - scale parameter configuration endows the model with a stronger representation ability, enabling it to learn more abundant knowledge and language patterns. In natural language generation tasks, the model can generate more natural, fluent, and logical text. In code generation tasks, it can generate more accurate and efficient code. There is a positive correlation between the model parameter scale and performance. As the parameter scale increases, the model's performance in various tasks is also significantly improved.

2.3 Multimodal Fusion Technology

DeepSeek has made remarkable progress in multimodal fusion and has developed models such as the DeepSeek - VL2 vision - language model and the Janus - Pro multimodal model [4]. These models achieve the effective fusion of visual and language information through ingenious designs. For example, the DeepSeek - VL2 model uses a hybrid visual encoder, which can efficiently process high - resolution images (1024x1024) within a fixed token budget while maintaining relatively low computational costs. This enables the model to perform well in tasks such as visual question - answering and image description. The Janus - Pro model further enhances multimodal understanding and visual generation capabilities by decoupling visual encoding for multimodal understanding and visual generation, optimizing the training strategy, expanding the data, and increasing the model scale[5]. In text - to - image generation tasks, it can generate high - quality images that conform to semantics according to text instructions.

3 TRAINING MECHANISM

3.1 Pretraining: Preliminary Knowledge Accumulation

In the pretraining stage, the DeepSeek model is trained using massive corpus data. These data come from a wide range of sources, including Internet texts, academic literature, code libraries, etc. Through learning on large - scale data, the model can master rich language knowledge, semantic information, and general knowledge. For example, in the training of code generation models, a dataset containing a large amount of code is used, enabling the model to learn the syntax and programming patterns of multiple programming languages. Pretraining enables the model to have basic language understanding and generation capabilities, laying a good foundation for subsequent fine - tuning.

3.2 Supervised Fine - Tuning and Reinforcement Learning: Optimizing Model Behavior

In the supervised fine - tuning stage, the model is fine - tuned on the instruction dataset. Each sample in the dataset consists of an "Instruction Q - Response A" pair. In this way, the model can better follow human instructions. In the reinforcement learning stage, DeepSeek has adopted a variety of innovative methods. Taking DeepSeek - R1 as an example, it uses a rule - based reinforcement learning method (Group Relative Policy Optimization, GRPO). In mathematical problems, a reward score is calculated based on the accuracy of the answer. In code - related problems, the compiler is used to generate feedback based on predefined test cases. At the same time, a format reward is used to ensure that the model outputs in a specific format. This method simplifies the training process, reduces costs, and enables the model to perform well in inference tasks.

3.3 Training Optimization Strategies: Improving Efficiency and Performance

DeepSeek has adopted a variety of optimization strategies during the training process to improve training efficiency and model performance. In terms of hardware resource utilization, the DeepSeek - V3 model only uses 2,048 GPUs for 2 months of pretraining, greatly reducing the training cost [3]. In terms of algorithm optimization, methods such as adjusting the learning rate and optimizing gradient calculation are used to make the model training more stable and efficient. For example, a dynamic learning rate adjustment strategy is adopted. In the initial stage of training, the model can converge quickly, and in the later stage, fine - tuning is carried out to improve the model's accuracy. [3]

4 PERFORMANCE

4.1 Natural Language Processing Tasks

The DeepSeek model performs outstandingly in natural language processing tasks. In language generation tasks such as text continuation and story creation, the generated text is coherent, logical, and has few grammatical errors. In machine translation tasks, the translation accuracy and fluency reach a high level. In the GLUE (General Language Understanding Evaluation) benchmark test, the DeepSeek model achieved excellent results, demonstrating its strong language understanding ability. This benchmark test includes a variety of natural language understanding tasks such as text entailment and sentiment analysis. The model's comprehensive performance in these tasks reflects its in - depth understanding and processing ability of language.

4.2 Code Generation and Programming

DeepSeek's code generation model has achieved advanced performance among open - source code models in multiple programming languages and various benchmark tests. It can generate high - quality code based on natural language descriptions and performs well in tasks such as code completion and code error correction. In the CodeXGLUE code generation benchmark test, the DeepSeek - Coder model is superior to many similar models in terms of the accuracy and functionality of the generated code. For a given programming problem description, the model can quickly generate correct and runnable code, effectively improving software development efficiency.

4.3 Multimodal Tasks

DeepSeek's multimodal models demonstrate excellent capabilities in multimodal tasks. In visual question - answering tasks, the DeepSeek - VL2 model can accurately understand the image content and answer related questions. [4] For an image containing multiple objects, when asked "What is the red object in the image?", the model can accurately identify and answer. In text - to - image generation tasks, the Janus - Pro model can generate high - quality images that conform to semantics according to text instructions. When inputting "Generate an image of several seagulls flying on a sunny beach", the generated image can well reflect the scene described in the text, with rich image details and coordinated colors.

4.4 Comparison with Other Models

Table 1 Comparison of DeepSeek and Other Well-known Models

Comparison Dimension	**DeepSeek - V3**	**Other Well - known Models (e.g., Claude - 3.5 - Sonnet - 1022)**
Performance	Performs close to the current excellent level in knowledge - based tasks (such as MMLU, MMLU - pro, GPQA, SimpleQA); surpasses other open - source and closed - source models in math competitions (such as AIME2024, CNMO2024).	Performs well, especially in knowledge - based tasks, being comparable to DeepSeek - V3, but may be inferior to DeepSeek - V3 in math competition tasks.
Cost - Efficiency	Low training cost, only using 2,048 GPUs for 2 months of training, costing about $5.576 million.	May require more computing resources and higher costs.
Generation Speed	The output speed has increased from 20 tokens per second to 60 tokens per second, providing a smoother user experience.	Not specifically mentioned, but it is implied that it may not be as smooth as DeepSeek - V3.

Compared with other well - known large - scale models, the DeepSeek model is competitive in terms of performance and cost - efficiency (Table 1).

5 APPLICATION FIELDS

5.1 Intelligent Customer Service and Dialogue Systems

In the field of intelligent customer service, the DeepSeek model has been widely applied. Many enterprises have integrated it into their customer service systems. The model can quickly and accurately understand user questions and provide corresponding answers. In e - commerce customer service, when users ask questions about product information, logistics status, etc., the model can respond quickly and provide accurate answers, improving customer service efficiency and user satisfaction. In dialogue systems, the model can conduct natural and smooth conversations, understand the context, and achieve multi - turn conversations, providing users with an intelligent interaction experience.

5.2 Code Development and Programming Assistance

In code development, DeepSeek's code generation model provides powerful programming assistance capabilities for programmers. It can generate code snippets based on natural language descriptions, helping programmers quickly implement functions. When developing a web application, a programmer can input "Generate the front - end code for a user login interface", and the model can generate the corresponding HTML, CSS, and JavaScript code, reducing development time and workload. The model can also perform code completion and code error correction, improving code quality and development efficiency.

5.3 Multimodal Content Creation

In the field of multimodal content creation, DeepSeek's multimodal models play an important role. In advertising design, designers can input text descriptions, such as "Design a poster to promote new energy vehicles, highlighting environmental protection and a sense of technology". The Janus - Pro model can generate corresponding images, providing creative inspiration and visual references for designers. In film and television production, the model can generate scene concept maps based on script descriptions, helping directors and art teams better conceive scene layouts. In the education field, based on image - based learning materials, when students ask questions about the image content, the model can understand the questions and combine image information to give accurate answers, assisting in teaching and students' independent learning.

6 CHALLENGES AND PROSPECTS

6.1 Challenges Faced

Although the DeepSeek large - scale model has achieved remarkable results, it still faces some challenges. In terms of model interpretability, due to the large number of model parameters and complex structure, it is difficult to understand the decision - making process of the model and the reasons for the output results.[6] When dealing with some sensitive information, how to ensure data security and privacy protection is also a problem that needs to be solved. The model's performance depends on large - scale data and powerful computing resources, and limitations in data quality and computing resources may affect the model's effectiveness. In addition, the generalization ability in different fields and tasks and how to better adapt to complex and changeable practical application scenarios are also directions that require further research.

6.2 Future Development Directions

In the future, DeepSeek is expected to make breakthroughs in model interpretability research, developing visualization tools or interpretive algorithms to help users understand the internal mechanisms of the model. In terms of data security and privacy protection, more advanced encryption technologies and privacy - protection algorithms will be explored to ensure the security of data during use. With the development of hardware technology, more efficient computing devices and distributed computing technology will be utilized to further improve the model's training efficiency and performance. In applications, the model will be more deeply integrated into various industries, and more customized solutions will be developed to meet the needs of different users. There may also be explorations in cross - modal fusion, knowledge graph fusion, etc., to enhance the model's comprehensive capabilities.

7 CONCLUSION

The DeepSeek large - scale model, with its innovative technical architecture, unique training mechanism, and excellent performance, demonstrates strong competitiveness and broad application prospects in the field of artificial intelligence. It has achieved excellent results in tasks such as natural language processing, code generation, and multimodal understanding, providing strong support for the development of various industries. Although it faces some challenges, with the continuous progress of technology and in - depth research, the DeepSeek large - scale model is expected to make greater breakthroughs in the future, promoting the development of artificial intelligence technology and bringing more innovative applications and value to human society.

CONFLICT OF INTEREST

The authors have no relevant financial or non-financial interests to disclose.

8 REFERENCES

[1] Rohan Paul. DeepSeek - V3's Architectural Revolution: Rewriting the Economics of Large Language Model Training. 2024. Retrieved from https://rohanpaul.substack.com/p/deepseek-v3-technical-report-they

[2] Vaswani, A, Shazeer, N, Parmar, N, et al. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017.

[3] DeepSeek-V3 Technical Report. It is authored by DeepSeek-AI, 2024. DOI: https://doi.org/10.48550/arXiv.2412.19437. Retrieved from https://arxiv.org/abs/2412.19437v1. Project homepage: https://github.com/deepseek-ai/DeepSeek-V3.

[4] Elmo. DeepSeek - VL: New Open Source Vision - Language Models! Medium (Medium Reviews). 2024. Retrieved from https://medium.com/@elmo92/deepseek-vl-new-open-source-vision-language-models-32bc77fa4647

[5] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, et al. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling, 2025. Retrieved from https://arxiv.org/abs/2501.17811v1

[6] DeepSeek-AI, Daya Guo, Dejian Yang, et al. DeepSeek - R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. 2025. DOI: https://doi.org/10.48550/arXiv.2501.12948. Retrieved from https://arxiv.org/pdf/2501.12948

[7] DeepSeek-AI, Xiao Bi, Deli Chen, et al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. 2024. Retrieved from https://arxiv.org/abs/2401.02954

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3036

DESIGN OF INTELLIGENT NURSING HOME MONITORING SYSTEM BASED ON WI-FI AND GSM COMMUNICATION TECHNOLOGY

HuiXian Chen, PengHui Liu, MengYang Zhang, Guan Wang*
Pingdingshan University, Pingdingshan 467000, Henan, China.
Corresponding Author: Guan Wang, Email: wangguan0123@163.com

Abstract: With the aging of the population, intelligent nursing homes have become the key to improving the quality of life of the elderly. In this paper, a smart nursing home monitoring system based on the STM32 microcontroller is proposed. The system carries out comprehensive data collection and monitoring through modules such as the DHT11 temperature and humidity sensor, sound sensor, infrared sensor, heart rate and blood oxygen sensor, and MPU6050 gyroscope. The system utilizes Wi-Fi wireless communication technology to upload the health monitoring data to the cloud, and at the same time combines with GSM communication technology to realize the SMS alarm and one-key dialing function of the GSM module. The system not only has efficient and real-time monitoring capability but also provides comprehensive health monitoring and security for smart nursing homes through OLED display, voice announcement, and buzzer alarm.

Keywords: Intelligent nursing home; Single-Chip microcomputer; Wireless communication; Health monitoring

1 INTRODUCTION

With the increasingly serious problem of global population aging, how to improve the quality of life and safety of the elderly has become the focus of social concern. Traditional nursing homes have problems such as rudimentary hardware and software facilities, unqualified supervision, and low service quality, so how to solve the existing issues to create a high-quality nursing home and effectively improve the quality of life of the elderly has become a social problem that needs to be solved in China [1]. Intelligent technology is actively introduced to improve the efficiency of nursing home services and the safety of the elderly. The smart nursing home monitoring system, as a solution integrating environment monitoring, physical signs detection, emergency response, and remote management, can collect real-time health data of the elderly and monitor environmental changes, to realize comprehensive management and rapid response to the health status of the elderly.

This paper proposes an intelligent nursing home monitoring system design based on Wi-Fi and GSM communication technology, aiming to realize efficient linkage and information transfer between system modules through wireless communication technology. The system focuses on monitoring the heart rate, blood oxygen, and body position of the elderly, keeping an eye on their physical health and alerting them to whether they fall. The system is supplemented by environmental detection and intelligent control, monitoring the air temperature and humidity, as well as the sound decibel, to provide a warm and comfortable environment for the elderly to live in, and the combination of the two builds an efficient and safe intelligent monitoring system. By exploring the design and implementation of the system in depth, this paper aims to provide a feasible technical solution for the construction of intelligent nursing homes, promote the thoughtful process of elderly services, and help the sustainable development of the aging society.

2 INTELLIGENT NURSING HOME MONITORING SYSTEM DESIGN

2.1 System Design Objectives

The requirement of this system design is mainly to create an intelligent nursing home with a comfortable environment, health, and safety through the combination of environmental detection and the detection of physical signs of the elderly. The functions to be realized include:

2.1.1 Real-time environment monitoring

The DHT11 temperature and humidity sensor module detects the environmental temperature and humidity, the sound sensor detects the sound decibels, and the infrared sensor monitors the entry and exit of personnel to ensure the comfort and safety of the nursing home environment.

2.1.2 Accurate monitoring of the signs of the elderly

Using heart rate and blood oxygen sensors to monitor the vital signs of the elderly in real-time, and the MPU6050 gyroscope to detect whether the elderly have fallen, to detect and deal with the abnormal problems of the elderly in time.

2.1.3 Abnormal Alarm and Broadcasting

When the old man's physical signs exceed the preset threshold, the voice broadcast module will broadcast the abnormal physical signs, and the buzzer alarm will remind the guardian to pay attention to the old man's condition in time.

2.1.4 Data display and uploading

The OLED display shows the monitoring data in real-time, which is convenient for on-site checking; meanwhile, the data can be uploaded to the cloud via ESP8266 to realize remote monitoring and data storage.

2.1.5 Emergency communication function

When the old man falls, the GSM communication module automatically sends a message to notify the relevant personnel; when the old man is uncomfortable, he can seek help by pressing the button to broadcast a call to ensure that timely help can be sought in case of emergency.

2.1.6 Stability and safety of the system

To protect the safety of the elderly, the system has a certain fault-tolerant design, such as automatically switching to the standby mode in the event of a communication failure, to ensure that the key functions are not affected. The system needs to ensure that the data uploaded to the cloud has sufficient security to prevent data leakage and protect the privacy of the elderly.

2.2 General Design Ideas

The system aims to realize the all-around monitoring of the intelligent nursing home by integrating multiple sensors and communication modules with the STM32 microcontroller as the core. Specifically, the system utilizes DHT11, sound sensors, and infrared sensors to monitor the environmental conditions; real-time monitoring of the elderly's physical signs and activity status through the heart rate and blood oxygen sensors and the MPU6050 gyroscope; the combination of voice announcements, buzzer alarms, OLED displays, and cloud uploads to ensure that the abnormal information is handled and recorded promptly; and is equipped with a GSM module to achieve rapid emergency It is also equipped with GSM module to realize fast communication response in emergency.

2.3 Overall System Framework

This system is an intelligent nursing home monitoring system based on an STM32 microcontroller, which mainly includes two modules: environmental detection and detection of elderly signs. It consists of an STM32F103C8T6 microcontroller core board, DHT11 temperature and humidity sensor, sound sensor, an infrared detection module, heart rate, and blood oxygen sensor, MPU6050 gyroscope, SU-03T voice announcing circuit, OLED display circuit, WIFI module, GSM communication module, key circuit and power supply.

The environmental detection module in this design integrates a DHT11 temperature and humidity sensor, a sound sensor, and an infrared sensor. The DHT11 sensor is used to monitor the temperature and humidity data in the nursing home in real-time, the sound sensor detects sound decibels in the environment to assess the noise level, and the infrared sensor monitors the entry and exit of the personnel and ensures the safety management of the nursing home.

The Elderly Signs Detection Module has a heart rate oximetry sensor and MPU6050 gyroscope. The heart rate oximetry sensor monitors the vital signs of the elderly in real-time, including heart rate and oxygen saturation, to assess the health of the elderly, while the MPU6050 gyroscope is used to detect changes in the posture of the elderly and determine whether a fall has occurred.

The system also contains a voice announcement module, a buzzer alarm device, an OLED display, an ESP8266 cloud upload module, and a GSM communication module. When the physical data of the elderly exceeds the preset threshold, the voice announcement module will announce the abnormal situation in time, and the buzzer will sound an alarm to alert the staff, the OLED display can show the monitoring data in real-time, which is convenient for on-site viewing, and the ESP8266 module is responsible for uploading the data to the cloud to realize remote monitoring and data analysis. If the old man falls, the GSM communication module will automatically send an emergency message to the preset contact person, and when the old man is not feeling well, he can trigger the broadcast call function by pressing the button to seek help. The general framework diagram of the design is shown in Figures 1 and 2.

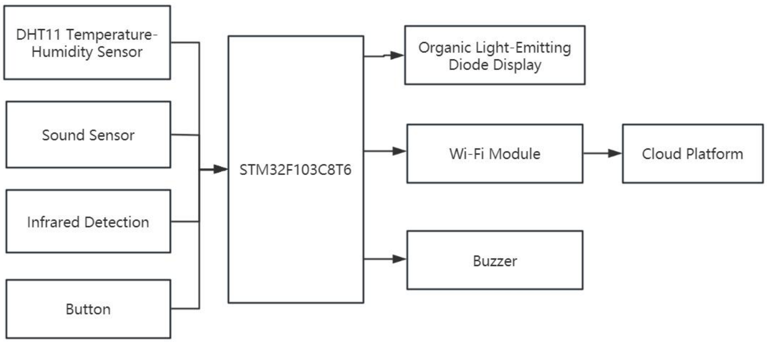


Figure 1 General Framework Diagram for the Design of the Environmental Module

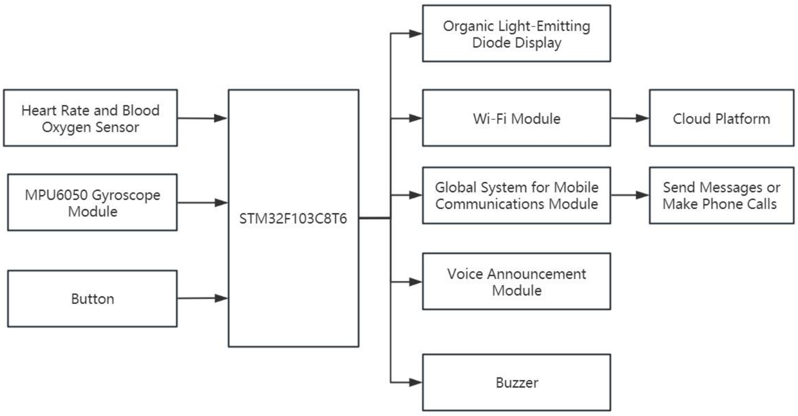

Organic Light-Emitting Diode Display
Heart Rate and Blood Oxygen Sensor
Wi-Fi Module
Cloud Platform
MPU6050 Gyroscope Module
STM32F103C8T6
Global System for Mobile Communications Module
Send Messages or Make Phone Calls
Voice Announcement Module
Button
Buzzer

Figure 2 General Frame of the Design of the Physical Signs Detection Module for the Elderly

3 SYSTEM HARDWARE DESIGN

3.1 Master Module

According to the function of the system, the STM32F103C8T6 minimum system board is selected for this design, which is based on the ARM Cortex-M3 core with high performance and low power consumption, and the main frequency of the 32-bit architecture can be up to 72MHz, with highly efficient operation, rich peripheral interfaces (such as multiple communication interfaces and timers, etc.) and reasonable memory configuration (64KB flash memory, 20KB RAM), and programming is more efficient through the SWD interface. Reasonable memory configuration (64KB Flash, 20KB RAM), programming is convenient with the help of official libraries, and debugging is carried out efficiently through the SWD interface; compared with the 51 series of microcontrollers, it has a faster computing speed, richer memory resources, stronger peripheral functions and high programming efficiency, which gives STM32F103C8T6 a significant advantage in complex applications.

3.2 DHT11 Temperature and Humidity Sensor Module

In this system design, the DHT11 sensor module is used to collect the temperature and humidity data in the air and detect the changes in temperature and humidity data in real-time. The DHT11 sensor module is a high-performance temperature and humidity composite sensor, which is built-in a resistive humidity-sensitive element and an NTC (Negative Temperature Coefficient thermistor) temperature measurement element to complete the acquisition and processing of temperature and humidity signals through the internal microcontroller. The DHT11 sensor module is a high-performance temperature and humidity composite sensor with a built-in resistive moisture-sensitive element and an NTC (Negative Temperature Coefficient thermistor) temperature-measuring element, and the temperature and humidity signals are collected and processed through the internal microcontroller [2]. The module can accurately measure the temperature and humidity of the environment in real-time through the dedicated digital module acquisition technology and output the results as calibrated digital signals. It has a wide measurement range, high accuracy, excellent long-term stability, and anti-interference capability. In addition, the DHT11 sensor module uses a single bus digital signal transmission protocol, simplifying the external circuit design and facilitating communication with a variety of microcontrollers, which is ideal for embedded system applications. The circuit diagram and physical diagram of the DHT11 module are shown in Figures 3 and 4 respectively.

