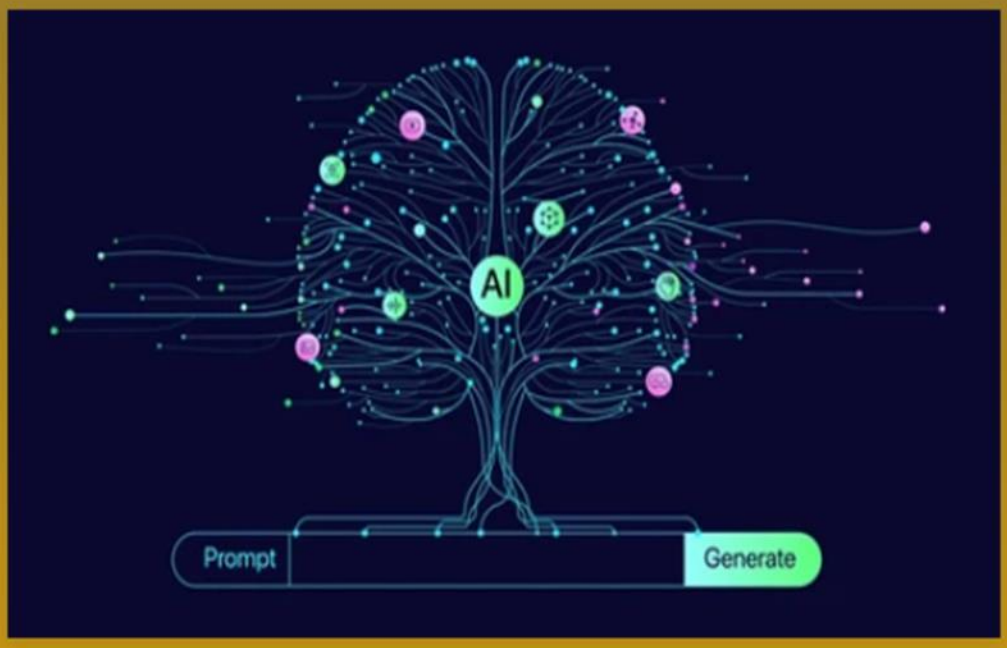
Volume 2, Issue 1, 2025 ISSN: 3078-7394

AI and Data Science Journal



AI and Data Science Journal

Volume 2, Issue 1, 2025

Published by Upubscience Publisher

AI and Data Science Journal
Online ISSN: 3078-7394
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3006

ADAPTIVE PHYSICAL SECURITY: REDEFINING ORGANIZATIONAL SAFETY IN DYNAMIC ENVIRONMENTS

Putri Ayu Nur Rizki*, Sasha Havisa Samosir, Iskandar Muda
Department of Accounting, Universitas Sumatera Utara, Medan, Indonesia.
Correpsponding author: Putri Ayu Nur Rizki, Email: putriayunurrizkiramadhani@gmail.com

Abstract: Organizations today face an unprecedented array of security challenges, driven by rapidly changing environments and increasingly sophisticated threats. Traditional static physical security systems, while effective in addressing known risks, are ill-equipped to handle the complexity and unpredictability of modern organizational landscapes. This paper introduces and explores the concept of adaptive physical security, a dynamic approach that leverages advanced technologies such as artificial intelligence (AI), machine learning, real-time data analytics, and Internet of Things (IoT)-enabled devices to create flexible and resilient security frameworks.

Adaptive physical security systems are designed to respond to evolving threats in real time, enabling organizations to detect, assess, and mitigate risks before they escalate. Central to this approach are predictive threat modeling, scalable access control mechanisms, and automated incident response strategies, all of which work in tandem to ensure comprehensive protection. Through a review of recent advancements, case studies, and practical applications, this study demonstrates the potential of adaptive security systems to significantly enhance organizational resilience.

Key challenges in implementation, such as integration with existing systems, data privacy concerns, and cost implications, are also discussed. Furthermore, the paper highlights the importance of a proactive security culture, emphasizing that technology must be complemented by human vigilance and strategic planning. The findings underscore the need for organizations to shift from reactive to adaptive security paradigms, redefining safety standards to address the demands of today's dynamic environments. This research provides a foundation for future studies and offers actionable insights for organizations aiming to strengthen their physical security measures.

Keywords: Adaptive physical security; Organizational safety; Dynamic environments; Real-time threat detection

1 INTRODUCTION

The concept of security has long been a cornerstone of organizational operations, safeguarding physical assets, personnel, and sensitive information from a variety of threats. Traditionally, physical security systems were designed with a static approach, relying on fixed measures such as locks, surveillance cameras, and human security personnel to deter and respond to risks [1]. While effective in controlled and predictable settings, such systems often fall short in addressing the multifaceted and rapidly evolving threats of today's dynamic environments. The limitations of static security measures have become increasingly evident as organizations navigate a landscape shaped by technological advancements, global interconnectedness, and unpredictable disruptions.

In recent years, the emergence of adaptive physical security has signaled a paradigm shift in how organizations approach safety and risk management. Unlike static systems, adaptive physical security leverages advanced technologies to create a flexible and responsive framework capable of adjusting to real-time changes in the environment. This approach incorporates cutting-edge innovations such as artificial intelligence (AI), machine learning, predictive analytics, and Internet of Things (IoT) devices [2]. Together, these technologies enable organizations to anticipate potential risks, dynamically adjust security measures, and respond to emerging threats with precision and speed.

The need for adaptive physical security has grown more urgent as organizations face increasingly complex challenges [3]. Cyber-physical convergence, for example, has introduced new vulnerabilities where physical and digital systems intersect, requiring an integrated approach to security. Additionally, the rise of global threats such as pandemics, terrorism, and climate-related disasters has underscored the importance of building resilient systems that can adapt to unforeseen circumstances. In this context, adaptive physical security offers a forward-looking solution, providing organizations with the tools to safeguard their operations while maintaining flexibility in uncertain conditions.

This paper seeks to explore the concept of adaptive physical security, examining its key components, benefits, and challenges. By analyzing recent advancements and real-world applications, this study aims to provide a comprehensive understanding of how adaptive systems can redefine organizational safety in dynamic environments. The discussion will focus on critical elements such as predictive threat modeling, scalable access control mechanisms, and automated incident response, highlighting their role in creating a cohesive and proactive security framework.

Furthermore, the paper will address the practical considerations involved in implementing adaptive physical security, including technological integration, cost-effectiveness, and organizational readiness. These factors are crucial for

organizations aiming to transition from traditional static systems to adaptive models. By emphasizing the synergy between technological innovation and strategic planning, this research aims to contribute to the ongoing evolution of security practices, providing actionable insights for organizations seeking to enhance their resilience in the face of modern challenges.