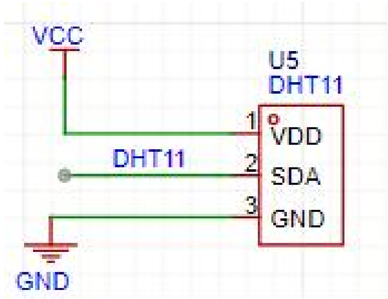


Figure 3 DHT11 Module Circuit Diagram

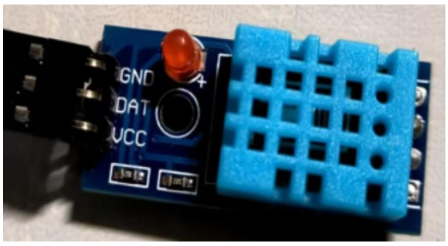


Figure 4 Physical Diagram of DHT11 Module

3.3 KY-037 Sound Sensor Module

In this system design, the KY-037 sound sensor module is used to detect the sound decibels in the environment, if the sound exceeds the threshold value to become noise, the management personnel can check the processing in time. The KY-037 sound sensor module is a high-sensitivity sound detection module with a built-in electret condenser microphone and amplifier circuit, which provides analog and digital dual outputs. The KY-037 module is a highly sensitive sound detection module, capable of capturing small sound changes, and also has the advantages of adjustable sensitivity, low power consumption, and ease of use. The KY-037 module has a high sensitivity and can capture small sound changes, and also has the advantages of adjustable sensitivity, low power consumption, and ease of use. It is very suitable for the environmental monitoring of the intelligent nursing home. The circuit diagram and physical drawing of the KY-037 sound sensor module are shown in Figures 2-3 and 2-4 respectively.

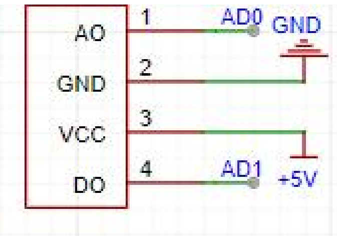


Figure 5 KY-037 Sound Sensor Module Circuit Diagram

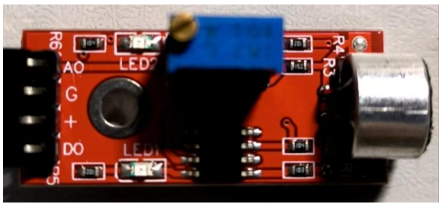


Figure 6 KY-037 Sound Sensor Module Physical Diagram

3.4 Infrared Detection Module

In this system design, an infrared detection module is used for in and out-person detection to prevent the elderly from accidentally getting lost when they go out alone. An infrared detection module is an electronic module that uses infrared light-sensing technology to detect changes in objects or the environment. The module consists of an infrared emitting diode (LED) and an infrared receiving diode (photodiode), which detects the presence, distance, or movement of an object by emitting infrared light of a certain wavelength and receiving the reflected infrared signal. The output of the module can be analog or digital signals, easy to interface with microcontrollers (such as microcontrollers), infrared detection module has non-contact detection, fast response speed, detection distance, high detection accuracy, and anti-interference capability. At the same time, the infrared detection module also has a small size, low power consumption, ease of integrate, and other advantages. Infrared detection module circuit diagram, the physical figure shown in Figure 7, 8, respectively.

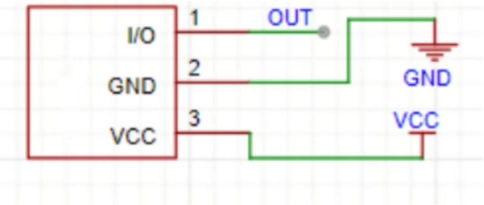


Figure 7 Infrared Detection Module Circuit Diagram

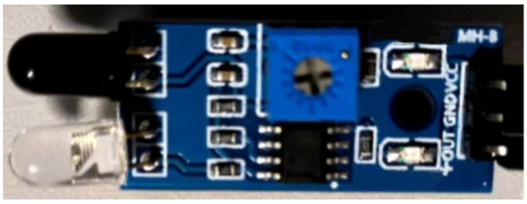


Figure 8 Physical Diagram of Infrared Detection Module

3.5 MAX30102 Heart Rate Oximeter Sensor Module

In this system design, the MAX30102 heart rate and blood oxygen sensor module are used to detect the basic body signs of the elderly. The MAX30102 sensor is a high-performance biosensor module. The MAX30102 is an optically integrated chip that carries red LEDs and infrared LEDs, in addition to being equipped with an optoelectronic detection device and an optimized low-noise AFE. The MAX30102 can regulate the power supply for the built-in LEDs via a 5.0V supply voltage and a 1.8V supply voltage and is mainly used in portable wearable applications, where it can detect the skin on multiple parts of the body, such as the fingertips, wrist joints, and earlobes, through the built-in included blood oxygen and heart rate monitoring solutions. The built-in I2C communication port of the sensor transmits the data received from the sensor to the master MCU, which performs a calculation function to figure out the actual heart rate and blood oxygen values [3]. This optical sensing technology can accurately and non-invasively monitor the physiological parameters of the human body. The circuit diagram and physical diagram of the module are shown in Figures 2-7 and 2-8, respectively.

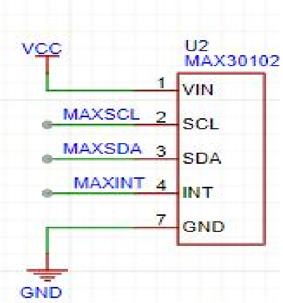


Figure 9 Circuit Diagram of the MAX30102 Heart Rate Oximetry Sensor Module

Figure 10 Physical Diagram of MAX30102 Heart Rate Oximetry Sensor Module

3.6 MPU6050 Gyro Module

In this system design, to monitor whether an elderly person falls or not, the MPU6050 gyroscope module is used. MPU6050 is a three-axis accelerometer sensor that senses the tilt condition of the user's body and calculates the tilt angle of the user's module, which in turn detects whether the user has fallen or not by comparing it to a set threshold [4]. The MPU6050 module is a three-axis gyroscope and a three-axis accelerometer integrated with an Inertial Measurement Unit (IMU) that provides motion and attitude measurements by sensing the angular velocity and acceleration of an object. The gyroscope measures the rotational angular velocity of the object, while the accelerometer measures the acceleration of the object along the three axes (X, Y, and Z). The MPU6050 communicates with an external microcontroller via an I2C interface to convert the acquired sensing data into digital signal output. Its working principle is based on vibration sensing technology and angular velocity detection, which provides accurate attitude estimation and dynamic motion detection by monitoring and calculating the object's motion status in real-time. It can detect the body posture of the elderly in real-time and determine whether the elderly have fallen. The circuit diagram and physical drawing of the MPU6050 gyroscope module are shown in Figures 11 and 12, respectively.

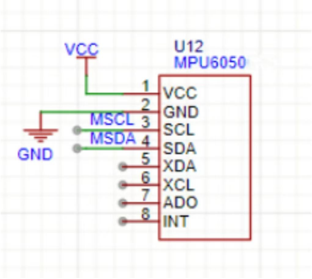


Figure 11 MPU6050 Gyroscope Module Circuit Diagram

Figure 12 Physical Diagram of MPU6050 Gyroscope Module

3.7 OLED Display Module

In this system design, an OLED display module is used to display the data transmitted by each sensor in real-time on the OLED display. The OLED display module is a display module based on Organic Light Emitting Diode (OLED) technology, which has the advantages of high contrast ratio, wide viewing angle, and low power consumption. Unlike traditional liquid crystal displays (LCDs), OLEDs do not rely on a backlight source but are self-illuminated by organic materials excited by an electric current, so each pixel point can be controlled independently, providing sharper colors and deeper blacks. It also excels in terms of response speed, with extremely short response times for pixels, allowing for fast screen switching and effective reduction of ghosting. The module is usually equipped with a controller chip that supports communication with a microcontroller through an I2C or SPI interface. The circuit diagram and physical drawing of the OLED display module are shown in Figures 13 and 14 respectively.

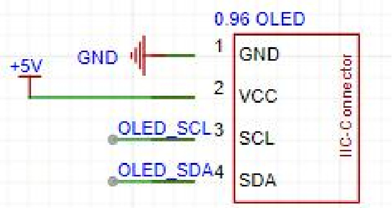


Figure 13 OLED Display Module Circuit Diagram

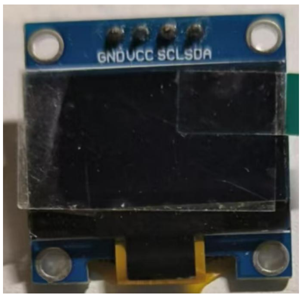


Figure 14 Physical Diagram of OLED Display Module

3.8 SU-03T Voice Announcement Module

In this system design, the SU-03T voice module is used to achieve the effect of voice announcement. When the old man's physical signs are abnormal, the voice module immediately sends out a voice announcement to remind the staff to check the old man's physical condition in time.SU-03T voice module is a kind of integrated voice processing equipment. Its principle is to receive external voice signals through the built-in voice recognition chip, convert them into digital signals, and then use internal algorithms to identify and process the voice content; at the same time, it can also convert pre-stored digital voice information into analog voice signals for playback. It has the characteristics of fast response speed, high stability, strong security, etc. The circuit diagram and physical diagram of the SU-03T voice module are shown in Figures 15 and 16 respectively.

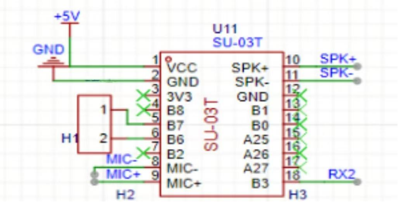


Figure 15 SU-03T Voice Module Circuit Diagram

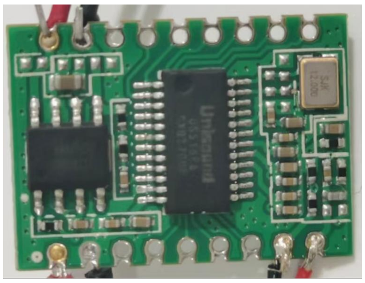


Figure 16 SU-03T Voice Module Physical Drawing

3.9 Buzzer Module

In this system design, a buzzer module is used to play an alarm role. When the environmental data or the signs of the elderly have abnormalities buzzer will sound an alarm to remind the staff to check and solve the problem promptly. This design module uses an active buzzer, The active buzzer module is commonly used in electronic devices in the sound output components, mainly through the built-in oscillator circuit to generate a fixed frequency sound signal. Its working principle is based on the voltage-excited oscillator principle, when power is added to the electrodes of the module, the oscillator inside the module starts to work, thus driving the buzzer to emit sound. Specifically, the module contains an electronic oscillator circuit, a speaker, and a driver circuit, which causes the buzzer to vibrate at a particular frequency through a change in voltage to produce sound. It is characterized by small size, low power consumption, easy integration, and control. The circuit diagram and physical diagram of the buzzer module are shown in Figures 17 and 18, respectively.

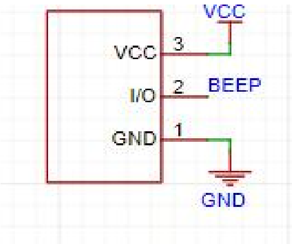


Figure 17 Buzzer Module Circuit Diagram

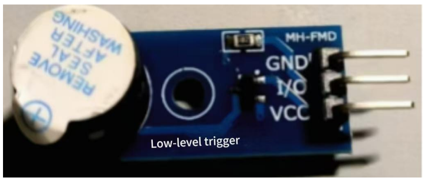


Figure 18 Physical Diagram of the Buzzer Module

3.10 ESP8266 WiFi Module

In this system design, the wireless communication uses the ESP8266 WiFi module, whose function is to connect to the cloud platform to realize the wireless transmission of information data and real-time monitoring. The ESP8266 is a low-cost, high-performance Wi-Fi microcontroller module that is widely used in the wireless communication of IoT devices. It integrates Wi-Fi function and microprocessor, supports multiple communication protocols (e.g., TCP/IP, UDP, etc.), and has strong processing capability to connect directly to the Internet for data transmission.ESP8266 has the advantages of low power consumption, easy to be embedded, easy to be developed, etc. The circuit diagram and physical diagram of the ESP8266 WiFi module are shown in Figures 19 and 20, respectively.

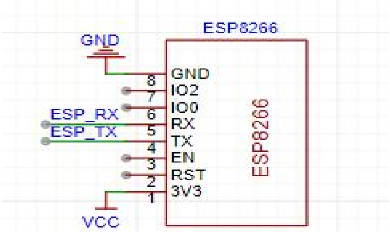


Figure 19 ESP8266 WiFi Module Circuit Diagram

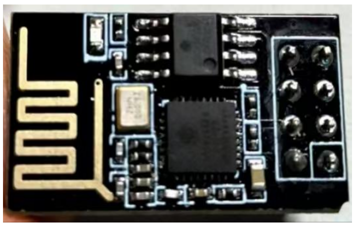


Figure 20 Physical Diagram of the ESP8266 WiFi Module

3.11 GSM Communication Module

In this system design, the wireless communication also uses a GSM communication module. Its function is to send SMS and make phone calls to the staff in time when the elderly fall condition occurs. The communication module is a wireless communication module based on the Global System for Mobile Communications (GSM) standard, which is mainly used to realize data transmission, voice communication, and SMS functions. It realizes the connection with the mobile network by putting the mobile SIM card into the SIM card slot and connecting to the GSM antenna, which can be remotely controlled and monitored, and supports sending and receiving SMS, making voice calls, and serial communication with other devices. The module has an internal integrated GSM modem, RF unit, and protocol stack, and supports serial ports (e.g. UART) for data exchange with external microcontrollers. When the system receives an alarm signal, the wireless communication module is controlled by sending AT commands through the serial port, and the module sends the alarm information to the staff's cell phone after receiving the commands [5]. The circuit diagram and physical diagram of the GSM communication module are shown in Figures 21 and 22, respectively.

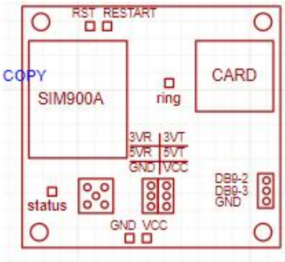


Figure 21 Circuit Diagram of GSM Communication Module

Figure 22 Physical Diagram of GSM Communication Module

4 SYSTEM SOFTWARE DESIGN

According to the system function, the system can be divided into two sub-systems, the main program of sub-system 1 contains two major sub-programs, i.e.: infrared in/out detection sub-system and environment detection sub-system; and the main program of sub-system 2 contains three major sub-programs, i.e.: GSM communication sub-system, elderly sign detection sub-system, and MPU6050 fall detection sub-system.

First of all, in subsystem 1 after connecting the power supply, the whole system is initialized, by the instructions to enter the corresponding sub-system program system overall flow chart, through the infrared sensor to detect whether there is the entry and exit of the person, to remind the guardian to check whether there is an old man out, to ensure the safety of the elderly, at the same time, temperature and humidity sensors, noise sensors to detect the environmental status of the old people living in the environment, will be collected to display the data on the OLED screen, and through the WiFi module to upload the collected data to the AliCloud platform, real-time view data. The main flow chart is shown in Figure 23.

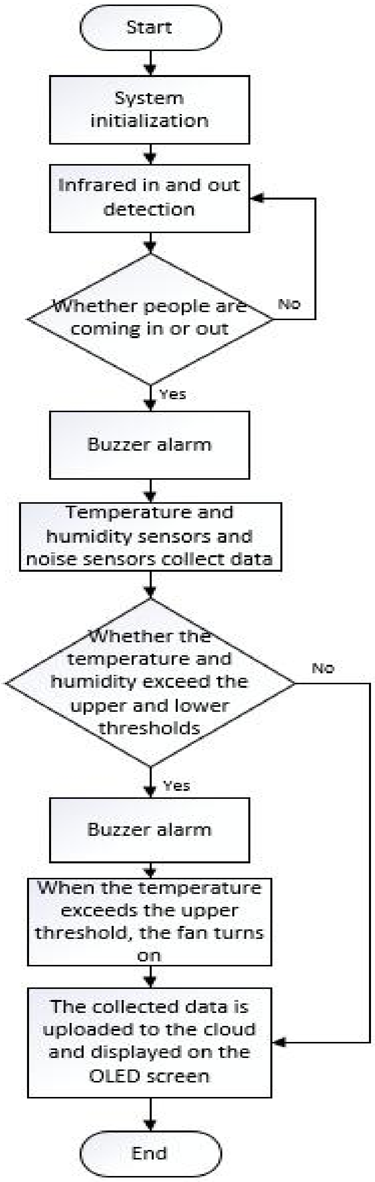

Start
System initialization
Infrared in and out detection
Whether people are coming in or out
No
Yes
Buzzer alarm
Temperature and humidity sensors and noise sensors collect data
Whether the temperature and humidity exceed the upper and lower thresholds
No
Yes
Buzzer alarm
When the temperature exceeds the upper threshold, the fan turns on
The collected data is uploaded to the cloud and displayed on the OLED screen
End

Figure 23 STM32 Main Program 1 Flowchart

4.1 Environmental Detection Subsystem

According to the system function, first, initialize the corresponding functional modules and their pins, collect the temperature and humidity information through the temperature and humidity sensor, and at the same time display the temperature and humidity data on the OLED screen, and determine whether the temperature and humidity value is within the normal range. If the temperature and humidity value exceeds the set upper and lower thresholds, the buzzer alarms, and when the temperature value exceeds the set upper-temperature threshold, the drive fan is turned on to achieve the role of temperature regulation. The flowchart of the environment detection subsystem is shown in Figure 24.

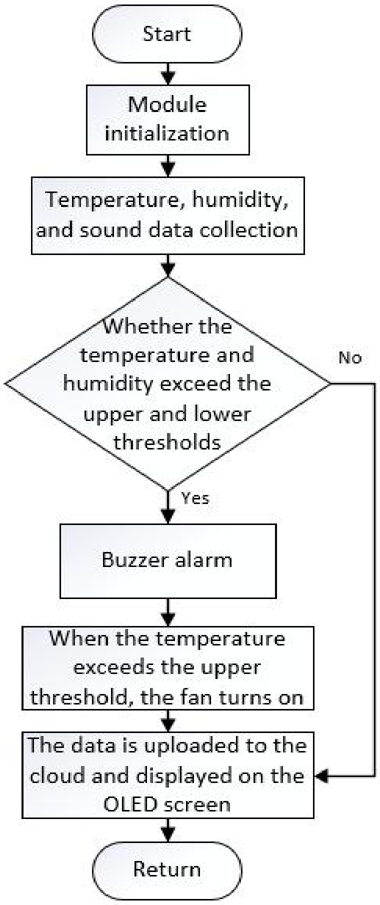

Start
Module initialization
Temperature, humidity, and sound data collection
Whether the temperature and humidity exceed the upper and lower thresholds
No
Yes
Buzzer alarm
When the temperature exceeds the upper threshold, the fan turns on
The data is uploaded to the cloud and displayed on the OLED screen
Return

Figure 24 Environmental Detection Subsystem Program Flowchart

4.2 Infrared Access Detection Subsystem

According to the system function, after the system is powered on, the corresponding functional modules and their pins are initialized, and then infrared transmission and reception are to detect whether someone passes through, if someone passes through, the buzzer alarm, the state of in and out on the cloud platform is reversed, and if no one passes through then return to the detection of infrared in and out of the detection sub-system flowchart, as shown in Figure 25.

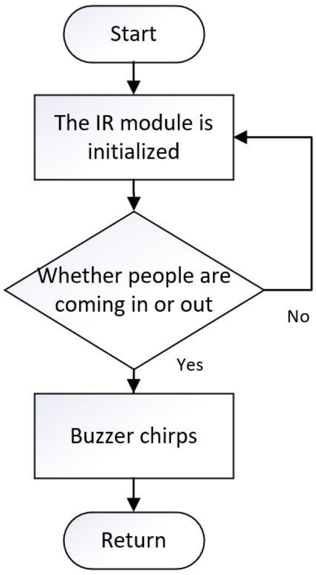

Start
The IR module is initialized
Whether people are coming in or out
No
Yes
Buzzer chirps
Return

Figure 25 Infrared in/out Detection Subsystem Program Flow Chart

Secondly, after connecting the power supply to subsystem 2, the whole system is initialized and enters the corresponding subsystem program in order according to the instructions. The overall flowchart of the system is to measure the signs of the elderly through the MAX30102 heart rate oximetry sensor, and if the signs of the elderly are not within the normal range, then drive the voice module to broadcast the message of “Abnormal Signs of the Elderly” and the collected signs will be uploaded to the cloud platform. The collected values will be uploaded to the cloud platform; then the MPU6050 sensor detects the posture of the old man to determine whether the old man falls, if the old man falls, then the GSM communication module to send a message to the cell phone “Old man falls”, if the old man does not feel comfortable, he can dial a phone number through the key module. If the old man feels uncomfortable, he can make a call through the key module. The main flow chart is shown in Figure 26.

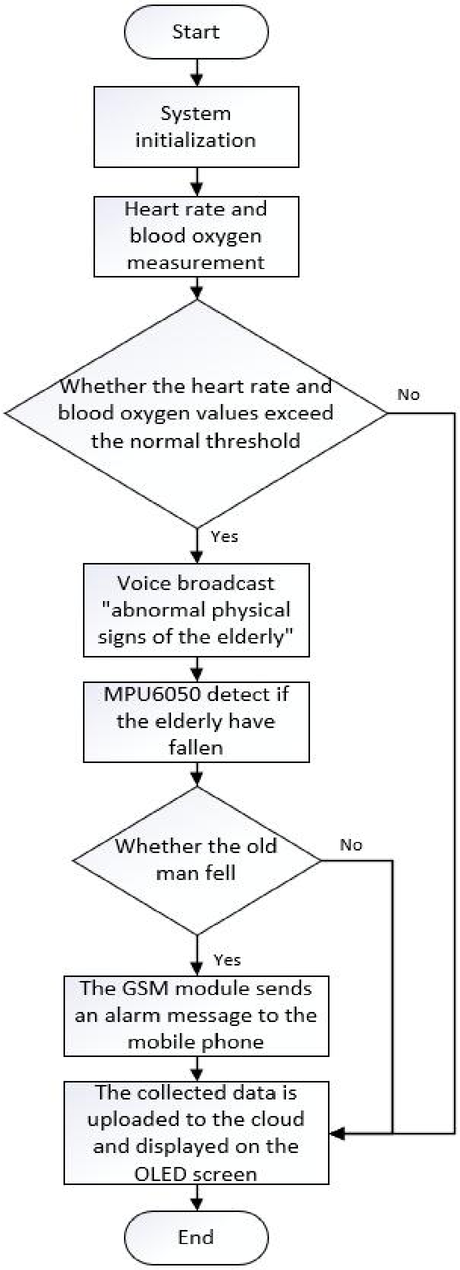

Start
System initialization
Heart rate and blood oxygen measurement
Whether the heart rate and blood oxygen values exceed the normal threshold
No
Yes
Voice broadcast "abnormal physical signs of the elderly"
MPU6050 detect if the elderly have fallen
Whether the old man fell
No
Yes
The GSM module sends an alarm message to the mobile phone
The collected data is uploaded to the cloud and displayed on the OLED screen
End

Figure 26 STM32 Main Program 2 Flowchart

4.3 GSM Communication Subsystem

According to the functional requirements of the system, insert the SIM card, after the system is powered on, the system is initialized, and the corresponding functional modules and pins are initialized, and then press the key 1 to receive the signal through the built-in antenna of the GSM module [6], and then demodulate the received signal to convert the analog signal into a digital signal, and the demodulated digital signal will be further processed and decoded to extract the voice, SMS and other information, the GSM module can send the processed data to other devices or servers, send the voice data to the cell phone user, and finally convert the processed digital signal to radio wave through the built-in radio transmitter to realize the functions of telephone call, sending SMS, etc. The program flow chart of the GSM communication subsystem is shown in Figure 27.