Ultimately, the exploration of adaptive physical security is not merely an academic exercise; it is a response to the pressing need for organizations to protect their assets and personnel in an era of unprecedented change. By adopting adaptive approaches, organizations can move beyond the limitations of static systems, fostering a culture of preparedness and resilience that is essential for success in today's dynamic and unpredictable world.

2 LITERATURE REVIEW

Adaptive physical security has become an essential approach in addressing the challenges faced by modern organizations in dynamic threat environments [4]. The shift from static to adaptive security systems is driven by the increasing complexity and interconnectedness of physical and digital risks. Traditional security methods, such as static surveillance and manual access controls, are often inadequate when dealing with contemporary challenges like cyber-physical threats, which can lead to severe physical damage through digital breaches. For example, cyberattacks on critical infrastructure, such as the Ukraine power grid incident, highlight the vulnerabilities present in conventional systems that lack adaptability.

2.1 The Concept of Adaptive Physical Security

Adaptive physical security is a model that combines advanced technologies like IoT, artificial intelligence (AI), and real-time analytics to dynamically respond to evolving threats. Unlike static systems, adaptive frameworks are characterized by their ability to detect, assess, and mitigate risks as they arise. These systems are increasingly employed in environments where real-time responsiveness and integration across domains are critical, such as in healthcare facilities, corporate offices, and critical infrastructure.

2.2 Components and Functionalities

Several core components define the effectiveness of adaptive security systems: [5]

1. **Predictive Analysis**: By leveraging AI and machine learning, adaptive systems predict potential risks based on patterns and behaviors, providing preemptive measures against emerging threats. This approach significantly reduces the likelihood of successful intrusions .
2. **Integrated Systems**: Adaptive frameworks often merge physical security with IT systems, creating a cohesive network that improves threat detection and enhances response capabilities. For instance, access controls integrated with behavioral analytics provide higher levels of security and flexibility
3. **Real-Time Adjustments**: Unlike traditional systems, adaptive security measures can escalate or de-escalate their responses in real-time, ensuring that security levels match the perceived risk

2.3 Implementation Challenges

Despite the advantages, adopting adaptive security frameworks involves several challenges:

- **High Costs**: Implementing advanced technologies, such as AI-driven analytics and IoT-enabled systems, demands significant financial investment and technical expertise.
- **Complexity and Training**: Integrating adaptive systems into existing infrastructure requires skilled personnel and comprehensive training to manage their functionality effectively.
- **Privacy Concerns**: The use of pervasive monitoring and data analytics in adaptive systems raises ethical questions regarding user privacy and data protection

2.4 Research Gaps

While existing studies emphasize the technological and operational aspects of adaptive physical security, there is limited exploration of its organizational adoption and user perspectives. Qualitative studies could provide insights into the cultural, managerial, and human factors that influence the successful implementation of these systems.

2.5 Contribution to the Field

This literature review highlights the transformative potential of adaptive physical security in creating resilient organizations. By blending traditional methods with advanced technological capabilities, adaptive systems provide a proactive approach to mitigating threats. The findings underscore the need for further research into organizational strategies and best practices for integrating adaptive frameworks.

3 METHODOLOGY

This research adopts a qualitative approach to examine the role of adaptive physical security in organizations. A qualitative methodology is particularly suited for investigating the subjective experiences of stakeholders involved in security practices and for exploring complex phenomena, such as the integration of adaptive security systems in dynamic organizational settings.

3.1 Research Design

The study follows a descriptive qualitative design, aiming to capture the rich, detailed perspectives of individuals who are involved in the implementation or management of physical security measures within their organizations. Unlike quantitative methods, which focus on numerical data, qualitative research offers deeper insights into the meaning and nuances behind security practices and challenges.

3.2 Data Collection Methods

The primary data collection methods include semi-structured interviews and focus group discussions with key personnel in organizations. These participants will include security officers, IT experts, facility managers, and other relevant individuals who contribute to or oversee security operations. The semi-structured interview format allows flexibility, enabling participants to discuss their personal experiences, views, and challenges related to adaptive security systems. Focus groups will provide additional insight into group dynamics and collective opinions regarding security strategies within organizations.

The open-ended nature of these data collection techniques helps the researcher explore complex issues such as organizational resistance, technology integration, and the evolving nature of security threats. Participants will be encouraged to elaborate on their responses, offering a deeper understanding of the topic.

3.3 Sampling Strategy

A purposive sampling strategy will be employed to select participants who have direct experience with security systems. This non-random sampling technique ensures that the study focuses on individuals who can provide valuable insights into the adaptive security processes within their organizations. The study will draw participants from a range of industries, including healthcare, manufacturing, and technology, to capture diverse perspectives on the implementation of adaptive security.

3.4 Data Analysis

The data from the interviews and focus groups will be analyzed using thematic analysis, a widely-used method for identifying patterns and themes within qualitative data. This process involves coding the data, identifying recurring themes, and interpreting them to understand the shared experiences and challenges faced by organizations in implementing adaptive security. According to Braun and Clarke, thematic analysis allows researchers to construct meaningful patterns from qualitative data, helping to develop a comprehensive understanding of the topic.

3.5 Ethical Considerations

This study will adhere to strict ethical standards to ensure the privacy and confidentiality of participants. Informed consent will be obtained from all participants, outlining the purpose of the study and their right to withdraw at any time without penalty. Additionally, all responses will be anonymized, and care will be taken to ensure that sensitive information is protected throughout the research process.

3.6 Limitations

As is the case with many qualitative studies, the findings from this research will not be generalizable to all organizations. The sample size and the specific industries chosen for the study may limit the transferability of the findings to other contexts. Moreover, since qualitative research focuses on understanding individual experiences and perspectives, the results may reflect subjective viewpoints that are not universally applicable.

4 RESULTS

This section presents the findings of the study on adaptive physical security in organizations. The analysis of the semi-structured interviews and focus group discussions revealed several key themes that highlight the challenges and successes organizations experience when integrating adaptive security measures.

4.1 Integration of Technology in Adaptive Security

A dominant theme that emerged was the integration of advanced technologies such as biometric access control systems, AI-driven surveillance, and real-time monitoring tools. Participants consistently noted that these technologies provided a flexible and scalable approach to security, allowing organizations to adapt to evolving threats [6]. For example, one security manager from a healthcare organization stated, "Biometrics have revolutionized our access control, allowing us to ensure that only authorized personnel are accessing sensitive areas in real-time, without the delays traditional systems create."
The use of Internet of Things (IoT) devices also surfaced as a significant component of adaptive security [7]. These devices enable organizations to monitor various physical assets and detect potential security breaches more effectively. However, the implementation of IoT-based systems was seen as challenging due to concerns over data security and the integration of these systems into existing security infrastructure.

4.2 Challenges in Adoption

Another significant theme was the resistance to change from employees and management, which often hindered the adoption of adaptive security measures. Many participants discussed the difficulty in shifting from traditional, static security systems to more dynamic, adaptive ones [8]. A facilities manager from a manufacturing plant mentioned, "It's hard to convince upper management that we need to invest in adaptive systems when the old systems have worked for years without much disruption."
Additionally, budget constraints were frequently mentioned as a barrier to the implementation of adaptive security systems. Many organizations, particularly small to medium-sized enterprises (SMEs), lacked the financial resources to invest in the latest security technologies. As one respondent from a technology firm explained, "We recognize the importance of adaptive security, but funding for upgrading our systems is always a challenge, especially when we have to justify the cost to stakeholders."

4.3 Benefits of Adaptive Security

Despite the challenges, several organizations reported significant benefits from adopting adaptive security measures. Key advantages highlighted included improved incident response times and the ability to manage security more efficiently in real-time. Security managers indicated that adaptive systems allowed them to tailor their security responses based on the nature of the threat, rather than relying on predefined protocols.
Participants also noted that adaptive security systems facilitated proactive risk management. With the ability to integrate real-time data and predictive analytics, organizations could anticipate potential risks and adjust their security protocols before incidents occurred [9]. For instance, a participant from a large retail chain shared, "Our new adaptive system allows us to monitor patterns in real-time, and we can predict when and where security breaches are more likely to occur. This predictive capability has made a huge difference in minimizing theft."

4.4 Organizational Culture and Training

A recurring theme in the results was the importance of organizational culture and staff training in the successful implementation of adaptive physical security systems. Participants stressed the need for a shift in organizational mindset to recognize security as a dynamic, ongoing process rather than a static set of procedures. One security director emphasized, "It's not just about technology; we need to foster a culture where everyone understands the importance of being vigilant and adaptable to new threats."
Training was also identified as a critical factor for the successful adoption of adaptive security systems. Without proper training, staff members may struggle to effectively use new technologies, leading to inefficiencies and potential vulnerabilities.

4.5 Policy and Regulatory Challenges

Finally, the study found that regulatory compliance and policy development were also significant factors impacting the adoption of adaptive security systems. Organizations, particularly those in sectors such as healthcare and finance, face strict regulatory requirements concerning security and data protection. Participants noted that keeping up with changing laws and regulations was a constant challenge when trying to implement new, adaptive systems. One respondent from a financial institution stated, "We need to ensure that our adaptive security solutions comply with industry standards, which can sometimes slow down our ability to implement cutting-edge technologies.

5 DISCUSSION

The findings of this study provide valuable insights into the implementation of adaptive physical security systems in organizations, highlighting both the benefits and challenges encountered in adopting such measures. The results confirm several key trends and suggest practical implications for improving the security landscape in dynamic environments.

5.1 Integration of Technology and Real-time Monitoring

A primary benefit identified in this study was the integration of advanced technologies into adaptive security systems, such as biometric access control and AI-based surveillance. These technologies allow organizations to be more proactive and responsive to emerging security threats. As noted by the participants, technologies such as real-time monitoring and IoT devices provide the flexibility required to adjust security protocols to changing environments. [10] who emphasize that the future of security lies in systems capable of adapting to both physical and digital threats. The shift from traditional security measures to tech-driven, adaptive systems enables organizations to better manage risk and respond to incidents more swiftly. However, the integration of technology also presents challenges. Despite the clear benefits, many participants reported difficulties in fully implementing new systems due to resource constraints, particularly among smaller organizations. This finding echoes the concerns raised in previous studies, such as those by Patton, who note that budget limitations often hinder the adoption of advanced security technologies, especially in industries with tight financial margins. Thus, while technology plays a crucial role in enhancing adaptive security, organizations must consider the financial and infrastructural costs involved in upgrading their security measures.

5.2 Organizational Resistance and Culture

Another significant finding was the resistance to change observed within organizations, particularly in adapting to more dynamic and flexible security protocols. Several participants noted that upper management often viewed traditional security measures as sufficient, which created barriers to adopting new systems. This resistance aligns with the theory of organizational change, which suggests that individuals and organizations may resist change due to factors such as perceived threats to established routines, lack of awareness, or uncertainty about the new systems' efficacy. The reluctance to move away from established practices can slow the transition to adaptive security, making it important for organizations to foster a culture that is open to change and innovation.

As suggested by one of the respondents, training and awareness campaigns are essential to overcoming resistance. The emphasis on organizational culture in this study underscores the importance of aligning security practices with organizational values, which, as stated by [11], is critical for the successful implementation of any new system. Organizations that invest in creating a security-conscious culture will likely experience smoother transitions to adaptive security systems, with employees more likely to understand the importance of evolving security measures.

5.3 The Role of Training and Staff Engagement

The study also revealed that comprehensive training is key to ensuring the successful implementation of adaptive security systems. Respondents highlighted the importance of ongoing staff education and engagement to maintain the effectiveness of new technologies. Without proper training, even the most sophisticated security systems can become underutilized or mismanaged, leading to vulnerabilities. This finding is consistent with research by [11], who emphasizes that training and continuous professional development are crucial for creating a workforce capable of managing complex security systems.

5.4 Compliance with Regulations

A notable challenge for many organizations, particularly in regulated industries such as healthcare and finance, is maintaining compliance with external security and privacy regulations while implementing adaptive security systems. As mentioned by one participant in the study, adhering to regulatory requirements often delays the adoption of advanced security technologies. This finding highlights the tension between innovation and compliance, a topic discussed by [13], who note that organizations must navigate a balance between adopting new technologies and ensuring they meet industry-specific regulatory standards. This challenge suggests that security strategies must be carefully crafted to meet both internal and external requirements, particularly when regulatory bodies impose strict data protection and security mandates.

5.5 Implications for Future Security Strategies

The study's findings suggest several key implications for organizations looking to implement adaptive security systems. First, it is clear that technological integration plays a pivotal role in improving security responsiveness and efficiency. However, organizations must weigh the benefits of advanced systems against the financial and operational challenges associated with their deployment. Additionally, organizational culture and staff training must be central to the process, as

overcoming resistance to change and ensuring effective use of security systems are crucial for long-term success. Finally, organizations must remain vigilant about regulatory changes and ensure that adaptive security measures are compliant with industry standards to avoid legal or financial repercussions.

6 CONCLUSION

This study has explored the implementation and challenges of adaptive physical security systems within organizations, shedding light on the evolving nature of security management in dynamic environments. Through qualitative data collected from interviews and focus groups with security professionals, several key themes emerged regarding the integration of technology, organizational resistance, the role of training, and compliance with regulations.