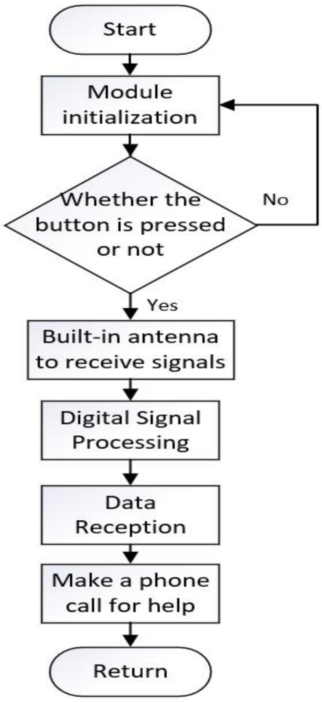

Start
Module initialization
Whether the button is pressed or not
No
Yes
Built-in antenna to receive signals
Digital Signal Processing
Data Reception
Make a phone call for help
Return

Figure 27 GSM Communication Subsystem Program Flowchart

4.4 Elderly Signs Detection Subsystem

According to the functional requirements of the system, the corresponding modules and pins are initialized first, and then MAX30102 heart rate and the blood oxygen sensor are used to collect the signs data of the elderly, and the heart rate and blood oxygen data are displayed on the OLED screen at the same time, and then the collected data are uploaded to the AliCloud platform through the WiFi module, so that the data can be viewed in real-time, and it can determine whether the signs value is within the normal range. If the value is out of the normal range, it will drive the voice module to broadcast “The old man's physical signs are abnormal”. The program flowchart of the elderly sign detection subsystem is shown in Figure 28.

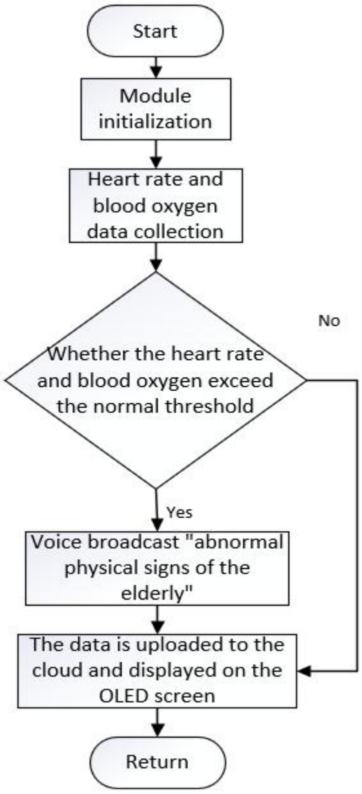

Start
Module initialization
Heart rate and blood oxygen data collection
Whether the heart rate and blood oxygen exceed the normal threshold
No
Yes
Voice broadcast "abnormal physical signs of the elderly"
The data is uploaded to the cloud and displayed on the OLED screen
Return

Figure 28 Flowchart of the Program of the Subsystem for Detecting Signs of the Elderly

4.5 MPU6050 Fall Detection Subsystem

According to the functional requirements of the system, the corresponding modules and pins are initialized first, and then the three-axis accelerometer can detect the acceleration in three directions at the same time, and the gyroscope [7] detects the rotational motion of the object, which can help to determine the attitude and direction of the object, and through the built-in Kalman filtering algorithm to process the data from the gyroscope and the accelerometer to determine whether the old man has fallen down or not, and if the old man falls, the GSM communication module will be driven to send the message “Old man falls” to the mobile phone. If the old man falls, it drives the GSM communication module to send the message “Old man falls” to the cell phone. The program flow chart of the MPU6050 fall detection subsystem is shown in Figure 29.

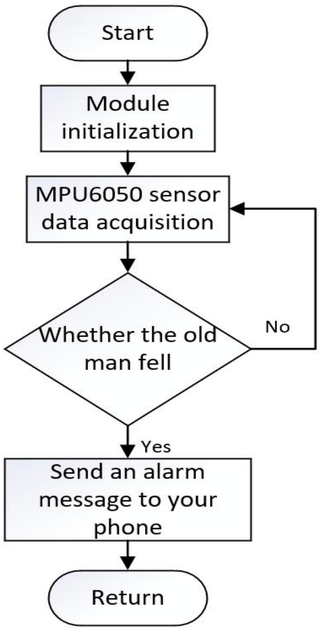

Start
Module initialization
MPU6050 sensor data acquisition
Whether the old man fell
No
Yes
Send an alarm message to your phone
Return

Figure 29 MPU6050 Fall Detection Subsystem Program Flowchart

5 CONCLUDING REMARKS

This paper designs a smart nursing home system based on the purpose of research and application, focusing on alleviating the existing problems such as the increasing trend of population aging, the mismatch between supply and demand of traditional institutionalized elderly care, and inefficient resource allocation [8], etc. This design innovates the use of low-cost materials to create a safe and comfortable smart elderly care environment and greatly reduces the waste of resources and duplication of inputs. However, this design still has some limitations, and there is still a certain gap with the practicality, but still hope that this design provides some ideas for the application of smart aging, and helps the senior care service industry to high-quality development. Meanwhile, the next research direction of this paper will focus on further optimizing the degree of intelligence of the system, monitoring the elderly's physical characteristics data more finely, improving the user experience to a greater extent, and exploring a more efficient and sustainable resource utilization model.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] LIU Xiaoxu, Lv Zhihua, HE Runhua. Intelligent system design of nursing home based on narrowband internet of things. Modern Information Technology, 2024, 8(02): 154-157.

[2] LIANG Taohua, ZHOU Jiang, XIAO Guannan. Design and realization of temperature and humidity meter based on DHT11. Electronic Fabrication, 2024, 32(15): 88-90.

[3] Zhao JC. Research on health detection system based on multifunctional sensors. Qingdao University, 2022.

[4] ZHU Min, XIN Yiyang, ZHAO Yangguang, et al. Fall alarm system for the elderly based on STM32 microcontroller control. University Physics Experiment, 2023, 36(05): 72-76.

[5] LI Ningning, WU Xingyu, LI Youxue, et al. Design and realization of human body alarm based on wireless communication. Technology and Innovation, 2024(23): 46-48.

[6] SHI Yongkang, LIU Tao, LIU Chen. Design of home burglar alarm system based on GSM SMS. Wireless Interconnection Technology, 2024, 21(18): 73-75.

[7] Du Peng. A low-power attitude and depth measurement device based on six-axis sensor MPU6050. Naval Electronics Engineering, 2022, 42(06): 168-170.

[8] Tian ZP. Research on Technology Application Enabling M Nursing Home Smart Elderly Service. Shanxi University of Finance and Economics, 2024.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3037

THE APPLICATION OF MULTIPLE IMPUTATION METHOD BASED ON HYBRID MULTI-STRATEGY IN HANDLING MISSING AIR QUALITY MONITORING DATA

ZhiQuan Zheng[1*], WenYong Zhang[1,2], ZhongChen Luo[3]
[1]School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, Guizhou, China.
[2]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, Guangdong, China.
[3]School of Nursing, Guizhou Medical University, Guiyang 550001, Guizhou, China.
Corresponding Author: ZhiQuan Zheng, Email: zhengzhiquan@gzmu.edu.cn

Abstract: Air quality monitoring data is a crucial basis for assessing air pollution levels and formulating control measures. However, missing data is a prevalent issue due to instrument malfunctions, human factors, and other reasons, significantly compromising data integrity and usability. To address this problem, this study collected nearly 1 million air quality monitoring records from 12 monitoring stations between 2015 and 2023, summarizing and analyzing the mechanisms and characteristics of missing data in such datasets. Data imputation experiments were conducted using R. Through missing mechanism control and imputation experimental design strategies, the imputation performance of algorithms was evaluated under the criteria of MAE, RMSE, and WMAPE based on completely random missingness. Specifically, data imputation experiments under different missing scenarios were repeated N times, and the mean values were used to evaluate four multiple imputation algorithms, with 95% confidence intervals provided. The experimental results show that: (1) the hybrid multi-strategy imputation method MNPRF demonstrates significant advantages across all datasets, with the smallest confidence limits and interval widths; (2) this method not only inherits the strengths of parent algorithms, substantially improving data quality, but also mitigates the weaknesses of the original algorithms to some extent.

Keywords: Air quality monitoring; Missing mechanism; Data imputation; Confidence interval; Multiple imputation; Hybrid multi-strategy imputation

1 INTRODUCTION

At present, with the increasing application of information analysis, data mining, and neural network model training in various industries, data loss has become an important problem in most application fields, such as statistical investigation [1], environmental protection [2], medical research [3], etc. Some studies have shown that more than 40% of the data sets in the international open database UCI have missing observations [4-5]. There are many reasons for missing data. Taking air quality monitoring data as an example, data may be lost due to equipment failure, environmental conditions, human management and other factors, such as sensor failure or damage caused by extreme weather conditions, data transmission loss caused by communication equipment failure, and failure to properly process and collect caused by software errors. In addition, the particularity of geographical location will sometimes affect the stability of the monitoring equipment, resulting in the inability to continuously record data. Missing data will not only lead to deviations in statistical results, but also lead to the unavailability of the original model [6].

The processing methods of data missing mainly include deletion method and filling method. Considering different missing scenarios, there are many specific ways to delete data, such as deleting columns with missing values, or deleting observed samples with missing values. However, while the deletion method is simple, it can miss useful information in the original data set and lead to incorrect statistical results. For this reason, statisticians and scholars in related fields do not recommend the use of erasure to deal with missing data. Aiming at the problem of missing data, many scholars devote themselves to the research of missing value filling, and have laid the theoretical foundation of this problem in statistics [7-9].

For the same missing value, the Data Imputation Algorithms is classified according to the number of its filling values, and the Data Imputation Algorithms is divided into two categories: Single Imputation(SI) and Multiple Imputation(MI)[10]. Single value filling includes mean value filling, mode filling, regression filling, etc. These methods can effectively solve the problem of missing data due to their advantages of high computational efficiency and strong interpretability. However, single value filling fills only one possible estimate for each missing value, ignoring the uncertainty of the missing data, and such methods will change the original distribution of the data, resulting in the distortion of the statistical characteristics of the data (such as variance and covariance). Multiple Imputation can effectively reflect the uncertainty of the data while dealing with the missing data. As a method used to process missing data, the core principle of Multiple Imputation is to fill in missing data by generating multiple possible interpolation values, and improve the accuracy and reliability of statistical analysis. Based on Bayesian statistics and sampling theory, the method generates Multiple Imputation values from the posterior distribution of missing data and simulates the possible distribution of missing data, thus improving the accuracy of estimates. The implementation methods of

Multiple Imputation fall into two main categories: Joint Modeling (JM) and Fully Conditional Specification (FCS) [11]. Joint modeling assumes that all variables (including missing and observed variables) obey some joint distribution (such as a multivariate normal distribution) and generates interpolation values from that distribution [8,12-13]. The Complete conditional gauge (FCS) is an iterative interpolation method, also known as Multiple Imputation by Chained Equations (MICE) [14]. Its core idea is to construct conditional models for each missing variable, such as regression model, random forest, etc., and update the interpolation value through iteration [15]. As a filling idea, Multiple Imputation not only reflects the uncertainty of missing data, but also can handle multiple types of data and is applicable to different Missing Data Mechanisms. In subsequent studies, scholars have proposed different versions of JM and FCS methods [16-22].

2 ALGORITHM INTRODUCTION AND IMPROVEMENT

2.1 Introduction of R Mice Package

The R mice package is an implementation tool for Multivariate Imputation by Chained Equations that is very useful in working with missing data. The package allows users to interpolate different types of variables through multiple models, thus improving the quality and reliability of data analysis. On the R platform, the mice package is installed through the install.packages() function and loaded using the library() function. The mice package provides a series of methods for managing and analyzing datasets that contain missing values. The main process consists of several stages such as creating multiple interpolating objects, performing the actual interpolating process, and summarizing the results. The main parameters of the mice() function include dataframe (the data set to be filled), m (the number of interpolations), maxit (the upper limit of iterations), method (the name of the specific algorithm used), and seed (random seed setting to ensure reproducibility).

Once the mice() function has been filled, you can use the complete() function to obtain one of the complete datasets, or you can use the loop structure to traverse all possible combinations of results. This paper uses norm.predict (Regression Prediction method) and RF (Random Forest) filling algorithms in the mice package of R language to perform data filling experiments on air quality monitoring data, and tries to improve the algorithm.

2.2 Norm.Predict Filling Algorithm Based on Multiple Imputation (MNP)

The MNP method is a realization of the regression prediction method. When dealing with missing data, this method utilizes complete covariates to construct a linear model and predict the missing values. Specifically: for the target variable with missing values, a linear regression model under a multivariate normal distribution is first established using the data without missing values. Then, for each observation record with missing values, the expected mean μ and standard deviation σ are obtained by substituting the known covariate values into the above trained model. Finally, a random number is drawn from the normal distribution $N(\mu,\sigma^2)$ as the filling result. This method can well maintain the original data structure characteristics while introducing reasonable uncertainty estimation.

2.3 Random Forest Filling Algorithm Based on Multiple Imputation (MRF)

Random forest is widely used in the field of machine learning. As one of the classical classification algorithms, it has good robustness and accuracy. The algorithm evolved from the decision tree, reduces the risk of overfitting in the decision tree, and is not sensitive to noise or outliers in the data set, so it has good prediction and generalization ability. The MRF method combines the idea of Multiple Imputation with a random forest model in the field of machine learning to estimate missing values by building multiple decision trees. For each tree, the importance of multiple features is taken into account during the node splitting process, and the model is trained using the unmissing data. When a missing worth sample is encountered, the most likely value is deduced according to the existing feature information as the filling result. This process is repeated many times to produce a stable and reliable prediction.

2.4 Characteristics Analysis of Air Quality Monitoring Data

The incomplete air quality monitoring data is mainly caused by the failure of acquisition equipment and other reasons, which brings difficulties to further experimental analysis. It is very important to analyze the Missing Data Mechanisms and reason of the data and choose the appropriate filling algorithm for its processing. In 1987, Little and Rubin [8] proposed the concepts of Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR), classifying the complex and diverse reasons for data missing. However, in view of the specific data missing problem, it is often necessary to explore the correlation among variables and the distribution of missing values in the data set to be filled. For this reason, before filling in the data, this paper pre-analyzed the missing rate of about 1 million air quality monitoring data from 12 stations under different observation indicators, each of which had 7 monitoring indicators, and plotted a stack bar chart (left of Figure 1) and a percentage stack bar chart (right of Figure 1). The experimental results are shown in Figure 1:

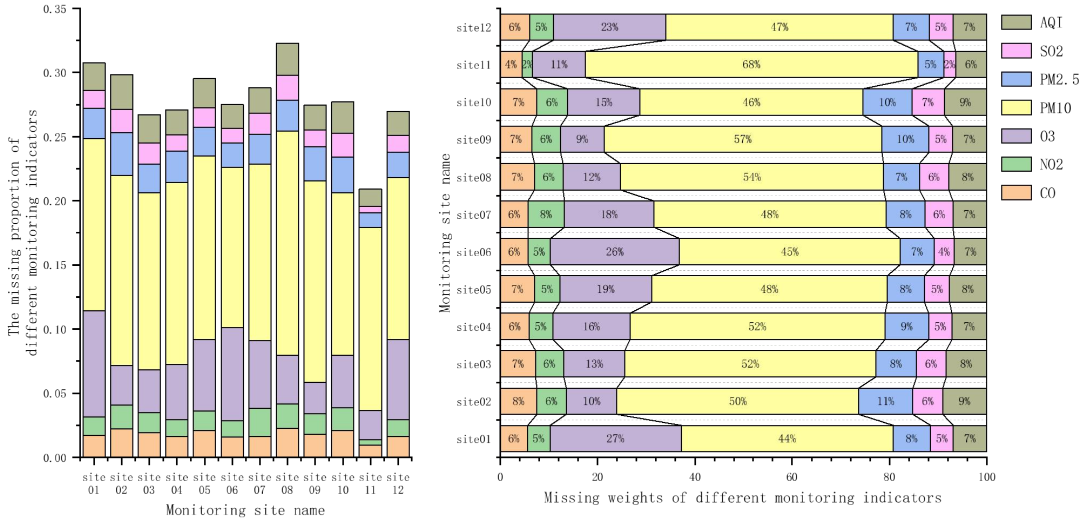


Figure 1 The Proportion of Missing Values of Different Observation Indicators at Different Monitoring Stations

The results of Figure 1 show that: First, in the air quality monitoring data, not only do all monitoring dimensions have missing data, but the missing rates of different monitoring dimensions are different in the air quality monitoring data of each station; Second, there are certain similarities in the missing rate of different stations under the same observation index. As can be seen from the experimental results in right of Figure 1, the missing rate of observation indicators CO, NO2, PM2.5, SO2 and AQI accounted for 5%-10% of the cumulative missing rate in all stations, while the missing rate of PM10 accounted for half of the total missing rate. Third, the experimental results of left of Figure 1 show that in all sites, the missing rate of monitoring indicator PM10 is above 15%, while the missing rate of the remaining vast majority of monitoring indicators is below 5%. In summary, for the air quality monitoring data, the missing trend of the internal observation indicators follows a similar rule, that is, in the air quality monitoring data of each station, only a few missing values of the monitoring dimension account for a relatively high proportion, more than 15%, while the missing values of the remaining most dimensions account for a very low proportion, less than 5%.

2.5 Hybrid Multi-strategy Interpolation Method Based on MICE

According to the missing proportion of different observed variables in the data set to be filled, a more targeted Data Imputation Algorithms is adopted to fill in different dimensions of the original data, which is the core idea of the algorithm improvement in this paper. Since the missing proportion of different monitoring dimensions of air quality data is highly differentiated, and there is a certain correlation between some observed variables, combined with the proportion distribution of missing values in each monitoring dimension of the data set to be processed, the observed variable with the largest missing proportion is given priority to be filled, and based on the initially filled data set, Another Data Imputation Algorithms was used to fill in the remaining observed variables. Considering the general filling effect of the algorithm, this paper improves the algorithm based on MRF and MNP. For the column with the largest proportion of missing in the same data set, the above two algorithms are used to fill in the first stage, and the result set after the initial filling is filled with another algorithm to fill in the missing values in the remaining observed variables. According to the processing order of the same data set by different filling algorithms, MRFNP(Random forest imputations and Linear regression mixed filling based on Multiple Imputation) and MNPRF(Linear regression and Random forest based on Multiple Imputation) are proposed imputations Mixed filling) algorithms are proposed. For ease of description, M represents the data set to be processed, $M_{i,j}, i \in [1,n], j \in [1,m]$ represents the element in the i th column of the j th row in M , n represents the number of sample points, m represents the total number of variables, $P_j, j \in [1,m]$ represents the proportion of missing values of the observation variable in column j of M . The specific filling process of the algorithm is as follows:
Step 1: Find the column with the largest missing rate by formula (1) :

$$j_{\max} = \left\{ j \mid P_j = \max\left(\frac{is.na(M_{i,j})}{nrow(M)} \right), i \in [1,n], j \in [1,m] \right\} \tag{1}$$

In formula (1), $\max(\cdot)$ is the maximum function, $is.na(\cdot)$ is a function for counting the number of observations in the jth column variable of M that are marked as NA, and $nrow(\cdot)$ is the sample quantity function in statistics;
Step 2: Generate the initial dataset M' to be filled, where $M' = M_{i,(-j_{\max})}$, and $M_{i,(-j_{\max})}$ represents the removal of the $j_{\max}$ column from M ;

Step 3: The data of M' is filled in by using MRF or MNP algorithms, and a complete data set M'_c is obtained;

Step 4: Concatenate M'_c and $M_{i,j_{\max}}$ column-wise to obtain a new dataset M'' that needs to be filled in;
Step 5: Based on the filling algorithm adopted in the third step, corresponding to another algorithm (namely MNP or MRF), data filling processing is carried out on M'' respectively, and then the final complete data set M_c is obtained.

3 RESEARCH METHODS

3.1 Data Sources and Experimental Environment

The operating system of all experiments in this paper is Windows 11, the program writing software is RStudio 2023.03.0 Build 386, the program execution kernel is R version 4.2.3, and the drawing tool is Origin. Considering that there may be some differences in air quality Monitoring data in different regions, in order to verify the effect of the algorithm under different data sets, this paper obtained data from CNEMC (China National Environmental Monitoring Centre, https://www.cnemc.cn/en/) collected air quality data of 12 monitoring stations. The monitoring indexes were CO, NO2, O3, PM10, PM2.5, SO2 and AQI, and the data collection frequency of each station was once an hour. The period is from 0:00 on Jan 1, 2015 to 23:00 on Dec 31, 2023. In the real data collection process, due to equipment failure, extreme weather impact, communication transmission problems and other factors, the collected data contains missing values, and even blank data for a period of time. In this paper, the collected data will be preprocessed, and the processing method is divided into the following two parts:
(1) For the data without recording time, this paper does delete processing; As for the data with time records, even if all monitoring indicators are missing, they are still retained, so as to obtain 12 data sets with missing values to be filled, and a total of 914,100 air quality monitoring records are obtained. In order to facilitate the description of subsequent experiments, this group of data is marked as $A_i, i \in [1,12]$;
(2) The records containing missing observed values were deleted for processing, so as to obtain 12 complete data sets, with a total of 741,679 air quality monitoring records. In order to facilitate the subsequent experimental description, this set of data was marked as $B_i, i \in [1,12]$. As the B_i records are all real data values, therefore, in the third part of the paper "comparison of experimental results", this set of data will provide a comparison basis for the advantages and disadvantages of MRF, MNP, MRFNP and MNPRF algorithms under different evaluation criteria. Taking into account the reasons for the absence of air quality monitoring data, this paper will conduct missing data processing for B_i based on complete random missingness by using computer simulation. The simulation method steps of the Missing Data Mechanisms are presented in Section 2.2.

3.2 Simulation of Missing Data Mechanisms

The absence of observed values in air quality monitoring data meets the definition of completely random absence. Therefore, all computer simulation experiments in this paper are conducted on the premise of completely random absence. The detailed simulation steps of the Missing Data Mechanisms are as follows:
Step 1: According to the experimental results shown in Figure 1, the number of records with missing values in different data sets varies. To better simulate the proportion of missing values in actual problems, this paper sets a general range for the proportion of missing values in the complete data set M, while the overall missing rate P_n^t of data set M is randomly generated within this range. That is to say, the proportion of records with missing values is determined from the row perspective. The formula is as follows:

$$P_n^t = runif\left(p_n \mid p_n \in [p_{\min}, p_{\max}]\right) \tag{2}$$

In Formula (2), P_n^t represents the proportion of observation records with missing values among the total number of records; $runif(\cdot)$ is a random number generation function; $p_{\min}$ and $p_{\max}$ respectively denote the upper and lower limits of the values that P_n^t can take, indicating the random selection of n values of p from $[p_{\min}, p_{\max}]$. Here, $n=1$, that is $P_n^t = \left(p_1^t\right)$.

Step 2: Based on the value of P_1^t, randomly select data rows from dataset M. The set of row numbers corresponding to these observation records is denoted as R_m, and R_m is determined by formula (3):

$$R_m = sort(sample(row(M), floor(p_t \times nrow(M)))) \tag{3}$$

In formula (3), $row(\cdot)$ is the extraction row number function, which is used to obtain the row number set corresponding to each observation data in M. $row(M)$ is a vector and is denoted as $V_\alpha, \alpha \in [1,n]$. $nrow(\cdot)$ is used to obtain the total number of records in M, $floor(\cdot)$ represents the floor function, $floor(p_t \times nrow(M))$ indicates the total number of records in M that contain missing observations, and this is counted as N_m. $sample(\cdot)$ is a random sampling function, indicating the random extraction of N_m elements without replacement from V_α ; $sort(\cdot)$ represents a sorting function, here in ascending order, used to arrange the randomly selected elements in the order from small to large.

Step 3: In real-world application scenarios, the randomness of whether each observation record is missing a certain observation indicator is also completely random. Based on this, in this paper, from the perspective of columns, the random missing processing is carried out for each observation indicator of each record corresponding to R_m , in order to simulate the missing situation of the air quality monitoring dataset in real scenarios to the greatest extent. There are a total of 11 observed variables in M . Among them, 4 are time-related records, namely year, month, day and hour; there are 7 air quality indicators, and the corresponding column numbers are denoted as $V_\phi^m, \phi \in [5,11]$. During the experiment execution, only the missing values of the air quality indicators were handled. Firstly, through formula (4), a set of random missing weight combinations P_β^w is generated for the 6 observed variables in M . The larger the value is, the greater the possibility of missingness for the current record in the corresponding observed variable is; conversely, the smaller the value is, the lower the possibility is.

$$P_\beta^w = runif\left(p_\beta^w \mid p_\beta^w \in [p_{\min}, p_{\max}]\right), \beta \in [1,6] \tag{4}$$

In formula (4), $runif(\cdot)$ is a random number generation function, while $p_{\min}$ and $p_{\max}$ respectively represent the upper and lower limits of the values that P_β^w can take, that is, randomly select 6 values of p^w from $[p_{\min}, p_{\max}]$, and at this time $P_\beta^w = \left(p_1^w, p_2^w, \cdots, p_6^w\right)$.

Step 4: Taking into account the missing characteristics of the air quality monitoring data shown in Figure 1, it is necessary to assign a higher probability of missingness to one of the remaining observation variables in M , so as to make the computer simulation experiment more closely resemble the real application scenarios. The initial probability value P_β^w is optimized through formula (5) to obtain the probability value $P_{\beta+1}^w$.

$$P_{\beta+1}^w = sample\left(\frac{c\left(P_\beta^w, a \times sum\left(P_\beta^w\right)\right)}{(a+1) \times sum\left(P_\beta^w\right)}\right) \tag{5}$$

In formula (5), $sample(\cdot)$ is a random sampling function, $sum(\cdot)$ is a summation function, $c(\cdot)$ is a vector concatenation function, a is a constant term, and $a > 0$. In the subsequent experiments, by controlling the value of a , the superiority of the proposed hybrid imputation algorithm in this type of missing scenario can be verified under different combinations of missing columns.

Step 5: For each record corresponding to R_m , the actual missing column $C_\varphi^{R_m}, \varphi \leq 7$ is generated through formula (6).

$$\begin{cases} C_\varphi^{R_m} = sort\left(sample\left(U, \varphi, W_1\right), l_1\right) \\ \varphi = sample\left(\left[1, length(U)\right], 1, sort\left(W_2, l_2\right)\right) \\ W_2 = sample\left(c\left(P_\beta^w, sum\left(P_\beta^w\right)\right) / 2sum\left(P_\beta^w\right)\right) \end{cases} , U = V_\phi^m, W_1 = P_{\beta+1}^w, l_1 = 0, l_2 = 1 \tag{6}$$

In formula (6), $C_\varphi^{R_m}$ represents the specific missing columns in each row record corresponding to R_m ; U represents the data source to be extracted; φ represents the number of samples extracted from U ; and the value of φ determines the number of elements contained in C_φ^R . W_1 represents the missing weights or probability distribution of each element in U . In this paper, non-uniform random selection operations are realized through W_1 . $sum(\cdot)$ represents the summation function; $sort(\cdot)$ represents the sorting function. When the value of parameter l is 0, it indicates ascending order; when it is 1, it indicates descending order. It is used to arrange the randomly selected elements in the prescribed order.

Step 6: To facilitate the comparison of subsequent experimental results, first, the corresponding observed values in M are saved in sequence. Then, the missing values in the data are handled by setting the corresponding values in $M_{i,j}, i = R_m, j = C_\varphi^{R_m}$ as "NA".

3.3 Experimental Methods

In real life, the overall proportion of missing data is not fixed, and there is uncertainty about whether the observed index is missing in each record. In order to verify the accuracy and stability of the Data Imputation Algorithms, this paper conducts experiments in five steps:

Step 1: Select an arbitrary data set from for the experiment, and denote this data set as M .
Step 2: On the R platform, the Missing Data Mechanisms method described in this paper is adopted to conduct a completely random missing treatment on the overall missing proportion P_n^t of M within the range of proportions $[3\%,7\%]$, $[13\%,17\%]$ and $[23\%,27\%]$, thereby obtaining a non-complete data set M^*.
Step 3: Step 3: For each M^* under current P_n^t , apply 4 different imputation algorithms to fill in the missing values once. After the imputation is completed, evaluate the deviation degree of the imputation results from the original values in M under different evaluation criteria for different imputation algorithms when the current missing rate is considered.
Step 4: Taking into account the randomness of computer simulation, for each different range of P_n^t , this paper repeats the third step experiment operation 100 times for each case, thereby obtaining the mean, standard deviation and confidence interval of the evaluation results after Multiple Imputations [23].
Step 5: To verify the effectiveness of the data imputation algorithm, the experimental method from step 2 to step 4 was repeated for each data set in $B_i, i \in [1,12]$. The mean and median of the evaluation results after multiple imputations were given to verify the superiority of the algorithm under different data sets.

3.4 Evaluation Criteria

In order to compare the effects of different filling algorithms, this paper assumes $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n\}$, $Y = \{y_1, y_2, \cdots, y_n\}$, where $\hat{Y}$ represents the filling value, Y represents the original value, and n represents the number of missing values in the dataset M^* that needs to be filled. Based on this, the experiment makes a comparison of the results from two aspects of absolute error and relative error. Three evaluation criteria, namely MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and WMAPE (Weighted Mean Absolute Percentage Error), are selected to evaluate the algorithm presented in the paper. Among them, MAE and RMSE respectively represent the absolute error between $\hat{Y}$ and Y . The formulas are defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{8}$$

In formula (7)-(8), it can be seen that the ranges of MAE and RMSE are both within $[0,+\infty]$. The closer the values are to 0, the smaller the deviation between the imputed values and the true values, which indicates that the performance of the data imputation algorithm is better. Compared with MAE, RMSE also reflects the stability of the deviation degree. It is more sensitive to outliers. A smaller value not only indicates a better imputation effect but also reflects the stability of the imputation algorithm.
In real life, since the value range of the observed variables to be filled may be very different, if only MAE and RMSE are considered, the degree of error between the filled value and the true value is often unable to reflect the degree of error relative to the true value itself. Based on this, this paper adopts the statistical quantity WMAPE, which can represent the relative error between $\hat{Y}$ and Y , to evaluate the superiority of different algorithms. The definition of WMAPE is as follows:

$$WMAPE = \sum_{i=1}^{n}|\hat{y}_i - y_i| \Big/ \sum_{i=1}^{n} y_i \tag{9}$$

Compared with $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$, WMAPE is less sensitive to outliers in Y . That is, when y_i approaches 0 or equals 0, it can still effectively reflect the relative error results among different algorithms. The range of WMAPE remains $[0,+\infty]$, and the closer the value is to 0, the better the filling effect is.

Taking into account the experimental methods of this paper, after Multiple Imputations under the same P_n^t range, different evaluation result vectors will be obtained, as follows:

$$MAE_\lambda = \{MAE_1, MAE_2, \cdots, MAE_N\} \tag{10}$$

$$RMSE_\lambda = \{RMSE_1, RMSE_2, \cdots, RMSE_N\} \tag{11}$$

$$WMAPE_\lambda = \{WMAPE_1, WMAPE_2, \cdots, WMAPE_N\} \tag{12}$$

The evaluation criteria used in this paper are derived from traditional MAE, RMSE and WMAPE calculations, and are defined as follows:

$$\overline{MAE} = \frac{1}{N}\sum_{\lambda=1}^{N} MAE_{\lambda} \tag{13}$$

$$\overline{RMSE} = \frac{1}{N}\sum_{\lambda=1}^{N} RMSE_{\lambda} \tag{14}$$

$$\overline{WMAPE} = \frac{1}{N}\sum_{\lambda=1}^{N} WMAPE_{\lambda} \tag{15}$$

In Formulas (13)-(15), N represents the number of fillings within the current P_n^t range, λ represents the λ th filling experiment, MAE_{λ} , $RMSE_{\lambda}$, and $WMAPE_{\lambda}$ respectively represent the values of MAE, RMSE and WMAPE obtained in the λ th experiment, and $\overline{MAE}$, $\overline{RMSE}$, and $\overline{WMAPE}$ respectively represent the mean values of the results under different evaluation criteria in multiple rounds of experiments.

4 RESULTS AND DISCUSSION

4.1 Experimental Results Based on Air Quality Monitoring Data in a Certain Region

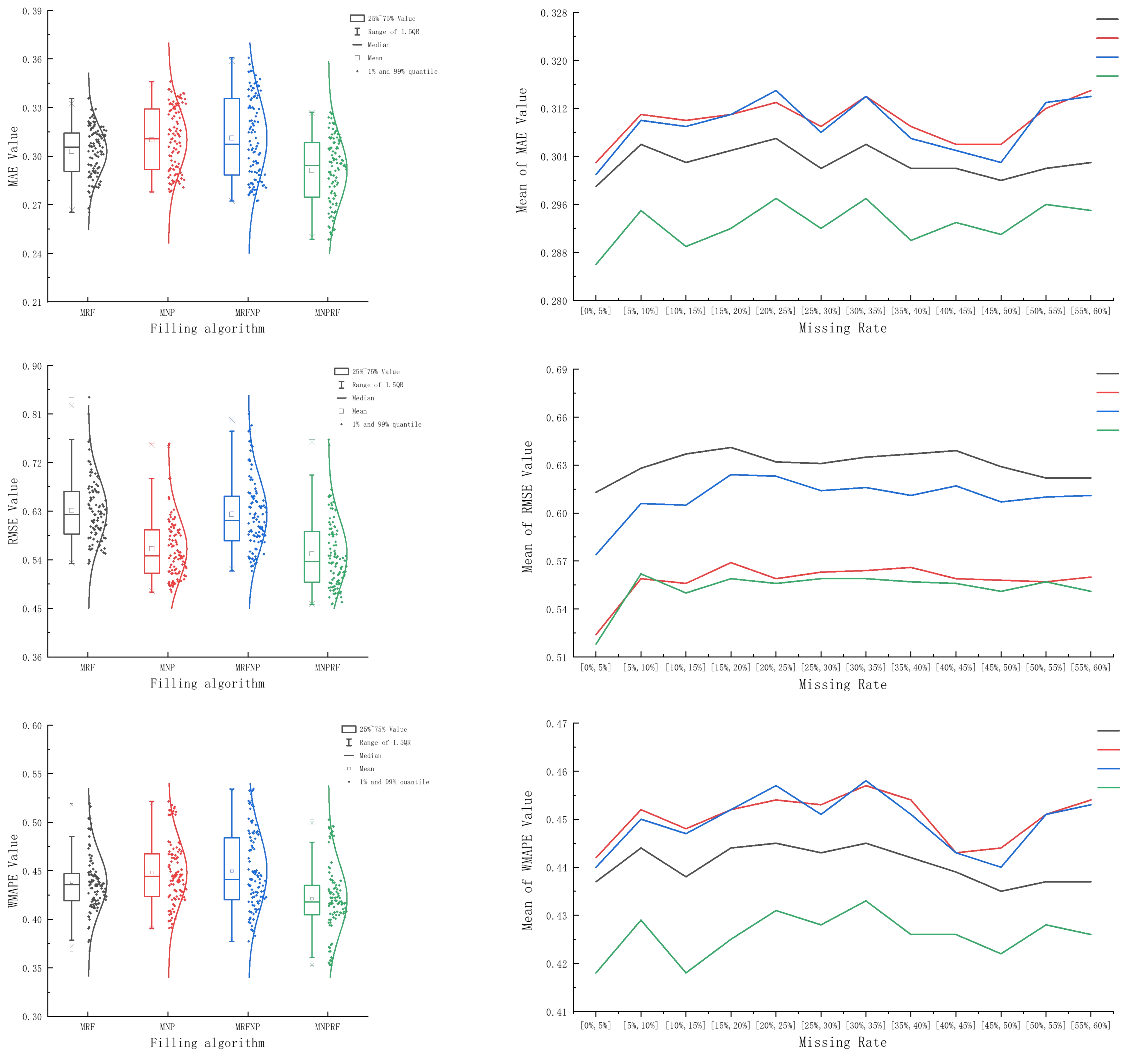
To facilitate the subsequent experimental description, let the mean value of the overall missing rate range be $\tau = \frac{p_{\min} + p_{\max}}{2}$, and the maximum weight of the missing column be $\omega = \frac{a}{a+1}$. Considering that in real scenarios, the overall missing rate of the dataset to be filled is not fixed, in order to simulate the effectiveness of the imputation algorithm under different missing rate ranges, this paper stipulates $p_{\max} - p_{\min} = 4\%$, and the values of $p_{\min}, p_{\max}$ are all accurate to 0.01, and a is a positive integer. During the experiment, $\tau = 0.05, 0.15, 0.25$, $\omega = 0.5, 0.8, 0.9$, the number of experimental repetitions $N = 100$, and the experimental results are as shown in Table 1:

Table 1 Under Different Loss Rates and Loss Weights, Four Filling Algorithms Fill 100 Times under Different Evaluation Criteria

Weight	Algorithm	$\tau = 0.05$			$\tau = 0.15$			$\tau = 0.25$		
		$\overline{RMSE}$	$\overline{MAE}$	$\overline{WMAPE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{WMAPE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{WMAPE}$
$\omega = 0.5$	MRF	0.634	0.306	0.442	0.625	0.300	0.437	0.634	0.301	0.436
	MNP	0.550	0.309	0.447	0.556	0.306	0.445	0.560	0.309	0.448
	MRFNP	0.616	0.310	0.449	0.613	0.304	0.442	0.616	0.306	0.444
	MNPRF	0.548	0.297	0.430	0.556	0.294	0.428	0.556	0.297	0.430
$\omega = 0.8$	MRF	0.634	0.307	0.446	0.631	0.305	0.441	0.636	0.303	0.442
	MNP	0.548	0.309	0.449	0.553	0.306	0.443	0.552	0.307	0.448
	MRFNP	0.606	0.309	0.449	0.614	0.307	0.444	0.605	0.307	0.449
	MNPRF	0.541	0.292	0.423	0.547	0.290	0.420	0.549	0.290	0.423
$\omega = 0.9$	MRF	0.633	0.306	0.445	0.627	0.301	0.438	0.629	0.303	0.439
	MNP	0.555	0.308	0.447	0.553	0.307	0.446	0.554	0.310	0.449
	MRFNP	0.610	0.308	0.448	0.608	0.309	0.448	0.610	0.311	0.451
	MNPRF	0.545	0.287	0.418	0.542	0.287	0.417	0.548	0.292	0.423

The experimental results in Table 1 show that under 3 different miss rate ranges and weight distributions, 4 algorithms fill 100 randomly generated incomplete data sets respectively, and then obtain the corresponding mean value of evaluation results. The above experimental results show that: (1) under MAE, RMSE and WMAPE evaluation criteria, the filling effect of MRF, MNP, MRFNP and MNPRF algorithms will not vary widely with the gradual increase of the missing rate, which preliminarily proves that the filling effect of the multi-filling algorithm based on the multi-interpolation idea is stable under different missing rates; (2) Under different ranges of missing rates and weight distributions, the MNPRF algorithm achieved the lowest A, B, and C results, and its overall filling effect was the best. The MNP algorithm's filling effect was second-best; (3) When $\omega = 0.5$ is true, the superiority of the MNPRF algorithm is not significant under the three evaluation criteria. However, as ω increases, the filling advantage of this algorithm becomes greater. In summary, the proposed algorithm improvement idea can inherit and extend the advantages of the existing algorithm, and has certain portability, and improve the filling optimization method from different perspectives.

In order to further verify the filling effects of the four algorithms under different missing rates, in this paper, the range of missing rates $[p_{\min}, p_{\max}]$ of the data set to be filled is gradually increased from $[0\%,5\%]$ to $[55\%,60\%]$. The step size of the upper limit $p_{\max}$ and the lower limit $p_{\min}$ of the missing rate range is set to 5% respectively, and the number of experiments for each missing rate range is $N=100$. The results of $\overline{MAE}$, $\overline{RMSE}$ and $\overline{WMAPE}$ are plotted under three criteria, and the experimental results are shown in Figure 2 (right). More importantly, in order to explore the stability of multiple experiments of different algorithms under MAE, RMSE and WMAPE criteria, this paper presents the box plot, result drop point and distribution curve of the miss rate similar to the real data. The experimental results are shown in left of Figure 2.



(**a**) (**b**)

Figure 2 (**a**) When $\tau=0.15,\omega=0.8,N=100$ is Set, Box Plots of Four Filling Algorithms under Different Evaluation Criteria; (**b**) When $\omega=0.8,N=100$ is Set, the Mean Line Graphs of MAE, RMSE and WMAPE Results Of The Four Filling Algorithms Under Different Missing Rates Are Presented

Figure 2 of (**a**) presents the box plots and distribution curves of the evaluation results obtained from 100 data imputation experiments conducted by MRF, MNP, MRFNP and MNPRF respectively when the range of missing rate is $[13\%,17\%]$ and the maximum weight of the missing column is 80%. It can be seen that: (1) MNPRF algorithm always maintains the minimum mean, median, upper quartile, lower quartile and other statistics under MAE, RMSE and WMAPE evaluation criteria, and has the best filling effect; (2) Under the MAE evaluation criterion, the box length of the MNPRF algorithm is only greater than that of the MRF algorithm. Moreover, the MAE_{λ} distribution of the MNPRF algorithm is more similar to the normal distribution compared to that of the MRF algorithm. This further validates that although the MNPRF algorithm is based on the MNP algorithm, it still retains a significant amount of the filling advantages of the MRF algorithm; (3) Under the RMSE evaluation criteria, the distance between the top and bottom quarterback values of the MNP algorithm is the smallest, followed by the MNPRF algorithm, and the distribution curves of the evaluation results of the two algorithms are similar, which indicates the system stability of the

MNPRF algorithm; (4) WMAPE value reflects the relative error of the algorithm. Under this evaluation criterion, MNPRF not only obtains the smallest spacing between top and bottom quarterbacks, but also the mean and median WMAPE values of the algorithm are even smaller than the lower quarterback values of the other three algorithms, showing significant filling advantage.

The experimental results in Figure 2 of (**b**) more directly demonstrate: (1) the filling advantage of MNPRF algorithm under different miss rates is significantly better than the other three algorithms under MAE and WMAPE criteria, while the filling effect under RMSE criteria is slightly better than MNP algorithm, but still significantly better than MRF and MRFNP algorithms; (2) The filling effect of MRF, MNP, MRFNP and MNPRF algorithms based on the idea of Multiple Imputation does not increase significantly with the increase of missing columns, and the filling effect is stable.