The findings highlight that adaptive security systems, such as real-time monitoring, biometric access control, and IoT integration, offer significant advantages in improving organizational security. These technologies enable a more flexible and responsive approach to security management, allowing organizations to quickly adjust their protocols to meet emerging threats. However, the adoption of such systems is not without its challenges, particularly concerning budget constraints and resistance to change from within the organization. Moreover, the importance of training staff and fostering a culture of security awareness is critical to ensuring the successful implementation and effectiveness of these adaptive security measures.

The research also underscores the necessity of navigating regulatory frameworks to ensure compliance while adopting advanced security technologies. Balancing innovation with regulatory requirements remains a complex task for organizations, particularly in highly regulated industries such as healthcare and finance.

In conclusion, while adaptive physical security systems provide enhanced protection against evolving threats, successful implementation requires a comprehensive approach. This includes technological investment, overcoming internal resistance, ensuring ongoing staff training, and maintaining compliance with industry standards. Organizations that can effectively integrate these elements into their security strategies will be better positioned to handle future security challenges in an increasingly dynamic and complex landscape.

CONFLICT OF INTEREST

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

REFERENCES

[1] Tambunan, B, Sihombing, H, Doloksaribu, A. The effect of security transactions, easy of use, and the risk perception of interest online buying on the e-commerce tokopedia site (Study on Tokopedia. id site users in Medan city). In IOP Conference Series: Materials Science and Engineering, IOP Publishing. 2018, 420(1): 012118. http://iopscience.iop.org/article/10.1088/1757-899X/420/1/012118/meta

[2] Rajesh, S, Abd Algani, Y M, Al Ansari, M S, et al. Detection of features from the internet of things customer attitudes in the hotel industry using a deep neural network model. Measurement: Sensors, 2022, 22, 100384. DOI: https://doi.org/10.1016/j.measen.2022.100384.

[3] Muda, I, Afrina, A, E. Influence of human resources to the effect of system quality and information quality on the user satisfaction of accrual-based accounting system (Implementing of adaptive behavior assessment system theory, case in Indonesia). Contaduría y Administración, próxima publicación, 2019, 63(4): 1-25. https://www.journals.elsevier.com/contaduria-y-administracion

[4] Abdelkader, S, Amissah, J, Kinga, S, et al. Securing modern power systems: Implementing comprehensive strategies to enhance resilience and reliability against cyber-attacks. Results in engineering, 2024, 102647.

[5] Shulha O, Yanenkova I, Kuzub M, et al. Banking Information Resource Cybersecurity System Modeling. Journal of Open Innovation: Technology, Market, and Complexity, 2022, 8(2): 80. DOI: https://doi.org/10.3390/joitmc8020080.

[6] Ilca, L F, Lucian, O P, Balan, T C. Enhancing cyber-resilience for small and medium-sized organizations with prescriptive malware analysis, detection and response. Sensors, 2023, 23(15): 6757.

[7] Repiso, E, Garrell, A, Sanfeliu, A. Adaptive social planner to accompany people in real-life dynamic environments. International Journal of Social Robotics, 2024, 16(6): 1189-1221.

[8] Sas, M, Reniers, G, Ponnet, K, et al. The impact of training sessions on physical security awareness: Measuring employees' knowledge, attitude and self-reported behaviour. Safety science, 2021, 144, 105447.

[9] Shandilya, S K, Datta, A, Kartik, Y, et al. Advancing Security and Resilience. In Digital Resilience: Navigating Disruption and Safeguarding Data Privacy. Cham: Springer Nature Switzerland. 2024, 459-529.

[10] Surya, B, Hadijah, H, Suriani, S, et al. Spatial transformation of a new city in 2006–2020: Perspectives on the spatial dynamics, environmental quality degradation, and socio—economic sustainability of local communities in Makassar City, Indonesia. Land, 2020, 9(9): 324.

[11] Braun, V, Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

[12] Schein, E H. Organizational culture and leadership. John Wiley & Sons. 2010, 2.
[13] Chang, J, Rabosky, D L, Smith, S A, et al. An R package and online resource for macroevolutionary studies using the ray-finned fish tree of life. Methods in Ecology and Evolution, 2019, 10(7): 1118-1124.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3007

EXAMINING THE MOTIVATION-PERFORMANCE LINK AMONG CONSTRUCTION EMPLOYEES IN WESTERN CAPE, SOUTH AFRICA

Ndukuba Samuel Nnadoziem[1*], Uwa Chukwunonso Aghaegbulam[2]
[1]*Department of Civil Engineering and Geomatics, Cape Peninsula University of Technology, Cape Town, South Africa.*
[2]*Department of Mechanical and Mechatronics Engineering, Tshwane University of Technology, Pretoria, South Africa.*
Corresponding Author: Ndukuba Samuel Nnadoziem, Email: Ndusam4christ@yahoo.com

Abstract: The construction industry, serving as a cornerstone in economic development, necessitates a comprehensive understanding of the determinants influencing the productivity of construction workers. This study delves into the motivating drivers and challenging factors impacting construction workers' productivity in Western Cape, South Africa. The research context underscores the pivotal role of motivation in augmenting worker productivity while acknowledging the adverse effects of challenging elements on job satisfaction and overall project success. Employing a survey research design, the study focused on 200 construction professionals in the Western Cape Province, South Africa. A stratified proportional random sampling technique was employed to ensure a representative sample of 200 participants and only 146 were retrieved. The data collection process encompassed a structured questionnaire, covering demographic characteristics, motivation, and factors influencing productivity. Rigorous analyses, including mean and standard deviation and Pearson Product Moment Correlation, were conducted using the Statistical Package for Social Sciences (SPSS) software, facilitating a systematic examination of the collected data. The noteworthy findings highlight the significance of factors such as recognition by authority, responsibility, provision of healthcare services, working conditions, transportation facilities, promotion opportunities, error tolerance, salary increments, adherence to company policies, and participation in decision-making processes as significant motivators. Conversely, challenges such as lack of cooperation among fellow workers, unfair reward practices, lack of appreciation for job performance, poor supervision, irregular salary payments, inadequate safety measures, lack of respect by supervisors, presence of incompetent crew members, and repetitive tasks were identified. This research highlights the drive relationship between motivation and productivity in the construction industry which emphasises the need to focus on motivational factors that improve productivity problems. In conclusion, it is stated that improving worker motivation and, by extension, productivity in the construction sector, requires addressing both challenging and motivating factors. Construction companies may create a productive and motivating environment for employees by implementing practice strategies such as guaranteeing safety, enhancing supervision, encouraging justice and respect, offering sufficient training, recognising accomplishments, and promoting teamwork. The research findings have significant implications for the Western Cape's construction firms and other industries, as they can help guide the development of motivational strategies that will increase worker output and project success.
Keywords: Construction employees; Motivation; Performance; Western cape

1 INTRODUCTION

The construction sector plays a crucial role in a country's economy, and human resources make the efficient use of other resources possible in large part. In most nations, the construction sector, which encompasses operations from project conception to completion, accounts for a sizeable share of gross capital and GDP. The industry plays a vital part in the South African economy, but in recent years, its performance has fallen short of expectations [1-2]. Decades of consistently poor productivity in the industry underscore the need for immediate action to motivate construction workers to increase productivity to achieve long-term economic growth [3-4]. Project delays, cost overruns, and a low GDP in the construction sector are the results of the global reduction in productivity in the industry, which includes South Africa. Improving labour effectiveness requires addressing elements including physical limitations, safety, ambient circumstances, and worker motivation. Despite this, the industry's practices for increasing worker involvement and efficiency are still lacking, leaving gaps that need attention [5-6].