In the process of the experiment, considering the factors such as the sample size and the time complexity of the algorithm, combined with the uncertainty brought by the computer random simulation, this paper only repeated the experiment 100 times under each missing rate range, which is slightly insufficient to evaluate the overall filling effect of the algorithm. To further verify the overall differences among the aforementioned imputation algorithms, this paper uses MAE_λ, $RMSE_\lambda$, and $WMAPE_\lambda$ as new samples to construct a 95% confidence interval, thereby evaluating the accuracy of the imputation effects of different algorithms. Since the overall variance of the new sample is unknown, and according to the experimental results in Figure 2, the data MAE_λ, $RMSE_\lambda$, and $WMAPE_\lambda$ generated by different algorithms in this sample do not fully satisfy the normal distribution. Therefore, in this paper, formulas (16)-(18) are used to construct the Confidence Interval for the mean absolute error (MAE_{CI}), the root mean square error ($RMSE_{CI}$), and the mean absolute percentage error ($WMAPE_{CI}$), and the confidence interval lengths CIL (Confidence Interval Length) are respectively given.

$$MAE_{CI} = \left(\overline{MAE} - Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=1}^{N} \left(MAE_\lambda - \overline{MAE} \right)^2 \Big/ N^2}, \overline{MAE} + Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=1}^{N} \left(MAE_\lambda - \overline{MAE} \right)^2 \Big/ N^2} \right) \tag{16}$$

$$RMSE_{CI} = \left(\overline{RMSE} - Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=1}^{N} \left(RMSE_\lambda - \overline{RMSE} \right)^2 \Big/ N^2}, \overline{RMSE} + Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=1}^{N} \left(RMSE_\lambda - \overline{RMSE} \right)^2 \Big/ N^2} \right) \tag{17}$$

$$WMAPE_{CI} = \left(\overline{WMAPE} - Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=1}^{N} \left(WMAPE_\lambda - \overline{WMAPE} \right)^2 \Big/ N^2}, \overline{WMAPE} + Z_{1-\frac{\alpha}{2}} \sqrt{\sum_{j=1}^{N} \left(WMAPE_\lambda - \overline{WMAPE} \right)^2 \Big/ N^2} \right) \tag{18}$$

In Equations (16)-(18), the value of α is set at 0.05. By calculation, $Z_{1-\frac{\alpha}{2}} = 1.96$ and N represent the sample quantities, which are the number of experiments conducted under different missing rate ranges in this paper. The experimental results are shown in Table 2 as follows:

Table 2 Under Different Missing Rates, Four Filling Algorithms Filled the Confidence Interval Generated under Different Evaluation Criteria 100 Times

Missing Rate	Algorithm	$RMSE_{CI}$	CIL	MAE_{CI}	CIL	$WMAPE_{CI}$	CIL
$\tau = 0.05$	MRF	[0.618, 0.650]	0.032	[0.304, 0.310]	0.006	[0.440, 0.452]	0.012
	MNP	[0.533, 0.563]	0.030	[0.306, 0.313]	0.007	[0.442, 0.455]	0.013
	MRFNP	[0.590, 0.623]	0.034	[0.304, 0.314]	0.010	[0.441, 0.457]	0.016
	MNPRF	[0.524, 0.559]	0.035	[0.288, 0.296]	0.009	[0.416, 0.431]	0.015
$\tau = 0.15$	MRF	[0.620, 0.643]	0.023	[0.300, 0.306]	0.006	[0.432, 0.444]	0.012
	MNP	[0.549, 0.572]	0.023	[0.306, 0.314]	0.008	[0.442, 0.455]	0.013
	MRFNP	[0.611, 0.637]	0.026	[0.306, 0.316]	0.010	[0.442, 0.458]	0.016
	MNPRF	[0.538, 0.564]	0.025	[0.287, 0.295]	0.008	[0.414, 0.428]	0.014
$\tau = 0.25$	MRF	[0.627, 0.645]	0.018	[0.300, 0.306]	0.005	[0.436, 0.448]	0.012
	MNP	[0.544, 0.560]	0.017	[0.303, 0.310]	0.007	[0.441, 0.454]	0.014
	MRFNP	[0.597, 0.614]	0.017	[0.303, 0.312]	0.009	[0.441, 0.456]	0.015
	MNPRF	[0.537, 0.560]	0.023	[0.286, 0.294]	0.008	[0.416, 0.431]	0.015

Statistical analysis of the experimental results in Table 2 with 95% confidence shows that:The MNPRF algorithm obtained the minimum lower and upper bounds of $RMSE_{CI}$, MAE_{CI} and $WMAPE_{CI}$ respectively under different missing rate ranges, which further verified the experimental results in Table 1; (2) Under different ranges of missing rates, the upper limit values of MAE_{CI} and $WMAPE_{CI}$ in the MNPRF algorithm are significantly lower than those of the other three algorithms; (3) Under different ranges of missing rates, the upper limit value of $RMSE_{CI}$ in the MNPRF algorithm is always smaller than the lower limit value of $RMSE_{CI}$ in both the MRF and MRFNP algorithms, and the starting point of the interval of the MNPRF algorithm is always smaller than that of the MNP algorithm; (4) From the perspective of CIL , under the 95% confidence level, the confidence interval length values of MNPRF and MRFNP algorithms are consistently greater than those of MRF and MNP algorithms in all three evaluation criteria. In conclusion, the improved MNPRF algorithm has some values of $RMSE_{CI}$, MAE_{CI} , and $WMAPE_{CI}$ increasing due to the influence of certain special values in the data set to be filled. This affects the overall filling accuracy of the algorithm. However, its overall filling effect is significantly better than that of MRF and MNP algorithms.

4.2 Experimental Results Based on Air Quality Monitoring Data of Different Stations

In order to verify the general applicability of the algorithm in the field of air quality monitoring data, this paper conducts experiments on 12 datasets collected by adopting the same experimental method. Considering the missing rate situation of the original real datasets, the P_n^t value of this part of the experiment refers to the missing rate of the real datasets, that is $p_{\min} = 0.1, p_{\max} = 0.16$. Each dataset is executed 100 times with the same experimental steps, and all experimental results are plotted under the MAE, RMSE, and WMAPE criteria. The details are shown in Figure 3, Figure 4, and Figure 5:

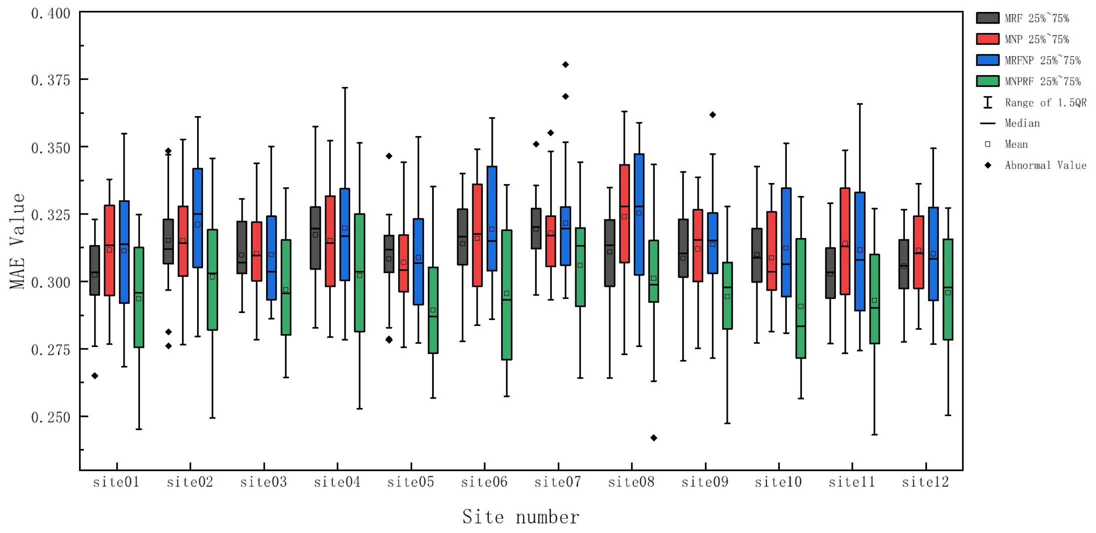


Figure 3 When $P_n^t \in [10\%,16\%], N = 100$ is Set, Box Plot of MAE Results from 4 Algorithms Based on Different Datasets for Imputation Experiments

Figure 3 Experimental results show that: (1) Under MAE evaluation criteria, the MAE mean, median, upper quartile, lower quartile and 1.5x interval of the MNPRF algorithm in all regional air quality monitoring data sets are significantly lower than the corresponding statistics of the other three algorithms; (2) In the experimental results of all regions, the MNPRF algorithm has a longer interquartile interval, while the MRF algorithm has the shortest interquartile interval. To sum up, MNPRF algorithm has significant advantages in filling different data sets.

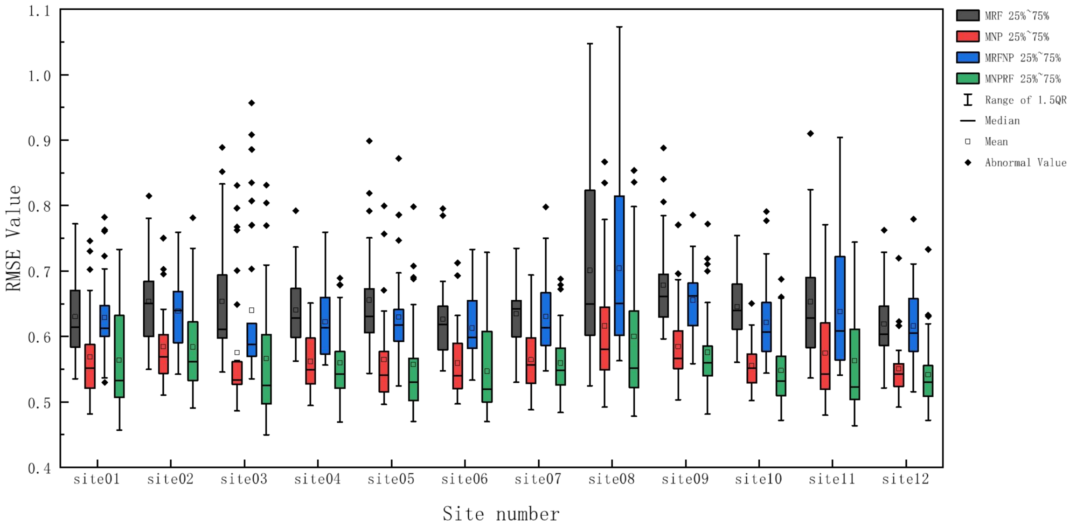


Figure 4 When $P_n^t \in [10\%,16\%], N = 100$ is set, Box Plot of RMSE Results from 4 Algorithms Based on Different Datasets for Imputation Experiments

Figure 4 Experimental results show that: (1) Under the RMSE evaluation criteria, the RMSE mean, median, upper quartile, lower quartile and 1.5x interquartile interval values of MNPRF algorithm in all regional air quality monitoring data sets are significantly lower than the corresponding statistics of MRF and MRFNP algorithms, and slightly lower than MNP algorithm; (2) In the experimental results of all monitoring stations, the interquartile spacing of the four algorithms had no obvious rule, and there were a few outliers in the RMSE experimental results of almost all stations. In summary, the MNPRF algorithm has the best system stability in different data sets.

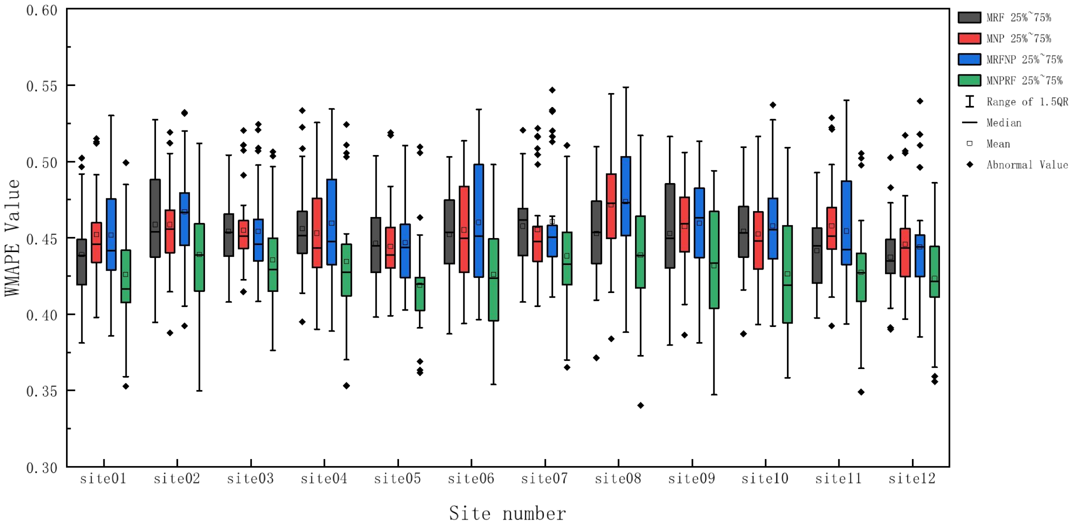


Figure 5 When $P_n^t \in [10\%,16\%], N = 100$ is set, Box Plot of WMAPE Results from 4 Algorithms Based on Different Datasets for Imputation Experiments

Figure 5 Experimental results show that: (1) Under the WMAPE evaluation criteria, the mean, median, upper quartile, lower quartile and 1.5x interquartile of WMAPE in all regional air quality monitoring data sets of MNPRF algorithm have the smallest values; (2) In almost all monitoring site experiments, the mean and median values of the evaluation results of the MNPRF algorithm were smaller than the lower quartile values of the other three algorithms; (3) In the experimental results of all monitoring sites, the quartile spacing of the four algorithms has no obvious rule, and the upper and lower quartile spacing of the MNPRF algorithm still has a good performance in the data filling experiments of some sites. In summary, the relative error of MNPRF algorithm in different data sets is the smallest, and the filling effect is the best.

5 CONCLUSIONS

With the application and popularity of machine learning and neural network algorithms in all walks of life, the scale and quality of data have become increasingly important, and missing value processing has become the most important part of data pre-processing. For the same data set to be filled, data analysts often need to choose a suitable Data Imputation Algorithms according to data characteristics, missing reasons, data scale and other factors. Considering the complexity of practical problems, high-dimensional data and multi-source heterogeneous data are becoming more and more common in the current real application scenarios, which leads to the cause of missing data of different dimensions becoming no longer single. Meanwhile, filling algorithms dealing with missing observed values of different dimensions in the same data set may also be different. Based on the original data imputation algorithms MRF and MNP, this paper makes improvements, tries to use different algorithms to deal with the missing value problem of different monitoring dimensions in air quality monitoring data, and proposes two algorithms MRFNP and MNPRF according to the kernel execution order of the improved algorithm. The experimental results of this paper show that the filling effect and accuracy of the algorithm can be greatly improved by selecting a more appropriate filling algorithm for different missing dimensions of the same data set. The algorithm improvement concept in this paper provides a new idea for the future development of the data filling field.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This research was funded by Guizhou Minzu University Fund Project (Youth Project) (No. GZMUZK[2023]QN12), the Guizhou Provincial Basic Research Program (Natural Science) Young Scholars Guidance Project (No. [2024]208), the Guizhou Provincial Basic Research Program(Natural Science) (No. zk2025-536).

REFERENCES

[1] Di Z, Guarnera U, Luzi O. Imputation through finite Gaussian mixture models. Computational Statistics and Data Analysis, 2007, 51: 5305-5316.

[2] Junninen H, Niska H, Tuppurainen K, et al. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 2004, 38: 2895-2907.

[3] Sabine V, Karlien V B, Peter G. Sequential imputation for missing values. Computational biology and chemistry, 2007, 31, 320-327.

[4] Pedro J, Garcia L, Jose-Luis S G, et al. K nearest neighbours with mutual information for simultaneous classification and missing data imputation, Neurocomputing, 2009, 72: 1483-1493.

[5] Wu S, Feng X D, Shan Z G. Missing Data Imputation Approach Based on Incomplete Data Clustering.Chinese journal of computer, 2012, 35: 1726-1738.

[6] Sethia K, Gosain A, Singh J. Review of Single Imputation and Multiple Imputation Techniques for Handling Missing Values. Lecture Notes in Networks and Systems, 2023, 730: 33-50.

[7] Rubin D B. Inference and Missing Data. Biometrika, 1976, 63: 581-592.

[8] Little R, Rubin D B. Statistical Analysis With Missing Data; Wiley and Sons Inc: New York, USA, 1987.

[9] Enders C K. Applied Missing Data Analysis. Guilford Press: New York, USA, 2010.

[10] Sebastian J, Arndt A, Felix B. A Benchmark for Data Imputation Methods. Frontiers in big data, 2021, 4, 674-693.

[11] Hakan D. Flexible Imputation of Missing Data. Journal of Statistical Software, 2018, 85: 1-5.

[12] Schafer J L. Analysis of Incomplete Multivariate Data. Chapman & Hall: Oxfordshire, UK, 1997.

[13] Schafer J L, Yucel R M. Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics, 2002, 11: 437-457.

[14] Stef V B. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 2007, 16: 219-242.

[15] Van B S, Brand J P L, Groothuis-Oudshoorn C G M, et al. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 2006, 76: 1049-1064.

[16] Yusuke Y, Toshihiro M, Kazushi M. A comparison of multiple imputation methods for incomplete longitudinal binary data. Journal of Biopharmaceutical Statistics, 2018, 28: 645-667.

[17] Kim H J, Reiter J P, Wang Q, et al. Multiple Imputation of Missing or Faulty Values Under Linear Constraints. Journal of Business & Economic Statistics, 2014, 32: 375-386.

[18] Enders C K, Keller B T, Levy R. A fully conditional specification approach to multilevel imputation of categorical and continuous variables. Psychological methods, 2018, 23: 298-317.

[19] Vincent A, Ndeye N. Clustering with missing data: which equivalent for Rubin's rules? Advances in Data Analysis and Classification, 2023, 17: 623-657.

[20] Van B S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 2007, 16: 219-242.

[21] Goldstein H, Carpenter J, Kenward M G, et al. Multilevel models with multivariate mixed response types. Statistical Modelling, 2009, 9: 173-197.

[22] Yang Z. Diagnostic checking of multiple imputation models. AStA Advances in Statistical Analysis, 2022, 106: 271-286.

[23] Zhi Q Z, Yan C, Meng M W, et al. Research on Stability of Data Imputation Algorithms With Different Miss Rates. Statistics and The Decision, 2023, 33: 12-17.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3038

PREDICTING AND ANALYZING THEFT CRIME THROUGH TEMPORAL AND SPATIAL DIMENSIONS

SuZhen Luo[1], ZhiSong Wu[2], LiNing Yuan[3*]
[1]Ministry of Public Sports, Guangxi Police College, Nanning 53028, Guangxi, China.
[2]School of Public Policy and Management, Guangxi Police College, Nanning 53028, Guangxi, China.
[3]School of Information Technology, Guangxi Police College, Nanning 53028, Guangxi, China.
Corresponding Author: LiNing Yuan, Email: yuanlining@gxjcxy.edu.cn

Abstract: The data on theft crimes exhibits characteristics such as dynamism, correlation, and uncertainty in both temporal and spatial dimensions. The factors influencing the occurrence of these crimes are complex and include various elements such as population density, education levels, poverty rates, employment status, and climate conditions. The volume and diversity of this data often pose challenges for traditional situational awareness technologies, which rely on criminological theories and case analyses, making it difficult to meet the actual needs of public security agencies. Consequently, crime data mining algorithms based on machine learning and deep learning have gradually become mainstream. This article analyzes the temporal and spatial characteristics of theft crimes, utilizes the Prophet model to predict future incidents, and employs kernel density estimation functions to identify spatial hotspots of theft crimes.

Keywords: Theft crime; Prophet model; Kernel density estimation; Spatial hotspots

1 INTRODUCTION

The crime of theft, a prevalent form of criminal activity, significantly impacts public safety and the quality of life for residents. In recent years, the rapid advancement of technologies such as big data, artificial intelligence, and geographic information systems (GIS), crime prediction and hotspot analysis increasingly important in criminological research. By conducting a thorough analysis of the temporal and spatial distribution of theft crimes, it is possible to enhance the efficiency of police resource allocation and provide a scientific foundation for developing targeted prevention and control measures. This article aims to explore the latest advancements in the temporal prediction and spatial hotspot analysis of theft crimes and to propose recommendations for improving existing methods in conjunction with relevant technological tools.

In recent years, scholars have made significant advancements in the field of crime prediction. Chainey et al. [1] achieved accurate predictions of theft crimes by introducing spatiotemporal pattern recognition technology. Travaini et al. [2] employed machine learning algorithms to analyze various factors influencing crime occurrence, thereby enhancing the accuracy of prediction models. Meanwhile, Jenga et al. [3] proposed a deep learning-based crime prediction model that substantially improved the predictive capability regarding the temporal distribution of theft crimes. In the realm of spatial hotspot analysis, Johnson et al. [4] utilized GIS technology to identify high-crime areas in urban environments and proposed corresponding early warning mechanisms. Mondal et al. [5] discovered the clustering effect of theft crimes in specific regions through spatial autocorrelation analysis, further validating the existence of crime hotspots.

Despite the advancements made in crime prediction and hotspot analysis, several pressing issues remain to be addressed [6~10]. First, the diversity and complexity of data necessitate that researchers thoroughly consider various influencing factors when constructing models, including socioeconomic conditions, population density, and public security investment. Second, effectively integrating prediction results with actual police operations continues to pose a significant challenge. Finally, the variations in crime characteristics and patterns across different regions underscore the importance of localizing and adapting models, which should be a key focus for future research.

Through a systematic study of temporal prediction and spatial hotspot analysis of theft crimes, this article aims to provide valuable insights for researchers in related fields and to offer theoretical support and technical guidance for social security management efforts. In the time dimension, the Prophet model [11] is employed to decompose theft crimes into a three-part structure consisting of trend, seasonal, and event components, thereby facilitating interpretable modeling of crime time series patterns. The trend component utilizes a piecewise linear function with adaptive change point detection, which effectively captures the long-term trajectory of crime rates influenced by exogenous variables. The seasonal component quantifies the inherent periodic characteristics of criminal activities through Fourier series expansion, while the event component enables the model to assess the impact of specific public security measures on crime suppression using custom functions. In the spatial dimension, kernel density estimation functions [12] are applied to visualize the spatial distribution characteristics of crime, providing a quantitative foundation for identifying crime hotspots and formulating prevention strategies.

2 MODELS AND ALGORITHMS

2.1 Prophet Model

The Prophet model is an open-source tool developed by Facebook for time series forecasting. The core of the algorithm includes an additive model and Bayesian inference. The additive part is typically represented as:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t \tag{1}$$

where, *g*(*t*) represents the trend component, which indicates the long-term growth or decline in the time series; *s*(*t*) represents the seasonal component, which refers to the periodic fluctuations in the time series (usually related to seasonal, monthly, and other periodic time characteristics); *h*(*t*) represents the holiday effect, which refers to the impact of holidays or special events on the forecast results in the time series; ε_t represents noise, which refers to the random fluctuations or disturbances in the time series that cannot be explained by the above components.

In practical applications, time series data may exhibit significant trend changes, such as a sudden increase or decrease in criminal behavior at specific time points. In the Prophet model, these time points where significant trend changes occur are referred to as "change points. Bayesian inference is utilized to automatically detect these change points in the data, allowing for adjustments to the trend. Furthermore, the Prophet model employs Bayesian inference to estimate model parameters, including trend components, seasonal components, and holiday components. For instance, it uses variational inference methods to sample the posterior distribution of these parameters. The incorporation of Bayesian methods enhances the Prophet model's capacity to quantify uncertainty, thereby making time series forecasting results more robust.

2.2 Kernel Density Estimation

The kernel density estimation (KDE) function estimates the probability density surrounding each data point through a smoothing process, generating smooth and continuous density values for each point. This quantifies the probability density distribution of the data in two-dimensional space, which can be intuitively visualized in the form of a heatmap. For theft crime data, the spatial information can be represented as a tuple (x_i, y_i), where x_i and y_i correspond to latitude and longitude, respectively, and *i* represents any data point in the dataset. In this case, the estimation function of the kernel density estimation can be expressed as:

$$\hat{f}(x,y) = \frac{1}{nh_x h_y}\sum_{i=1}^{n}\left(\frac{x-x_i}{h_x},\frac{y-y_i}{h_y}\right) \tag{2}$$

where, *n* is the sample size, h_x and h_y are the bandwidth parameters that control the smoothness in the *x* and *y* directions, and *K* represents the two-dimensional Gaussian kernel function:

$$K(u,v) = \frac{1}{2\pi}e^{-\frac{u^2+v^2}{2}} \tag{3}$$

where, $u = \dfrac{x-x_i}{h_x}$ and $v = \dfrac{y-y_i}{h_y}$ represent the influence of data points (x_i, y_i) on the target point (*x*, *y*).

Generate a density estimation map using the kernel density estimation function, where the density value of each point is represented by the intensity of color in a heatmap. The depth of color indicates the magnitude of the density value. Areas with a high concentration of data points and elevated density typically appear as darker regions (hotspot areas), while areas that are sparse or devoid of data generally appear as lighter regions (low-risk areas).

3 EXPERIMENTS AND RESULTS

In the relevant tasks, the open-source Chicago Crime dataset was utilized to predict the frequency of theft crimes over time and to conduct spatial hotspot analysis. For the time prediction task, theft crime data from 2012 to 2015 was employed to train the Prophet model, which achieved daily and monthly predictions for theft crimes in 2016. The experimental results of the time series prediction are presented in Figure 1. In the spatial hotspot analysis task, theft crime data from the first seven days of January through June 2016 was used to train a kernel density estimation function, resulting in a visualization of theft crime spatial hotspots. The experimental results of the spatial hotspot analysis are displayed in Figure 2.

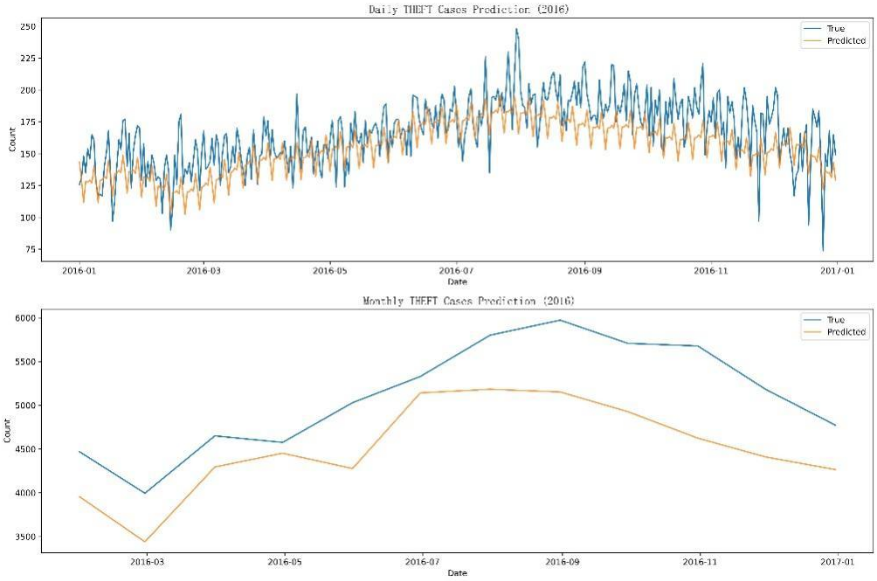
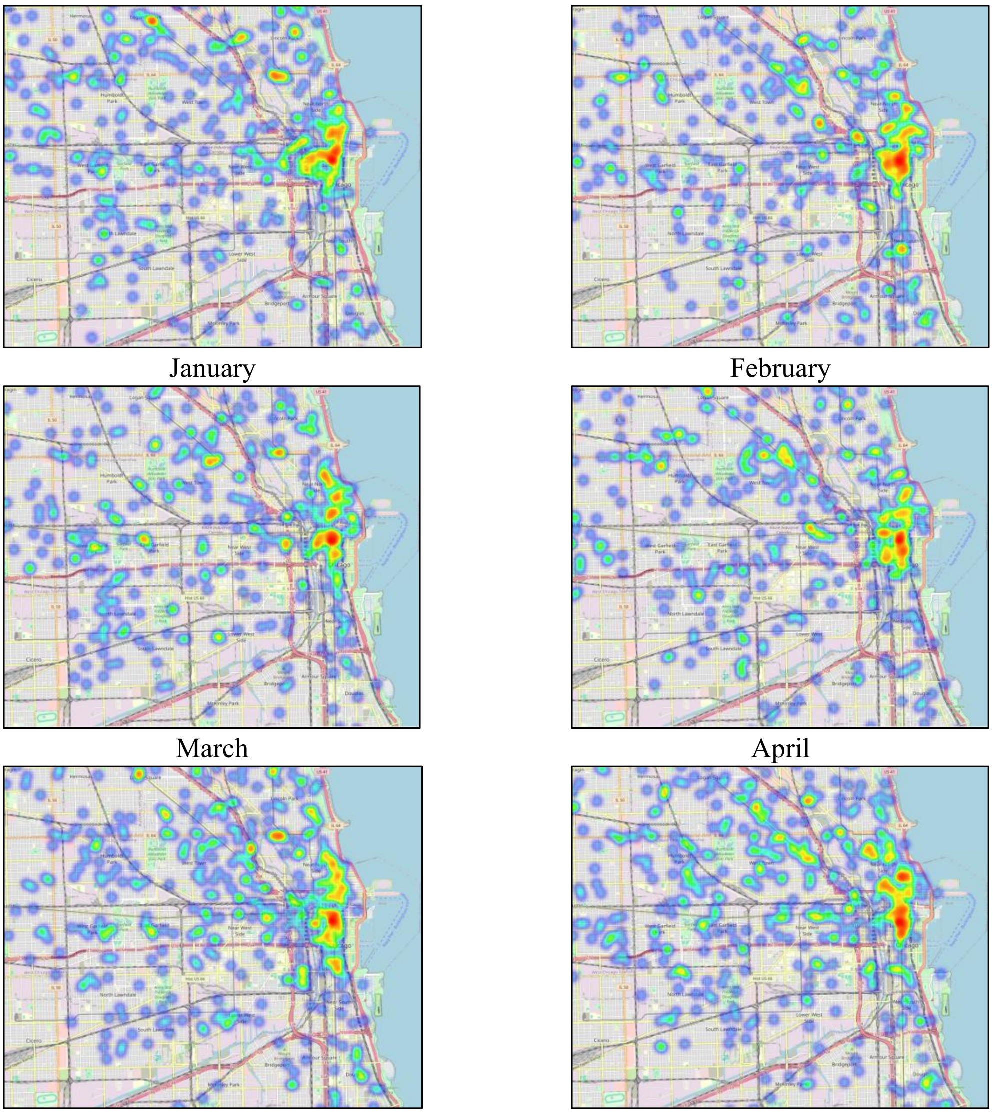


Figure 1 The Experimental Results of Theft Crime Time Series Prediction



January

February

March

April

May

June

Figure 2 Visualization of Hotspots for Theft Crimes (2016)

From a daily trend perspective, theft crimes exhibit periodic fluctuations throughout the week, with peaks generally occurring on weekends and lower numbers reported on Mondays and Tuesdays. From a monthly trend perspective, theft crimes demonstrate significant seasonal variations over the course of the year, with peak periods concentrated from August to November, followed by a notable decline from December to January of the subsequent year. In terms of spatial hotspots, theft crimes are typically concentrated in commercial districts and densely populated residential areas; for instance, bustling shopping areas often experience high foot traffic, making them attractive targets for theft. By employing the Prophet model to analyze the monthly and daily cycles of theft crimes, and integrating it with kernel density estimation to create a crime heatmap, a "time-space" two-dimensional situational awareness framework is established. Theoretically, time series forecasting validates the temporal fluctuation characteristics of crime opportunities as outlined in routine activity theory, revealing the potential influence of time variables on criminal behavior. Additionally, spatial density distribution supports the stability hypothesis of crime hotspots in criminology, identifying the fields that attract crime in high-risk areas.

4 CONCLUSION

Using KDE for theft crime prediction combines the advantages of time series forecasting with spatial data analysis. The Prophet model analyzes historical crime data to accurately forecast future crime trends, enabling law enforcement agencies to allocate resources effectively and implement preventive measures proactively. Meanwhile, the kernel density function leverages spatial data to identify areas with high crime concentrations, highlighting regions at elevated risk for criminal activity. The integration of these two methodologies not only provides insights into the temporal patterns of crime occurrence but also precisely identifies spatial distributions, thereby offering a foundation for developing more targeted crime prevention strategies. This approach demonstrates high predictive accuracy, enhances the efficiency of public safety management, reduces crime rates, and possesses significant social relevance and practical value.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported by the Social Science Fund of Guangxi under Grant 23FTQ005

REFERENCES

[1] Chainey S, Tompson L, Uhlig S. The utility of hotspot mapping for predicting spatial patterns of crime. Security journal, 2008, 21(1): 4-28.

[2] Travaini G V, Pacchioni F, Bellumore S, et al. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. International journal of environmental research and public health, 2022, 19(17): 10594.

[3] Jenga K, Catal C, Kar G. Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(3): 2887-2913.

[4] Johnson S D, Summers L, Pease K. Offender as forager? A direct test of the boost account of victimization. Journal of Quantitative Criminology, 2009, 25: 181-200.

[5] Mondal S, Singh D, Kumar R. Crime hotspot detection using statistical and geospatial methods: a case study of Pune City, Maharashtra, India. GeoJournal, 2022, 87(6): 5287-5303.

[6] Ratcliffe J. Crime mapping: Spatial and temporal challenges. Handbook of quantitative criminology, 2010: 5-24.

[7] Predictive policing and artificial intelligence. Routledge, Taylor & Francis Group, 2021.

[8] Craglia M, Haining R, Wiles P. A comparative evaluation of approaches to urban crime pattern analysis. Urban Studies, 2000, 37(4): 711-729.

[9] Mityagin S A, Ivanov S V, Boukhanovsky A V. Multi-factorial predictive modelling of drug addiction for large urban areas. 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT). IEEE, 2014: 1-7.

[10] Lin H. Unraveling Crime Dynamics: A Spatiotemporal Analysis of Crime Hotspots. Kent State University, 2024.

[11] Taylor S J, Letham B. Forecasting at scale. The American Statistician, 2018, 72(1): 37-45.

[12] Chen Y C. A tutorial on kernel density estimation and recent advances. Biostatistics & Epidemiology, 2017, 1(1): 161-187.

Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3040

AN EFFICIENT CROSS-DOMAIN AUTHENTICATION SCHEME FOR INTERNET OF VEHICLES (IOV) BASED ON REPUTATION

YuChen Shan
Guangzhou University, Cyberspace Institute Advanced Technology, Guangzhou 510006, Guangdong, China.
Corresponding Author: YuChen Shan, Email: shanyuchen0707@163.com

Abstract: With the rapid development of intelligent transportation technologies, the Internet of Vehicles (IoV) has become an essential component of modern transportation systems. However, as the scale of IoV continues to expand, achieving efficient and trustworthy cross-domain authentication has emerged as a critical technical challenge. To address this issue, this paper proposes a blockchain-based reputation framework for IoV, focusing on cross-domain authentication. Specifically, we propose an efficient cross-domain authentication scheme based on short signatures to address the efficiency and security issues in cross-domain authentication. Traditional cross-domain authentication schemes often incur high computational overhead and network bandwidth consumption, especially when there are significant differences in reputation values across different regions, making identity authentication inefficient. To this end, we design a cross-domain authentication algorithm based on reputation value coupling, which dynamically adjusts the reputation values of vehicle nodes according to the reputation weights of different regions. By integrating smart contract technology, the transparency and traceability of the authentication process are ensured. The dual-authentication mechanism based on reputation values further enhances the security and trustworthiness of nodes. Experimental results demonstrate that the proposed authentication scheme not only significantly improves authentication efficiency but also enhances the security and reliability of the IoV system.
Keywords: Blockchain; Reputation; Cross-domain; IoV

1 INTRODUCTION

With the rapid development of information technology, the IoV has become a core technology of the next-generation intelligent transportation system and is being widely applied globally. IoV enables information sharing and collaborative decision-making among vehicles through Vehicle-to-Vehicle (V2V), Vehicle-to-Roadside (V2R), Vehicle -to-Pedestrian (V2P), and Vehicle-to-Internet (V2I) communications[1]. This enhances traffic efficiency, reduces traffic accidents, and improves the driving experience. However, the rapid growth of IoV also brings significant security and privacy challenges, especially in cross-domain authentication and trust management[2].

In traditional IoV architectures, vehicle authentication and information exchange often rely on centralized Certificate Authorities (CAs). While this approach provides a certain level of security, it also suffers from several critical issues, such as single-point failure, low efficiency in cross-domain collaboration, and the formation of trust silos. In cross-domain scenarios, vehicle communication between different regions or service providers requires complex cross-domain authentication processes, which increase system overhead and authentication latency, thereby affecting the real-time requirements of IoV. Moreover, the lack of a vehicle behavior credibility assessment mechanism makes IoV systems vulnerable to Sybil attacks (identity forgery) and man-in-the-middle attacks, threatening the overall network security[3].

To address these challenges, blockchain technology has emerged as a promising solution for IoV security due to its decentralized, tamper-proof, and transparent characteristics[4]. Blockchain can effectively overcome the limitations of traditional centralized authentication mechanisms by recording vehicle identities and behavior information in a distributed ledger, thereby establishing a trustworthy cross-domain authentication system. However, the application of blockchain in IoV also faces new challenges, such as high storage overhead, low throughput, and communication efficiency issues caused by long signatures. Therefore, designing an efficient cross-domain authentication scheme that ensures security while maintaining high efficiency and scalability is a crucial research topic in the IoV field[5].

Short signature technology, as an efficient cryptographic tool, can significantly reduce signature length and computational overhead while ensuring security. Integrating short signature technology with blockchain can further optimize the efficiency of cross-domain authentication, enhancing the real-time and scalable nature of IoV[6]. Based on this, this paper proposes an efficient cross-domain authentication scheme for IoV based on blockchain and short signatures, aiming to address the efficiency and security issues in cross-domain authentication and build a trustworthy, efficient, and scalable IoV ecosystem[7].

In IoV systems, cross-domain authentication is a key link for vehicle information sharing and collaborative decision-making. However, existing cross-domain authentication schemes have several significant shortcomings:

(1) Single-point failure in centralized authentication: Traditional IoV authentication relies heavily on centralized CAs. If a CA is attacked or fails, the entire authentication system may collapse, threatening the security and reliability of IoV.

(2) Low efficiency in cross-domain collaboration: Communication between vehicles in different domains requires complex cross-domain authentication processes, leading to increased authentication latency and difficulty in meeting the real-time requirements of IoV.

(3) Trust silo problem: There is a lack of unified trust assessment mechanisms between domains, making it difficult to quantify vehicle behavior credibility and leaving the system vulnerable to malicious node attacks.

(4) Trade-off between signature efficiency and security: Traditional digital signature algorithms (e.g., RSA, ECDSA) are secure but have long signature lengths and high computational overhead, making them unsuitable for high-concurrency, low-latency IoV scenarios.

To address the aforementioned challenges, blockchain technology offers a decentralized solution for cross-domain authentication in IoV. However, blockchain itself faces several limitations, such as high storage overhead, low throughput, and communication inefficiencies caused by long signatures. Therefore, integrating blockchain with short signature technology to design an efficient and secure cross-domain authentication scheme has become an urgent issue.

In this paper, we propose an efficient cross-domain authentication scheme for IoV based on blockchain and short signatures, leveraging the decentralized nature of blockchain, the efficiency of short signatures, and the trustworthiness of reputation mechanisms to build a reliable, efficient, and scalable cross-domain authentication framework for IoV. Specifically, the main contributions of this paper are as follows:

(1) Partitioned Blockchain Architecture: We divide the IoV into different regions and propose a partitioned blockchain architecture to efficiently manage the complex IoV network. By recording vehicle identities and behavior information in a distributed ledger, we eliminate the single-point failure issue associated with traditional centralized authentication.

(2) Short Signature Algorithm: We employ an efficient short signature algorithm to significantly reduce the signature length and computational overhead while ensuring security. This enhances the efficiency of cross-domain authentication and enables rapid authentication in high-concurrency scenarios, meeting the real-time requirements of IoV. Additionally, we introduce a reputation-based dual-authentication mechanism to ensure the security of cross-domain nodes.

(3) Smart Contract Automation: We incorporate smart contract technology to automate the management of interactions between vehicle nodes, such as message passing and information updates. Suitable API interfaces are embedded in the onboard units to quickly update information to the blockchain system.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the preliminaries. Section 4 details the design of the cross-domain authentication scheme. Section 5 presents the simulation experiments. Section 6 concludes the paper.

2 RELATED WORK

In the field of cross-domain authentication for the Internet of Vehicles (IoV), researchers have proposed a variety of solutions tailored to different application scenarios. For example, Zhang et al.[8] proposed a distributed and scalable cross-domain vehicle authentication framework that optimizes the authentication process to reduce system overhead. Experimental data showed that this solution reduced computational resource consumption and communication latency. Wang et al.[9] designed a decentralized authentication architecture from the perspective of edge computing, innovatively incorporating a dynamic management mechanism for centralized certificate revocation tables. Simulation results indicated that compared to traditional solutions, their approach significantly reduced verification latency, making it particularly suitable for high-density vehicle environments.

In the context of cloud-based traffic monitoring, Wang et al.[10] studied a scenario where cloud servers and authoritative institutions should be able to verify the source of reports, i.e., checking whether traffic conditions are reported by legitimate vehicles. They theoretically analyzed the efficiency of their approach and experimentally demonstrated its practicality.

Regarding privacy protection, Zhang et al.[11] constructed a conditional privacy-preserving authentication model based on the Chinese Remainder Theorem, mathematically verifying the scheme's advantages in resisting replay and impersonation attacks. He et al.[12] innovatively used non-bilinear pairing encryption algorithms, proposing an authentication protocol that significantly shortened signature generation time while ensuring data integrity, making it suitable for security-related applications in IoV. Cui et al.[13] developed a one-time registration authentication system that uses pre-configured trusted institution certificate chains to ensure that vehicle cross-domain authentication response times meet the real-time requirements of IoV.

To address complex network threats, Xu et al.[14] proposed a cross-domain group authentication scheme that effectively resolves security issues. Sun et al.[15] designed a dual-layer authentication system capable of defending against DoS attacks, increasing authentication throughput while ensuring privacy by differentiating between intra-domain and cross-domain authentication processes. Meng et al.[16] integrated blockchain technology to build an authentication framework that automatically negotiates keys through smart contracts, demonstrating superior lightweight characteristics and efficiency. Zhang et al.[17] proposed a dual-blockchain-assisted conditional privacy-preserving authentication framework and protocol for IoV. Zhu et al. [18] introduced a certificateless signature scheme that pre-generates cross-domain communication credentials, significantly improving cross-domain authentication efficiency between vehicles, with simulation experiments verifying the protocol's security. Zhong et al.[19] developed a batch authentication mechanism that uses pre-computation techniques to reduce the cost of concurrent multi-vehicle authentication. Tan et al. [20] proposed a dynamic authentication protocol that leverages RSU cluster collaboration to achieve high authentication efficiency even at high vehicle speeds.

These studies collectively demonstrate the ongoing efforts to enhance the security, efficiency, and scalability of cross-domain authentication in IoV. However, challenges such as high computational overhead, communication inefficiencies, and privacy concerns still need to be addressed.

3 PREPARATIONS

3.1 Short Signature Technology

The Boneh-Lynn-Shacham (BLS) short signature scheme is a cryptographic signature mechanism based on elliptic curve cryptography and bilinear pairings. It is characterized by its short signature length, high computational efficiency, and strong security. In the context of low-bandwidth communication in IoV, BLS can significantly reduce the authentication overhead. During cross-domain authentication, vehicle nodes can perform concurrent authentication and aggregate signatures from multiple vehicles, thereby reducing the communication overhead associated with cross-domain interactions.