To solve the persistent problem of low construction worker productivity, motivation is crucial. It is a process that has its roots in human wants and creates an internal motivator that guides behaviour and activity toward goal completion. There are two types of motivation: internal, which stems from internal wants, and external, which is used by firms to satisfy a range of demands and habits among their workforce [7-8]. The significance of motivating workers to improve productivity is demonstrated through the reciprocal relationship between productivity and motivation. Many theories, notably Maslow's hierarchy of needs, challenge the idea that money is the only motivator for construction workers by stressing the complexities of motivation. In employment-intensive industries like construction, efficient HRM goes beyond financial incentives [9-11]. Al-Abbadi and Agyekum-Mensah [12] conducted a recent study to investigate the factors influencing workers' motivation and productivity in the construction industry. The study identified several challenges, including insufficient financial incentives, delayed wage payment, insecure jobs, and unfavourable working

conditions. These elements were shown to hurt employees' motivation, underscoring the critical need for better pay and working conditions to raise employee motivation and, in turn, productivity.

Furthermore, an extensive amount of research has identified important motivating factors that have positive effects on construction workers' productivity. According to research showing that competitive and fair pay increases employee motivation and satisfaction, pay and benefits are essential for attracting and keeping qualified employees [13-14]. Public acknowledgement and incentives based on work performance are highly effective in increasing motivation and efficiency among workers in the construction industry [15-17]. In addition, funding training and development programs enables opportunities for career advancement and skill improvement, which raises employee motivation and ultimately results in increased production [18].

Furthermore, although the importance of worker motivation on productivity is recognized, empirical research addressing the direct relationship between worker performance and motivation in the context of South Africa is noticeably lacking. To gain a comprehensive picture of the total influence of motivation on the construction industry, it is imperative to examine the degree to which highly motivated workers contribute to better project results, timely completion, and lower rates of project abandonment. The literature that currently exists mostly focuses on broad issues and challenges facing the South African construction sector. But to fully understand the unique potential and problems facing the local construction industry, more research needs to be conducted that focuses on the Western Cape, South Africa. Enhancing motivation and productivity in the Western Cape can be achieved through examining how incentive strategies may be tailored to the specific needs and conditions within the area.

Thus, by investigating the relationship between motivational factors and productivity in Western Cape construction firms, this investigation provides insightful information that may be utilised to develop strategies aimed at enhancing productivity and efficiency in the region's building and construction industry. By gaining an in-depth understanding of the relationship between motivation and productivity and identifying the primary motivators for workers, construction firms in the Western Cape can design focused strategies that will motivate employees, successfully address challenges, and promote a more productive work environment. The findings of this study have the potential to improve Western Cape construction firms as well as provide insightful analysis and suggestions that can apply to the broader South African construction industry. These insights can play a crucial role in maximizing employee productivity and achieving project success on a broader scope. For that reason, this research seeks to as ascertain the following objectives:

1. To ascertain the motivation factors that drive' construction workers' productivity in the Western Cape
2. To ascertain the negative factors that challenge construction workers and their influence on productivity in the Western Cape.
3. To establish the relationship between motivation and worker productivity in construction firms in the Western Cape.

2 THEORETICAL REVIEW MASLOW'S AND HERZBERG MOTIVATION

A psychological theory that describes human motivation and development called Maslow's hierarchy of needs has been implemented in several industries, including construction. Maslow's hierarchy of needs highlights basic needs like food, water, and shelter at the base.

Cover. These physiological demands are also met in the construction industry when workers are given safe working conditions, suitable housing, and access to basic amenities [19-21]. These fundamental needs are met in part by appropriate safety precautions, suitable tools, and a hygienic work environment, which gives construction workers a sense of stability and security.

Maslow underscores the importance of safety and belongingness needs, such as social ties and a sense of community, as he moves up the hierarchy. To accomplish these objectives, the construction sector must cultivate a cooperative and stimulating work environment [19]. Construction workers can form strong interpersonal bonds through team-building exercises, open lines of communication, and a favourable work atmosphere. People who experience a sense of support and belonging within the construction team not only have higher job satisfaction but also foster a more effective and productive work environment. When these basic requirements are met, construction workers feel more secure and stable because they have access to the right tools, safety precautions, and a healthy work environment.

Additionally, Maslow's theory can be applied to the requirements for self-actualization and esteem at the higher levels of the hierarchy [22]. These higher-order demands are met by offering chances for skill development, accomplishment recognition, and career growth paths in the construction sector. Construction professionals' self-actualization is facilitated by organizations that provide training programs, mentorship programs, and career advancement opportunities [23]. Workers in the construction business are more likely to be inspired, involved, and fully committed to their work if they have a feeling of personal growth and accomplishment in their roles.

3 HERZBERG MOTIVATION

Herzberg's Two-Factor Theory also known as the Motivation-Hygiene theory has found practical application in the construction industry by addressing key factors that influence employee motivation and job satisfaction [24]. In the construction sector where projects are often complex and demanding, job satisfaction is crucial for maintaining high levels of productivity and ensuring the successful completion of projects [25]. Herzberg's theory identifies two sets of factors: motivators (satisfaction factors) and hygiene factors (dissatisfaction factors).

Within the construction sector, motivators include things like accountability, success, and acknowledgement. Construction firms have adopted numerous strategies to acknowledge and incentivize their workers, including praising exceptional work, offering chances for skill enhancement, and assigning challenging tasks that provide a feeling of achievement [26]. Construction firms seek to establish a work environment that promotes intrinsic motivation by focusing on motivators. This approach is aimed at improving job satisfaction and, as a result, productivity levels [27]. In terms of hygienic factors, construction organizations have taken steps to prevent employee unhappiness by addressing aspects such as working conditions, company policies, and interpersonal relations. Reducing job dissatisfaction in the construction sector involves maintaining a safe and well-equipped work environment, developing transparent and equitable company rules, and encouraging strong interactions among team members [28]. Human resource practices in the industry have been guided by Herzberg's theory, which highlights the significance of addressing motivational elements as well as potential sources of unhappiness to establish a balanced and enjoyable work environment [29].