From a security perspective, BLS signatures are based on the mathematical hardness of bilinear pairings, ensuring the anonymity of vehicle requests. External attackers cannot infer the true identity of vehicles from the signatures. The main construction of BLS is as follows:

$$BB = \left(q, G_1, G_2, G_T, e, P_1, P_2, H(\cdot)\right) \tag{1}$$

In the IoV, the security level of the system is determined by the order q of the elliptic curve. The cyclic groups G_1, G_2 and G_T on the elliptic curve serve as the foundation for implementing bilinear pairings and signature verification. Specifically: G_1 is used for generating public keys, creating signatures, and verifying signatures. It forms the fundamental space for vehicle authentication. G_2 is used for hashing messages and generating/verifying cross-domain credentials. It serves as the basis for message authentication and cross-domain authentication. G_T is the target group of the bilinear pairing. The bilinear pairing function e: $G_1 \times G_2 \rightarrow G_T$ is employed to validate the authenticity of BLS signatures. $P_1 \in G_1$ and $P_2 \in G_2$ are base points on the elliptic curve, used for generating public keys and signatures. The hash function H: $\{0,1\}^* \rightarrow G_2$ maps messages to points in the elliptic curve group G_2, enabling the signing and verification processes.

3.2 Smart Contracts

Smart contract is a self-executing program that operates autonomously once deployed on a blockchain. It enforces predefined rules and executes operations without the need for human intervention. This automation significantly reduces the potential for manual errors and enhances the efficiency and reliability of the system.

Smart contracts are particularly well-suited for applications requiring high levels of transparency and immutability. The inherent properties of blockchain technology ensure that once a smart contract is deployed, its rules and executed operations cannot be altered or tampered with. This immutability guarantees data integrity and transparency, making every transaction publicly verifiable and auditable. The execution of smart contracts is typically governed by conditional logic encoded in the blockchain, often represented as "if...then..." statements. Upon fulfilling specified conditions, the blockchain updates accordingly, reflecting the executed operations.

In the context of cross-domain authentication within the IoV, the integration of smart contracts offers several key advantages. First, it streamlines the authentication process by automating tasks such as vehicle public key registration, reputation value adjustment, and cross-domain credential verification. This automation eliminates the need for intermediaries, thereby reducing operational costs and minimizing delays. Second, the complex logic required for cross-domain authentication can be efficiently implemented and enforced through smart contracts, enhancing both the efficiency and flexibility of the system. By leveraging these capabilities, smart contracts provide a robust and transparent solution for secure and efficient cross-domain authentication in IoV applications.

3.3 System Model

As shown in Figure 1, the IoV based on blockchain technology involves multiple domains, each comprising five main entities: Trust Authority (TA), Vehicle Nodes, Roadside Units (RSUs), Upper-layer Blockchain, and Lower-layer Blockchain. The communication process primarily includes intra-domain authentication and cross-domain authentication. Intra-domain authentication ensures the integrity and trustworthiness of messages within the same domain, while cross-domain authentication guarantees the security and trustworthiness of messages exchanged between different domains. The main functions of each entity are as follows:

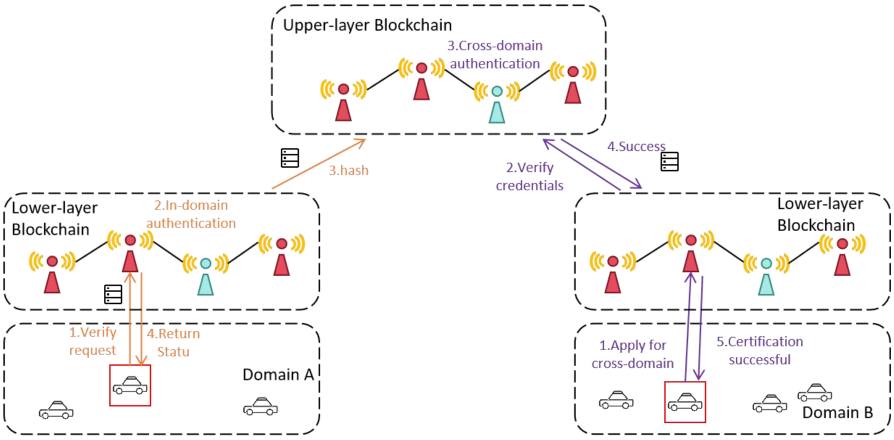


Figure 1 Cross-Domain Authentication Framework Diagram of the Internet of Vehicles

TA: Each domain is managed by a TA, which is considered a trusted institution by default. The primary responsibilities of the TA include managing members within its domain, such as registering nodes, generating public-private key pairs, and tracking member behavior. Additionally, TAs from different domains collectively maintain an upper-layer blockchain, which records their activities and forms a distributed ledger to ensure the overall security of the IoV.

Vehicle Nodes: Vehicle nodes are the fundamental communication units in the IoV, capable of communicating with other vehicle nodes, RSUs, and TAs. During V2V and V2R communications, messages sent by vehicles need to be authenticated by RSUs or TAs to ensure their integrity and trustworthiness. Specifically, if both the sender and receiver belong to the same domain, the message is verified by the local RSU and TA. Otherwise, cross-domain authentication is performed by the source TA and the target TA. The reputation of vehicle nodes is derived from the records of their behaviors, which will be detailed in the following section.

RSUs: RSUs are roadside infrastructure units responsible for collecting local IoV information. Given the vastness of the IoV, multiple RSUs may exist within a single domain to reduce communication latency. RSUs act as intermediaries between different participants, facilitating tasks such as signature verification, message forwarding, invoking smart contract APIs, and contacting TAs. As semi-trusted entities, RSUs do not deviate from predefined protocols but may attempt to access private information such as identities. Their temporary reputation is derived from the average reputation of all nodes within the domain in the previous consensus phase.

Upper-layer Blockchain: The upper-layer blockchain is a distributed database collectively maintained by TAs from different domains. It stores all relevant information of the IoV, including cross-domain authentication records, vehicle public-private key pairs, and other essential data.

Lower-layer Blockchain: The lower-layer blockchain is a distributed database maintained by all RSUs within the same domain. It records vehicle behavior information and serves as an intermediary node between vehicle nodes and the upper-layer blockchain, facilitating communication and data synchronization.

4 DETAILED DESIGN OF CROSS-DOMAIN AUTHENTICATION

This section provides a detailed introduction to the cross-domain authentication scheme based on reputation. The scheme consists of five main steps: Initialization, Registration, Intra-Domain Authentication, Cross-Domain Authentication, and Identity Traceability.

4.1 Initialization Phase

The initialization phase is the foundational setup of the system, involving the generation of system parameters, configuration of regional reputation weights, and deployment of smart contracts. The pseudocode is shown in Table 1.

Table 1 Initialization Algorithm

```
Algorithm 1: Initialization
Input: BB
Output: System Initialized
Get BB
e: $G_1 \times G_2 \rightarrow G_T$
H: $\{0,1\}^* \rightarrow G_2$
for each domains do
    Initialize TA(domain) according to(1)
    for each RSU in domain do
        Initialize RSU(domain) according to(2)
end for
```

```
    Get RRW according to(3)
    Initialize downblockchain()
end for
Initialize upblockchain()
return "System Initialized"
```

4.1.1 System initialization

Firstly, the IoV generates the basic domain parameters and values based on the bilinear pairing e and the elliptic curve BB. The vehicle authentication space G_1 and the message authentication space G_2 satisfy the bilinear pairing e, and the hash function H: $\{0,1\}^* \rightarrow G_2$ maps relevant authentication messages to G_2. Subsequently, the TA is initialized as the manager of the upper-layer blockchain. Using Equation (1), the TA generates its own public key PK_{TA}. The TA then initializes the RSUs in each region, generating their respective public keys PK_{RSU} using (2). Finally, the upper-layer blockchain deploys its own TA node, while the lower-layer blockchain is maintained by the RSU nodes in each designated region.

$GTA()$: Generates the public key PK_{TA} and private key SK_{TA} for the TA. The private key SK_{TA} is a random large prime number, and the corresponding public key PK_{TA} is derived from the generator g_2 in group G_2:

$$PK_{TA} = SK_{TA} \cdot g_2 \tag{1}$$

$GRSU()$: Generates the public key PK_{RSU} and private key SK_{RSU} for each RSU. The private key SK_{RSU} is a random large prime number, and the corresponding public key PK_{RSU} is derived from the generator g_2 in group G_2:

$$PK_{RSU} = SK_{RSU} \cdot g_2 \tag{2}$$

4.1.2 Regional reputation weight configuration

The Regional Reputation Weight (RRW) is a trust metric at the regional level. It measures the degree of trust a region has in vehicles (or other participants) during the authentication process, thereby determining whether the region is willing to accept a vehicle's authentication request and the extent to which the vehicle's reputation value is adjusted during authentication. The RRW facilitates trust propagation in cross-domain authentication systems. For example, when a vehicle moves from Region A (source region) to Region B (target region), the reputation value of the source region is adjusted in the target region based on the RRW. The RRW of the target region directly affects the vehicle's entry authentication in that region. During cross-domain authentication, a vehicle's reputation value is dynamically adjusted according to the RRW. This ensures fairness and rationality in trust propagation between regions. For instance, if a vehicle has a high reputation in the source region but enters a target region with a low RRW, its reputation value may decrease. Conversely, if the target region has a high RRW, the vehicle's reputation value may increase. These adjustments ensure that trust is fairly and reasonably propagated across regions.

The RRW is calculated based on a combination of multiple factors, including: The average reputation value within the region $\overline{R_{RSU}}$; Historical authentication data HH; The level of IoV activity within the region AA. The formula for calculating the RRW is defined as follows:

$$RRW = \omega_1 \cdot \overline{R_{RSU}} + \omega_2 \cdot HH + \omega_3 \cdot AA + \omega_4 \cdot \sigma \tag{3}$$

Where,

$$HH = \frac{H1}{H2} \tag{4}$$

$$AA = log_2 n \tag{5}$$

$$\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1 \tag{6}$$

$\overline{R_{RSU}}$ is the average reputation value of all RSUs within the region during a consensus phase. HH is the proportion of successful authentications in the region, representing the region's historical trust in vehicles. $H1$ is the number of successful authentications and $H2$ is the total number of authentication attempts. AA represents the level of IoV activity within the region, including communication frequency between vehicles and RSUs. A higher level of activity enhances the efficiency and accuracy of the authentication process. n denotes the number of active vehicles within the region. σ is a noise factor introduced to add randomness and enhance security. $\omega_1, \omega_2, \omega_3, \omega_4$ are weight coefficients for each factor, determining their respective impact on the RRW.

Regarding the visibility analysis of RRW, vehicles can only be aware of the weight of the current area, in order to prevent potential manipulation of the weight of the target area. The source area and the target area TA can know their respective RRW through application(Table 2). Below is a summary of the visibility of RRW.

Table 2 RRW Visibility Table

Participants	Source RRW	Target RRW
Source Vehicle	Visible	Invisible
Source RSU	Visible	Invisible
Source TA	Visible	Visible (Application)
Target TA	Visible (Application)	Visible
External Attacker	Invisible	Invisible

4.1.3 Smart contract deployment

In the initialization phase, vehicle public-private key pairs must be registered on the blockchain to facilitate subsequent signature verification. Smart contracts are deployed to automate and verify this process, which includes the following steps:

(1) Key Pair Generation and Registration

Key Pair Generation: Each vehicle generates a public-private key pair. The private key is used for signing messages, while the public key is used for verification.

Public Key Registration: Vehicles submit their public keys along with other identity information (e.g., vehicle ID) to the smart contract.

Public Key Storage: The smart contract verifies and stores the vehicle's public key on the blockchain, ensuring that it can be validated by other participants during future authentication processes.

(2) Management of Regional Reputation Weights

Regional Reputation Weights (RRWs) are crucial factors influencing vehicle reputation values in cross-domain authentication. The smart contract manages RRWs through the following functionalities:

Storage of Regional Reputation Weights: The smart contract stores the RRWs of each region.

Dynamic Update of Regional Weights: The smart contract updates RRWs dynamically based on predefined criteria and historical data.

Calculation of Vehicle Reputation Values: The smart contract calculates the vehicle's reputation value based on the RRWs of the source and target regions.

(3) Cross-Domain Authentication Process

During the cross-domain authentication phase, vehicles initiate authentication requests through the source region's Trust Authority (TA). The smart contract plays a vital role in this process:

Receiving Authentication Requests: Vehicles submit authentication requests to the smart contract via the source region's TA. The request includes the vehicle ID, source region's reputation value, target region ID, and a timestamp.

Signature Verification: The smart contract verifies the legitimacy of the signature using the vehicle's public key, ensuring the authenticity of the vehicle's identity.

Reputation Value Adjustment: Based on the RRWs of the source and target regions, the smart contract calculates the adjusted reputation value of the vehicle in the target region.

Storing Authentication Results: The authentication result (pass/fail) and the adjusted reputation value are stored on the blockchain via the smart contract.

All authentication processes, verification steps, and authentication results are recorded on the blockchain through the smart contract. This ensures transparency, traceability, and auditability of the entire authentication process. The smart contract automates key management, reputation weight adjustments, and authentication decisions, thereby enhancing the efficiency and security of cross-domain authentication in the Internet of Vehicles (IoV).

4.2 Registration Phase

In the registration phase, vehicles register their identities within the IoV and obtain their public-private key pairs. The pseudocode is shown in Table 3. The Data Flow Diagram is roughly illustrated in Figure 2. The detailed process is as follows:

4.2.1 Vehicle registration request

The vehicle node generates a registration request using its unique identity identifier ID_V and sends the request *GSetup()* to the local RSU. Upon receiving the request, the RSU forwards it to the upper-layer for registration.

4.2.2 Identity verification and key generation

The TA verifies the legitimacy of the vehicle's identity identifier *IDv*. If the verification is successful, the TA generates the vehicle's private key SK_V using Equation (7):

$$SK_V = H(ID_V) \cdot SK_{TA} \tag{7}$$

where *H()* is a cryptographic hash function, ID_V is the vehicle's unique identity identifier, and SK_{TA} is the private key of the TA.

The TA then generates the vehicle's public key PK_V using Equation (8):

$$PK_V = SK_V \cdot g_2 \tag{8}$$

where g_2 is the generator of the group G_2.

4.2.3 Recording and feedback

The TA records the vehicle's public keyPK_V and other relevant information on the upper-layer blockchain. Then, the TA then sends the registration status back to the RSU, which forwards it to the vehicle node.

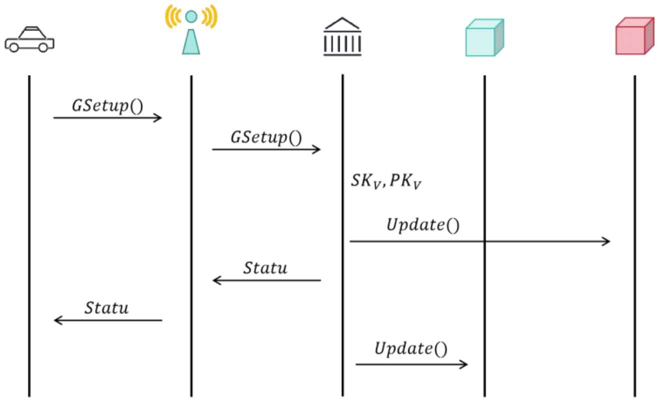


Figure 2 Data Flow Diagram for Registration Phase

Table 3 Registration Phase Algorithm

```
Algorithm 2: Registration algorithm
Input: Vehicles, RSU, TA, blockchain
Output: Registration Statu
for each vehicles do
    Get Vehicle $ID_V$
    Send $GSetup()$ to RSU and TA
    TA get $SK_V$ according to(7)
    TA gat $PK_V$ according to(8)
    Update $upblockchain()$
    Send statu to RSU and Vehicle
    Update $downblockchain()$
    return Registration Statu
end for
```

4.3 Intra-Domain Authentication Phase

The intra-domain authentication phase is a critical step where vehicles must be authenticated within their source region to ensure the legitimacy of their identity, the integrity of the information, and the validity of the authentication request. The pseudocode is shown in Table 4. The Data Flow Diagram is roughly illustrated in Figure 3. The detailed process is as follows:

4.3.1 Authentication request initiation

The vehicle node generates an authentication request and signs the request message before sending it to the local RSU. The authentication message format is:

$$m_V = \{ID_V, R_A, T, H(m_V), starttarget, endtarget\} \tag{9}$$

Where ID_V is the vehicle's unique identifier. R_A is the vehicle's reputation value. T is the timestamp. $H(m_V)$ is the hash of the message. starttarget and endtarget are the sender and receiver targets, respectively.

The vehicle uses its private key SK_V to generate the signature σ_V:

$$\sigma_V = SK_V \cdot H(m_V) \tag{10}$$

4.3.2 Signature verification by RSU

Upon receiving the authentication request, the RSU retrieves the vehicle's public key $PK_V$from the blockchain and verifies the signature using the following bilinear pairing equation:

$$e(\sigma_V, g_2) = ?\, e(H(m_V), PK_V) \tag{11}$$

If the signature verification fails, the RSU rejects the request and notifies the vehicle. If the signature verification succeeds, the RSU forwards the request to the TA.

4.3.3 Further verification by TA

The TA queries the vehicle's information stored on the blockchain to further verify the vehicle's identity, reputation value, and the validity of the authentication request. If the request passes all verifications, the TA returns an authentication success result to the vehicle.

4.3.4 Recording authentication results

All authentication requests and verification results are stored on the blockchain via a smart contract, ensuring the traceability and auditability of the authentication process.

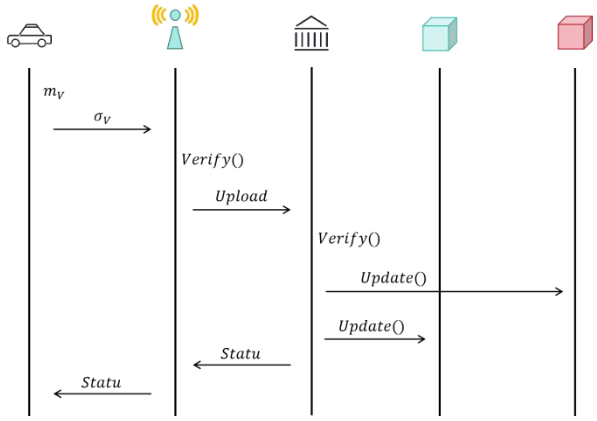


Figure 3 Data Flow Diagram for Intra-Domain Phase

Table 4 In-Domain Authentication Phase Algorithm

```
Algorithm 3: In-domain authentication algorithm
Input: Vehicles, RSU, TA, blockchain
Output: In-domain Statu
for each vehicle do
    Generate $m_V$
    Get $\sigma_V$ according to(10)
    Send to RSU
    if $verify(RSU)$ accoording to(11)
        upload TA
        if $verify(TA)$
            Update $upblockchain()$
        end if
        Update $downblockchain()$
    end if
    Send Statu to vehicle
    return In-domain Statu
end for
```

4.4 Cross-Domain Authentication Phase

The cross-domain authentication phase is initiated when a vehicle moves from one region to another and requests to join the vehicular network of the target region. This phase ensures that the vehicle is authenticated and its reputation is properly adjusted based on the regional reputation weights. The pseudocode is shown in Table 5. The Data Flow Diagram is roughly illustrated in Figure 4. The detailed process is as follows:

4.4.1 Authentication request initiation

The vehicle node in the source region (managed by Source TA) sends an authentication request to the Target TA. The request message format is:

$$m_{TA} = \{ID_V, R_A, RRW_A, T, H(m_{TA}), starttarget, endtarget, \sigma_V\} \tag{12}$$

Where ID_V is the vehicle's unique identifier. R_A is the vehicle's current reputation value. RRW_A is the Regional Reputation Weight of the source region. $T$is the timestamp. $H(m_{TA})$is the hash of the message. starttarget and endtarget are the sender and receiver targets, respectively. σ_V is the signature generated by the vehicle.

The vehicle generates the signature σ_{TA} using its private key SK_V:

$$\sigma_{TA} = SK_V \cdot H(m_{TA}) \tag{13}$$

4.4.2 Signature verification by target TA

The Target TA retrieves the vehicle's public key PK_V from the blockchain and verifies the signature σ_{TA} using the bilinear pairing equation:

$$e(\sigma_{TA}, g_2) = ?\, e(H(m_{TA}), PK_V) \tag{14}$$

If the signature verification fails, the request is rejected, and the vehicle is notified.

4.4.3 Reputation adjustment

The Source TA's Regional Reputation Weight RRW_A and the Target TA's Regional Reputation Weight RRW_B are used to calculate the adjusted reputation value R_B for the vehicle in the target region:

$$R_B = R_A \times \frac{RRW_B}{RRW_A} \tag{15}$$

4.4.4 Automated reputation calculation by vehicle

As the vehicle moves through different regions, it automatically calculates its adjusted reputation value $R_B{}'$ based on the known Regional Reputation Weights $RRW_A{}'$ and RRW_B:

$$R_B{}' = R_A \times \frac{RRW_B{}'}{RRW_A{}'} \tag{16}$$

4.4.5 Malicious behavior check

The system verifies whether the vehicle's calculated reputation value $R_B{}'$ matches the adjusted reputation value R_B computed by the TAs:

$$R_B = ?\, R_B{}' \tag{17}$$

If $R_B \neq R_B{}'$, the vehicle may be flagged for potential tampering or malicious behavior.

4.4.6 Recording authentication results

All information related to the authentication request (e.g., vehicle ID, source region reputation, target region reputation, signature) is stored on the blockchain via a smart contract. This ensures the transparency and traceability of the authentication process.

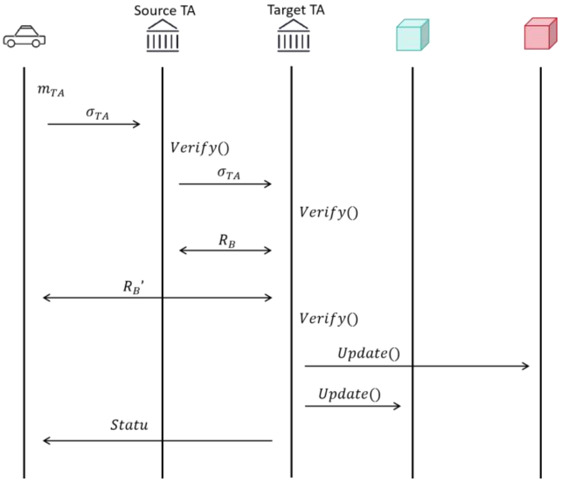


Figure 4 Data Flow Diagram for Cross-Domain Authentication Phase

Table 5 Cross-Domain Authentication Phase Algorithm

```
Algorithm 4: Cross-domain authentication algorithm
Input: Vehicles, Source TA, Target TA, blockchain
Output: Cross-domain Statu
for each vehicle do
    Generate $m_{TA}$
    Get $\sigma_{TA}$ according to(13)
    Send to SourceTA and TargetTA
    if $verify(TargetTA)$ according to(1114)
        TA get $R_B$ according to(15)
        vehilce get $R_B{}'$ according to(16)
        if $verify(Reputation)$ according to(17)
            Update $upblockchain()$
        end if
    end if
    Update $downblockchain()$
    return Cross-domain Statu
end for
```

4.5 Identity Traceability Phase

The primary objective of the identity traceability phase is to track the historical authentication records of vehicles during the cross-domain authentication process. This ensures the traceability and transparency of the authentication process. Detailed information from each authentication request, such as vehicle ID, reputation value, source region, target region, signature, and authentication status, is stored on the blockchain. This ensures the transparency of the authentication process. Additionally, multi-signature verification is employed to ensure that each step in the authentication decision-making process (e.g., source region TA, target region TA) is trustworthy. The signature records from each step are stored on the blockchain via a smart contract, allowing for the complete traceability of each authentication process. The pseudocode is shown in Table 6.