4 MOTIVATIONAL FACTORS THAT POSITIVELY INFLUENCE THE PRODUCTIVITY OF CONSTRUCTION WORKERS

To increase productivity and guarantee the successful completion of projects, construction workers must be motivated. Creating a secure and encouraging work environment is one important component. Employees are more likely to concentrate on their work without unneeded distractions or worries about their well-being when they feel safe and confident in their surroundings [30-31]. Ensuring access to appropriate safety equipment, conducting frequent training programs, and putting in place extensive safety measures all help to create a good work atmosphere that inspires construction workers to provide their best effort.

Recognition and appreciation for diligent labour are essential for increasing productivity. A sense of pride and pleasure is fostered when construction workers receive acknowledgement and compensation for their hard work and accomplishments, whether through verbal praise, formal recognition programs, or incentives [32] [33]. Employees who feel respected and appreciated are more likely to own up to their mistakes and strive for excellence in their work. Recognising both individual and team achievements inspires workers and fosters a collaborative productive environment on construction sites.

Manoharan et al. [34] suggest that efficient collaboration is another motivating element that raises productivity in the building sector. Workers are better able to organize and complete their work successfully when they are aware of the objectives, deadlines, and deadlines of the project. A transparent workplace reduces uncertainty and irritation through regular team meetings, updates, and open channels of communication with management and supervisors [35-36]. Consequently, effective communication fosters a sense of accountability and trust, which in turn motivates construction workers to take responsibility and contribute to the project's overall success.

In addition, Olanrewaju and Lee [37] claimed that providing opportunities for advancement in their careers is a progressive strategy for inspiring construction workers. Giving employees access to workshops, skill-building activities, and training programs improves their talents and shows support for their professional development. Employee loyalty and dedication are increased when they see that their company values and invests in their skill set [38]. A workforce that is knowledgeable and trained is not only more productive but also better able to deal with obstacles and support innovation and success in the construction sector across all industries.

5 MOTIVATIONAL CHALLENGES ON CONSTRUCTION WORKER'S PRODUCTIVITY

The nature of the labour and the requirements of the construction industry make it challenging to inspire workers to increase their productivity. According to Xing et al. [39], the physically demanding environment of construction labour can cause weariness and exhaustion, which can lower employee motivation. Exhaustion and lack of excitement can be caused by long hours, exposure to different climates, and having to work for frequent responsibilities [28]. The short-term and project-based nature of construction work can make it challenging to establish an efficient and involved workforce [40]. Time constraints are a common feature of construction projects, which puts additional strain on labourers. Ineffective management of this pressure might result in burnout and demotivation. It can also be challenging to sustain constant motivation due to the cyclical nature of construction work, which involves times of high activity interspersed by downtime [41].

Famakin et al. [42] claim that strict rules and safety concerns in the construction sector may unintentionally worsen problems with motivation. Strict safety regulations may be seen as a barrier, which would slow down production and maybe make workers unhappy. Maintaining a motivated workforce requires striking a balance between the demand for efficiency and production and safety precautions [43]

According to a study by Ayodele et al. [44], motivation is potentially affected by the construction industry's unidentified career development pathways. Because of this, employees may find it challenging to perceive the long-term results of the work they do without an established path for advancement, which could affect their dedication and desire. To address these issues, a comprehensive strategy that fosters a culture of motivation and productivity among construction workers is needed. This strategy should include creating a positive work environment, offering opportunities for skill development, and recognizing and rewarding exemplary performance [8, 45].

6 METHODOLOGY

The survey research design that has been selected has been purposefully adopted to facilitate the reliable collection of data about the intricate relationship between worker productivity and motivation in construction enterprises in the Western Cape Province of South Africa. Establishing the attitudes, beliefs, behaviours, and qualities that workers have is a very suitable application for this design. When evaluating numerical data, its quantitative method plays a crucial role in ensuring validity and accuracy in the findings. This methodology works well when examining a large workforce, which is a feature of the construction sector [46]. The research sample size selected was 200 construction professionals who are in the construction firm in Western Cape Province, South Africa, to understand how motivation influences workers' productivity. These selected survey professionals were engaged with ongoing projects, mostly owned by both private and corporate firms. Employing a stratified relative random sampling technique ensures the representation of various perceptions, resulting in a sample size of 146 respondents which constituted 73% of the targeted population size.

7 THE DATA COLLECTION

The data collection process for this study within the construction industry employed a structured questionnaire with two different sections. It should be noted that a five-point Likert-type scale was adopted to determine the relationship between workers' motivation and productivity. Section A captured demographic characteristics, while Section B focused on data concerning workers' motivational challenges and their perspectives on factors influencing productivity in the construction sector. A web-based survey and face-to-face administering were adopted because of the geographical spread of the construction professionals and firms involved in the study. The survey instrument with a supplementary personalised, signed cover letter was sent to the 200 survey respondents through e-mail, it is important to note that out of 200 sent e-mails and face-to-face administering, 146 were duly completed and returned using the web survey and face-to-face collection and hence an overall response rate of 73% was achieved. The direct delivery and retrieval method facilitated the efficient collection of data. In terms of the data analysis, descriptive analyses, encompassing measures such as mean and standard deviation, were utilized to rank the motivational drivers and challenges that influence construction worker's productivity. The data analysis approach incorporated a combination of descriptive and inferential statistical methods. Descriptive statistics were employed to assess central tendencies, including mode, median, and mean, as well as to evaluate data dispersion using measures such as standard deviation. Inferential statistics were utilised to validate the data obtained. These analyses provided a succinct summary and description of the study findings, offering insights into the motivational aspects of construction workers. The use of Statistical Package for Social Sciences (SPSS) software ensured a systematic and rigorous analysis of the collected data, emphasising the relevance of the findings to the construction industry and contributing to the advancement of knowledge in this field.

8 FINDINGS AND DISCUSSION

8.1 Section A: Demographic Profiles

Table 1 Distribution of Respondents Based on Age and Educational Qualification, Professional Roles, Educational Specialization and Years

Variables	Label	Freq.	Percentage %
Age of Respondents	Below 25yrs.	10	6.8
	25 – 30yrs	21	14.4
	31 – 35yrs	46	31.5
	36 – 40yrs	36	24.7
	41 – 45yrs	25	17.1
	46 - 50yrs	5	3.4
	51 above yrs	3	2.1
	Total	146	100
Educational Qualification	ND (National Diploma)	23	15.8
	BSc / BTech /Advanced Dip	63	46.6
	Honours degree	29	19.9
	Master's degree	13	8.9
	PhD	4	2.7
	Other	9	6.2
	Total	146	100
Professional Roles	Architect	11	7.5
	Project manager	34	23.3
	Site Engineer	15	10.3
	Quantity surveyor	63	43.2
	Supervising Builder (Foreman)	23	15.8
	Total	146	100
Years Spent with the firm	Less than 5 Yrs.	37	25.3
	6-10 Yrs.	62	42.5

	11-15 Yrs.	29	19.9
	16 Above Yrs.	18	12.3
	Total	146	100
Professional Roles	Architect	11	7.5
	Project manager	34	23.3
	Site Engineer	15	10.3
	Quantity surveyor	63	43.2
	Supervising Builder (Foreman)	23	15.8
		146	100

Table 1 above presents the distribution of respondents based on their age. The result shows that respondents below 25 years have a frequency of 10 representing 6.8%, followed by those whose ages are between 25—30 years with frequencies of 21 representing 14.4% while those between 31-35 have the highest frequencies of 46 representing 31.5% and 36 - 40 years is the second highest frequencies of 36 representing 24.7% while as 41 – 45 years, 46 – 50, 51 above years and above got the frequencies of 25, 5 and 3 representing 17.1%, 3.4% and 2.1% respectively. The percentage distribution shows that most of the respondents in the study area are below 30 years old. Further, the table presents the distribution of respondents based on their Educational Qualifications. The result shows that respondents with Matric certificates have a frequency of 15 representing 10.3%, followed by ND (National Diploma) holders with frequencies of 23 representing 15.8% while respondents with BSc / BTech /Advanced Dip holders have the highest frequencies of 68 representing 46.6% whereas Honours degree holders have the second highest frequencies of 29 representing 19.9%. The master's degree, PhD and Others got the lowest frequencies of 13, 4 and 9 representing 8.9%, 2.7% and 6.2% respectively. The percentage distribution shows that most of the respondents in the study are BSc / BTech /Advanced Dip holders.

Furthermore, the above illustrates the distribution of respondents based on their professional roles within the construction industry. The findings indicate that Quantity Surveyors recorded the highest frequencies, comprising 63 respondents, representing 42.3%. Following closely are Project Managers, with frequencies of 34, constituting 23.3%, while Supervising Builders (Foremen), Site Engineers, and Architects had the lowest frequencies at 23, 15, and 11, representing 15.8%, 10.3%, and 7.5%, respectively. Additionally, the table presents the distribution of respondents based on their tenure with the company. The results reveal that individuals with 6-10 years of experience garnered the highest frequencies, totalling 62 and accounting for 42.5%. In contrast, those with less than 5 years of experience comprised 37 respondents, representing 25.3%. Respondents with 11-15 years and 16 years and above had the lowest frequencies, with 29 and 18 respondents, constituting 19.9% and 12.3%, respectively. The percentage distribution highlights that a significant portion of respondents in the study area have spent 6-10 years with their respective construction firms.

Table 2 Motivational Drivers that Influence Construction Workers' Productivity in the Western Cape?

Motivational drivers	Means	Std Dev	Rank
Recognition by authority	3.89	0.72	1st
Responsibility	3.67	0.83	2nd
Provision of good Health Care services	3.54	0.76	3rd
Working Overtime	3.51	0.73	4th
Provision of Transportation facility	3.46	0.77	5th
Opportunity to be promoted	3.39	0.85	6th
Error Tolerance	3.33	0.74	7th
Increase in Salary	3.26	0.94	8th
Company policy	3.25	0.73	9th
Taking part in the decision	3.18	0.88	10th
Working condition	3.15	0.69	11th
The work itself	3.07	0.71	12th

8.2 Discussion of Findings One

The descriptive statistic as indicated in Table 2 revealed that recognition by authority is the most significant motivational factor influencing construction worker's productivity, ranked 1st with a MS of 3.89 and SD of 0.72. Motivational drivers with mean scores of 3.67 and 3.54, respectively, are placed second and third and include delegating responsibilities and offering high-quality healthcare services. The study's conclusions about the crucial motivational factors affecting the productivity of construction workers in the Western Cape are consistent with previous findings in the field. Important motivators in the construction industry have been highlighted, including acknowledgement by authority, responsibility, the availability of quality health care, and putting in extra hours of labour [47] [48]. These factors have an important influence on workers' overall health and satisfaction with their work, which in consequence improves their drive and productivity. Additionally, the research highlights how crucial it is to provide transportation, opportunities for advancement, error tolerance, pay raises, dedication to company policies, the ability to participate in decision-making processes, and favourable working conditions, all of which have a significant impact on the

productivity of Western Cape construction workers. The findings agree with previous research that emphasised the importance of monetary incentives, professional growth possibilities, and employee empowerment in inspiring construction workers [49][50]. The identification of specific motivators provides Western Cape-based construction firms with valuable knowledge. It underlines the need to identify and address these motivators in a way that promotes worker morale, commitment, and productivity. As a result, the study clarifies the key motivating factors influencing the productivity of construction workers in the Western Cape and provides recommendations to construction firms on how to create motivational strategies that will increase worker productivity and project success in general.

Research Question Two: What are the challenges to worker motivation in Western Cape construction firms that affect productivity?

Table 3 Motivational Challenges on Construction Worker's Productivity

Motivational challenges	Means	Std Dev	Rank
Lack of cooperation from fellow worker	3.51	0.91	1st
Unfairness in giving reward	3.43	0.72	2nd
Lack of appreciation for a job well done	3.40	0.80	3rd
Irregular salary	3.35	0.83	4th
Poor supervision	3.33	0.78	5th
Poor safety measures	3.31	0.86	6th
Doing the same work more than one time	3.28	0.79	7th
Little achievement	3.24	0.73	8th
Lack company policy	3.23	0.71	9th
Lack of respect by supervisors	3.17	0.70	10th
Poor working condition	3.12	1.09	11th
Incompetent crew members	3.10	1.12	12th

Descriptive data are shown in Table 3 of the following set of findings, which highlights key demotivating factors affecting construction workers' productivity in the Western Cape geographical area. This provides insightful information about the challenges that are addressed in this construction industry.

These findings agree with earlier studies that identified many different factors that hurt worker productivity. One important finding illustrates the negative impacts of a lack of teamwork among colleagues, underlining the important necessity of promoting teamwork in the construction industry. Strong teamwork dynamics must be established to address this issue, together with suitable resources and training [49-52].

The inequity of rewarding practices—both financial and non-financial—has also been noted as a challenging factor. This unfairness creates chaos among workers, which consequently lowers motivation and productivity. Implementing timely and equitable reward systems is essential to keeping employees motivated and productive. In the construction firm, there are also challenges of management and leadership that demotivate workers. These challenges include repeated tasks, insufficient supervision, irregular salaries, a lack of recognition for a job well done, and inadequate safety measures [53]. Understanding the importance of motivation and proactiveness will foster a positive work atmosphere whereby helps construction firms mitigate productivity challenges and enhance overall project success.