Table 6 Identity Traceability Algorithm

```
Algorithm 5: Identity traceability algorithm
Input: Vehicles, blockchain
Output: Identity history
for each blockchain do
Get $tranceID$
 if $Decrypt(SK_{TA}, EncrytedID) == tranceID$:
```

```
Add to $history$()
end if
if $history$
return $history$()
else
return "No Certification History Found"
end for
```

4.6 Security Analysis

4.6.1Signature Security

In the cross-domain authentication mechanism, the security of signatures is of paramount importance. We employ the BLS short signature algorithm, which is based on the computational difficulty of the discrete logarithm problem. This algorithm effectively prevents forgery attacks. Given the complexity of the discrete logarithm problem in large number fields, it is virtually impossible for attackers to forge a valid signature. Each vehicle's authentication request is encrypted using a BLS short signature, and only the private key held by a legitimate vehicle can generate a valid signature. Therefore, even if an attacker intercepts an authentication request, they cannot forge a valid signature for authentication, ensuring the reliability of identities and the accuracy of verification in cross-regional communications.

4.6.2 Reputation Value Security

The use of blockchain technology ensures the immutability of vehicle reputation records and authentication data. The authentication history and reputation values of vehicles in different regions are stored on the blockchain, and once written, these data cannot be modified or deleted. The distributed ledger of the blockchain not only enhances data transparency but also strengthens the system's defense against tampering attacks. For trust verification of various nodes in the Internet of Vehicles (IoV) (such as vehicles and roadside units), blockchain provides a reliable mechanism, ensuring that all vehicle behavior history and reputation calculations are traceable, public, and trustworthy.

Moreover, vehicles in different regions may face varying trust requirements and authentication standards. To ensure fair reputation adjustments during cross-domain authentication, the reputation values are adjusted based on the source region's reputation value and the target region's reputation weight. The source region's reputation value represents the trustworthiness of vehicles within that region, while the target region's reputation weight reflects the region's trust mechanism and its acceptance of incoming vehicles. Through this adjustment formula, the reputation values of vehicles are dynamically adjusted according to the trust requirements of different regions, ensuring fairness and rationality in the cross-domain authentication process.

4.6.3 Replay Attack Prevention

In the IoV environment, preventing replay attacks is crucial to protect the authentication mechanism from misuse. To this end, we introduce timestamps and unique authentication IDs for each authentication request. Timestamps ensure that each authentication request is accepted within a valid time window, and requests that exceed the time limit are rejected, effectively preventing attackers from replaying delayed requests. Additionally, unique authentication IDs ensure the uniqueness of each authentication request. Even if an attacker captures a legitimate request, they cannot reuse it to forge authentication. Through this mechanism, we not only prevent replay attacks but also ensure that each authentication request is unique and timely, which is particularly important for the high-frequency vehicle identity authentication in IoV.

5 EXPERIMENTAL EVALUATION

This chapter aims to demonstrate the practical usability of our protocol. We conducted simulation experiments using Veins (v 5.0), Omnet++ (v5.4.1), and urban traffic simulation (Sumo v1.11.0). The relevant map data is sourced from OpenStreetMap, and smart contracts were deployed using Solidity (v0.8.0). In the simulation experiments, we tested the write and query latency, message authentication latency, accuracy of introducing reputation authentication, and the message loss rate in the vehicular network. Next, we compared our protocol with FEDAS[9] and RCoM[10] through simulation experiments, with each simulation result being the average of 1000 trials. The map data comes from the streets around the school, and after integrating the road dataset using the Sumo tool, it is specifically shown in Figure 5.

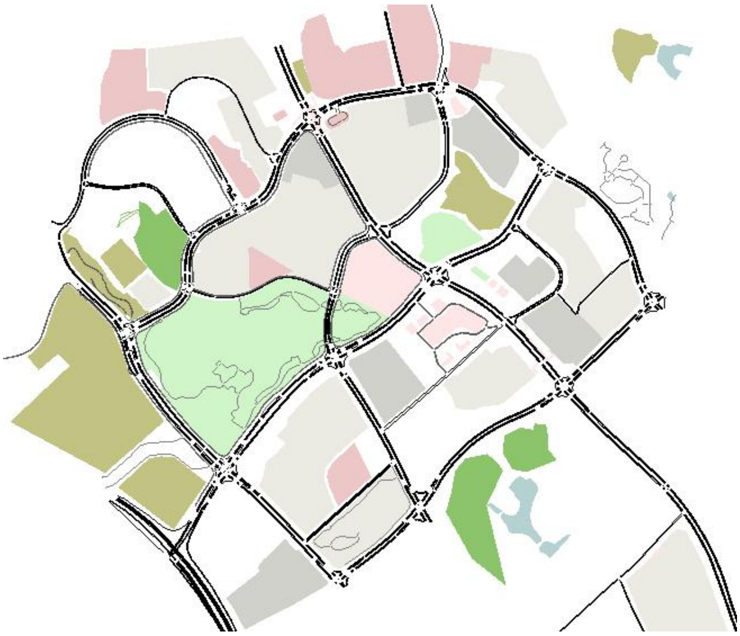


Figure 5 Simulates the Map

5.1 Write and Query Delays

In this protocol, given the introduction of blockchain technology, write delays and query delays have become important indicators for measuring system performance. Write delay refers to the time taken from when a vehicle node sends data until it is successfully written to the blockchain, while query delay refers to the time taken from when the requester sends a query request until the blockchain node returns a response. In this study, we primarily examined the impact of the number of regions (specifically 2, 4, 6, and 8) and vehicle density (ranging from 50 to 400 vehicles per square kilometer) on delays. The experimental results, as shown in Figure 6, indicate that as vehicle density increases, the write and query delays for vehicles within each region gradually increase. This is due to the fact that a higher vehicle density leads to an increase in the number of authentication requests, which in turn increases the load on the network and blockchain nodes, thereby increasing delays. As illustrated in Figure 7, an increase in the number of regions also leads to an increase in write and query delays. The primary reason for this phenomenon is that when vehicles perform cross-domain authentication between different regions, it results in more blockchain interactions, consequently leading to increased write and query delays.

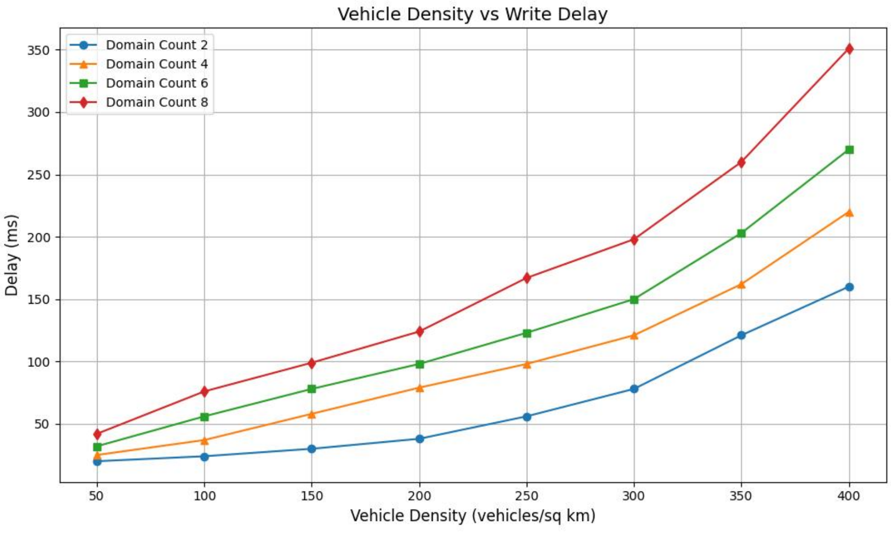


Figure 6 Changes in Vehicle Density and Write Latency

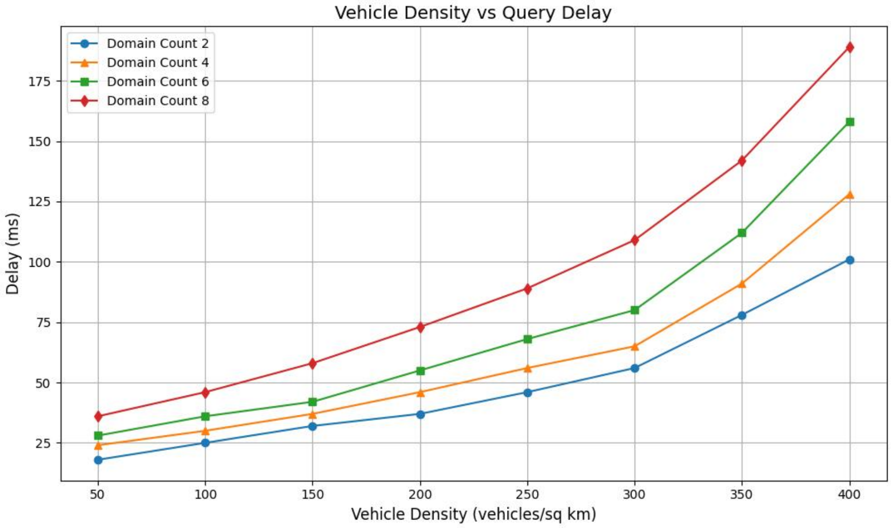


Figure 7 Changes in Vehicle Density and Query Latency

5.2 Delay in Message Authentication

In the Internet of Vehicles, message latency is one of the key indicators to measure system performance and user experience. Specifically, message latency refers to the total time that elapses between when the vehicle sends the certification request and when it receives the certification result. In order to evaluate the efficiency of different authentication protocols, we selected FEDAS[9] and RCoM[10] as the comparison experimental objects, mainly by analyzing the impact of the number of certified vehicles per unit time on the message authentication delay. As shown in Figure 8, the certification delay of the FEDAS and RCoM protocols slowly increases as the number of certified vehicles per unit time increases, until about 40 vehicles begin to increase significantly. This phenomenon may be due to the limited computing resources of centralized authentication, and when the number of authentication requests exceeds the processing capacity, the authentication efficiency of the system decreases rapidly. In contrast, the authentication latency of our solution is always between 200ms and 220ms. This is because we use short signature technology, which has the function of aggregate computation and can process multiple signature verifications in parallel, so as to ensure that the delay of the authentication process remains within a relatively stable range.

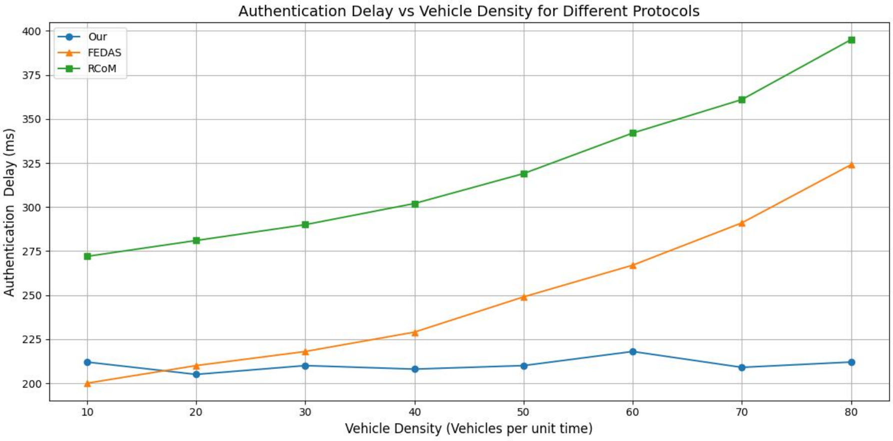


Figure 8 Changes in the Number of Vehicles Per Unit Time and the Certification Delay

5.3 The Introduction of Accuracy in Reputation-Based Authentication

In our cross-domain authentication scheme, reputation-based authentication rules have been introduced. To this end, we conducted comparative experiments between the scheme with a reputation authentication mechanism and the scheme that relies solely on short signatures, evaluating their performance in terms of true positive rate (TPR) and false positive rate (FPR). In the experiments, we assumed that there are 20% malicious vehicles in the network, which include attacks such as external intrusions and identity spoofing. Here, TPR represents the proportion of malicious vehicles correctly identified, while FPR indicates the proportion of normal vehicles misclassified as malicious. Specifically, it is expressed as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{18}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{19}$$

Here, TP represents the number of vehicles correctly identified as malicious, FNrepresents the vehicles that are truly malicious but are incorrectly classified as normal by the system. FP refers to the vehicles that are truly normal but are incorrectly classified as malicious by the system. TN is the number of vehicles correctly classified as normal. The experimental results are shown in Figure 4-9, indicating that the authentication scheme using the reputation mechanism significantly outperforms the cross-domain authentication scheme that only uses short signatures in terms of TPR, while exhibiting a lower FPR. This is because, in the vehicular network environment, malicious vehicles, when performing cross-domain authentication, cannot access the reputation weight information of the source area for malicious vehicles in the target area, leading to errors in the reputation authentication module in certain cases. This error results in a higher TPR and a lower FPR, thereby enhancing the security of cross-domain authentication.

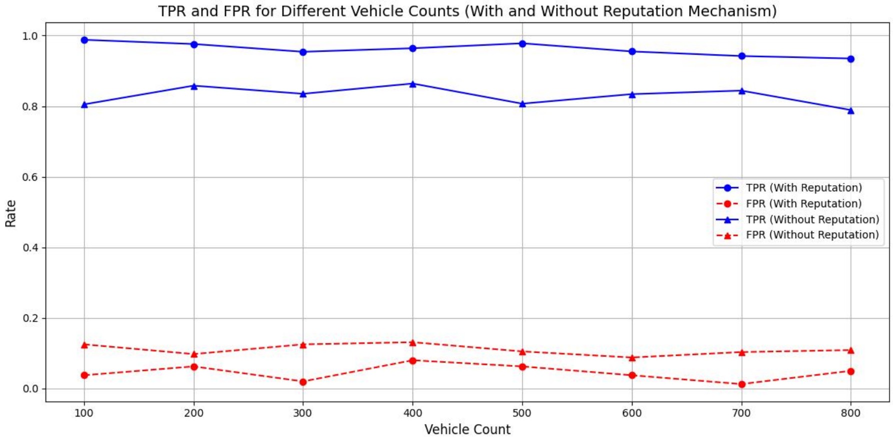


Figure 9 Changes between TPR and FPR

5.4 Packet Loss Rate

To evaluate the packet loss rate during the message transfer phase, we set the simulation experiment time to one hour. FEDAS[9] and RCoM[10] were used as comparison experiments to analyze the packet loss rate of the number of certifications per unit time. The simulation results are shown in 4-10, and we can observe that the influence of vehicle density on the packet loss rate is not significant. The packet loss rate of each protocol is stable within a fixed range, but our protocol has the lowest packet loss rate. The main reason is that each protocol has a fixed signature size, and the short signature technology has the smallest signature size.

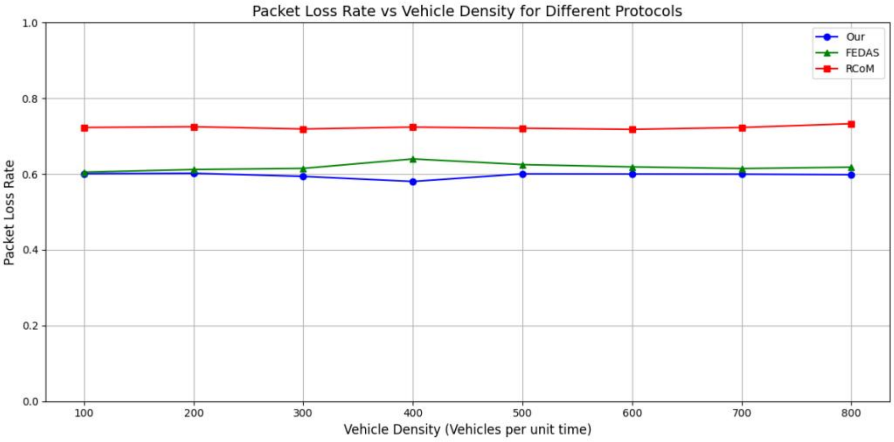


Figure 10 Comparison of Packet Loss Rates

6 CONCLUSION

This chapter discusses the efficiency and security issues faced by cross-domain authentication schemes in the Internet of Vehicles. Firstly, in view of the differences in the reputation evaluation system of different regions in the Internet of Vehicles, a reputation coupling scheme is proposed to ensure the fairness and effectiveness of the Internet of Vehicles. Then, combined with the practical application environment of the Internet of Vehicles, an efficient short signature technology is designed, and it is combined with the reputation mechanism to carry out double authentication, so as to improve the security of the vehicle node. In addition, through the introduction of blockchain and smart contract technology, the automatic execution of cross-domain authentication protocols is ensured, and the transparency and traceability of the authentication process are guaranteed. Finally, the effectiveness of the proposed scheme is verified by simulation experiments, and the experimental results show that the cross-domain authentication scheme of the Internet of Vehicles based on the short signature technology of reputation has improved the security and performance.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCE

[1] Hussain Q, Noor ASM, Qureshi MM. Reinforcement learning based route optimization model to enhance energy efficiency in internet of vehicles. Scientific Reports, 2025, 15: 3113.

[2] Yang Y, Chen Y, Liu Z, et al. Verifiable and redactable blockchain for internet of vehicles data sharing. IEEE Internet of Things Journal, 2025, 12(4): 4249-4261.

[3] Jiang Wenxian, Lv Xianglong, Tao Jun. A secure authentication framework for IoV based on blockchain and ensemble learning. Vehicular Communications, 2024, 50.

[4] Shen X, Ma R. A Blockchain Solution for the Internet of Vehicles with Better Filtering and Adaptive Capabilities. Sensors, 2025, 25(4): 1030.

[5] Ma Z, Jiang J, Wei H, et al. A Blockchain-Based Secure Distributed Authentication Scheme for Internet of Vehicles. IEEE Access, 2024, 12: 81471-81482.

[6] Liu Shuanggen, Zhou Xiayi, Wang Xu An, et al. A hash-based post-quantum ring signature scheme for the Internet of Vehicles. Journal of Systems Architecture, 2025, 160: 103345.

[7] Zhang X, Yang X, Zheng Y, et al. EACAS: An Efficient Anonymous Cross-domain Authentication Scheme in Internet of Vehicles. IEEE Internet of Things Journal, 2025, 160.

[8] Zhang J, Zhong H, Cui J, et al. CVAR: Distributed and Extensible Cross-Region Vehicle Authentication with Reputation for VANETs. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(1): 74-89.

[9] Wang Q, Gao D, Foh C H, et al. An Edge Computing-Enabled Decentralized Authentication Scheme for Vehicular Networks. ICC 2020 - 2020 IEEE International Conference on Communications (ICC). Dublin, Ireland: IEEE, 2020: 1-7.

[10] Wang Y, Ding Y, Wu Q, et al. Privacy-Preserving Cloud-Based Road Condition Monitoring with Source Authentication in VANETs. IEEE Transactions on Information Forensics and Security, 2019, 14(7): 1779-1790.

[11] Zhang J, Cui J, Zhong H, et al. PA-CRT: Chinese Remainder Theorem-Based Conditional Privacy-Preserving Authentication Scheme in Vehicular Ad-Hoc Networks. IEEE Transactions on Dependable and Secure Computing, 2021, 18(2): 722-735.

[12] He D, Zeadally S, Xu B, et al. An Efficient Identity-Based Conditional Privacy-Preserving Authentication Scheme for Vehicular Ad Hoc Networks. IEEE Transactions on Information Forensics and Security, 2015, 10(12): 2681-2691.

[13] Cui J, Zhang X, Zhong H, et al. Extensible Conditional Privacy Protection Authentication Scheme for Secure Vehicular Networks in a Multi-Cloud Environment. IEEE Transactions on Information Forensics and Security, 2020, 15: 1654-1667.

[14] Xu C, Ma M, Huang X, et al. A cross-domain group authentication scheme for LTE-A based vehicular network. In: 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN). Guangzhou, China: IEEE, 2017: 595-599.

[15] Sun C, Liu J, Xu X, et al. A Privacy-Preserving Mutual Authentication Resisting DoS Attacks in VANETs. IEEE Access, 2017, 5: 24012-24022.

[16] Meng X, Xu J, Liang W, et al. A lightweight anonymous cross-regional mutual authentication scheme using blockchain technology for internet of vehicles. Computers and Electrical Engineering, 2021, 95: 107431.

[17] Zhang J, Jiang Y, Cui J, et al. DBCPA: Dual Blockchain-Assisted Conditional Privacy-Preserving Authentication Framework and Protocol for Vehicular Ad Hoc Networks. IEEE Transactions on Mobile Computing, 2024, 23(2): 1127-1141.

[18] Zhu Y, Zhou Y, Wang J, et al.A Lightweight Cross-Domain Direct Identity Authentication Protocol for VANETs. IEEE Internet of Things Journal, 2024, 11(23): 37741-37757.

[19] Zhong Q, Zhao X, Xia Y, et al. CD-BASA: An Efficient Cross-Domain Batch Authentication Scheme Based on Blockchain With Accumulator for VANETs. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(10): 14560-14571.

[20] Tan H, Xuan S, Chung I. HCDA: Efficient Pairing-Free Homographic Key Management for Dynamic Cross-Domain Authentication in VANETs. Symmetry, 2020, 12(6): 1003.