While the findings pertain specifically to the Western Cape site, acknowledging potential limitations in generalisability to other regions or contexts, the consistent recognition of the relationship between motivation and productivity in construction across various studies reinforces the validity and relevance of these findings. This correlation emphasizes the need for prioritising and addressing motivational factors to elevate worker productivity in the construction industry. As motivation levels increase, the positive association suggests a concurrent rise in productivity levels among construction workers, underlining the pivotal role of motivation in the construction sector.

Discussion of Finding Three: Is there any relationship between workers' motivation and workers' productivity in construction firms in the Western Cape sites?

9 FINDINGS

Table 4 Factor analysis- Data Correlation Matrix

Indicators	Motivational drivers	Challenges factors	Workers Productivity
Motivational drivers	1.00		
Challenges factors	.73	1.00	
Workers Productivity	.64	.69	1.00

The information that was observed correlation matrix, as was generated by using factor analysis regarding each variable, is shown in Table 4 (i.e., Motivational drivers, challenges factors and workers' productivity).

The findings demonstrated that all the measured variables ranged from 0.001 to 1.00, indicating that there are sufficient variables that can measure productivity.
Relationship between Workers' motivation and productivity.

Table 5 SAS PCA Output

Eigenvalues of the Correlation Matrix: Total = 10 Average = 1				
Items	Eigenvalue	Difference	Proportion	Cumulative
Motivational drivers	2.21267568	0.475351250	0.31034	0.51453
Challenges factors	2.51384526	1.54423985	0.3152	0.8121

The results of the SAS PCA Output, which was obtained by conducting factors analysis for all variables (worker productivity, challenging factors, and motivational drivers), are shown in Table 5. The findings showed that all the variables are adequate for measuring productivity among workers.

Table 6 Correlation Analysis

Correlations				
		Motivational drivers	Challenges factors	Worker's productivity
Motivational drivers	Pearson Correlation	1		
	Sig. (2-tailed)	.000		
	N	146		
Challenges factors	Pearson Correlation	.572**	1	
	Sig. (2-tailed)	.000	.001	
	N	146	146	
Worker's productivity	Pearson Correlation	.612**	.653**	1
	Sig. (2-tailed)	.000	.000	.000
	N	146	146	146
**. Correlation is significant at the 0.01 level (2-tailed).				

Table 6 illustrates the correlation analysis examining the relationship between workers' motivations (both drivers and challenges) as independent factors and workers' productivity as the dependent factor. The findings indicate a significant positive correlation between motivational drivers and worker productivity, with a Pearson correlation coefficient of r= .572** ($p<0.01$). Similarly, there is a strong positive correlation between motivational challenges and worker productivity, as reflected by a Pearson correlation coefficient of r= .612** ($p<0.01$). Furthermore, the analysis reveals a positive and robust correlation between these factors, with a Pearson correlation coefficient of r= .653** ($p<0.01$). The positive correlation indicates that motivation links to the increased sustainable productivity of construction workers which supports the established theories and previous research that highlights the critical motivational role in driving employee performance [11, 47].
The implications of this finding for construction firms working in the Western Cape province add to its significance. Management and other decision-makers can better develop strategies to increase employee motivation in the construction industry by taking into consideration the fundamental link between productivity and motivation. Construction firms may successfully increase productivity among their workers by implementing programs that address motivating factors including pay, prospects for career advancement, working environment, recognition, and job autonomy.

10 CONCLUSION AND RECOMMENDATIONS

The study's findings demonstrate how certain motivational factors, such as having access to high-quality healthcare, receiving recognition from managers, and having a comfortable workplace, are critical in determining how productive construction workers are in the Western Cape. This underlines how important it is to deal with motivating problems facing the construction industry to improve the productivity of workers as well as overall project success. The study demonstrates a strong correlation between productivity and motivation, underscoring the need for construction firms to proactively address motivational issues. The study also identifies demotivating challenges, such as insufficient safety precautions, inconsistent wage payments, and a lack of respect. These findings demonstrate how important it is for Western Cape construction firms to address these issues and foster a positive work environment. By implementing the suggested approaches into practice, construction firms can solve issues with motivation among employees, create a positive work environment, and maximise worker productivity. This will consequently assist the Western Cape's construction industry to perform healthier in general and deliver more effective results from projects.

Recommendations and Future Research:

Western Cape province construction firms are strongly urged to give their full attention to implementing all-encompassing motivating strategies within their company's operations. Focus should be placed on important objectives including transparency, authority recognition, responsibility, access to quality healthcare, working overtime, and favourable working conditions. Construction firms can greatly increase motivation among workers and, as a result, overall productivity by focusing on these indicators. The firms also need to take proactive measures to overcome demotivating issues. This can be accomplished by enforcing tight adherence to safety regulations, establishing equitable

compensation structures, cultivating an environment of deference and gratitude, and making significant investments in training and development to improve worker effectiveness. By taking these steps, the construction industry will have an improved workplace environment, which will boost worker efficiency and drive.

Moreover, a key factor in motivating construction workers is skilled management and leadership. Construction firms should place a high priority on creating a work environment that values individual growth, promotes involvement in decision-making, strengthens teamwork, and facilitates effective collaboration. For the construction industry to become more productive overall, it is essential to create a work atmosphere that drives motivation.

Future research projects could investigate the long-term implications of these motivational techniques on worker retention rates and creative ways of integrating technological advances for increased productivity and satisfaction among employees. These studies could yield insightful information that will help the Western Cape and the construction industry continue to improve.

COMPETING INTERESTS

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

REFERENCE

[1] Alaloul WS, Musarat MA, Rabbani MBA, et al. Construction sector contribution to economic stability: Malaysian GDP distribution. Sustainability, 2021, 13(9): 5012.

[2] Owolabi J D, Ogunbodede E F, Afolabi A O. The Role of Leadership in Motivating Construction Workers in Nigeria. Journal of Construction in Developing Countries, 2020, 25(1): 33-50.

[3] Dixit S, Mandal SN, Thanikal JV, et al. Study of significant factors affecting construction productivity using relative importance index in the Indian construction industry. In E3S web of conferences, EDP Sciences, 2019, 140: 09010.

[4] George M L. Worker motivation and productivity in the construction industry. Journal of Construction Engineering and Project Management, 2019, 9(2): 17-27.

[5] González Navarro F, Selva Olid C, Sunyer Torrents A. The influence of total compensation on job satisfaction. Universitas Psychologica, 2022, 20: 1-15.

[6] Zong Z, Long T, Ou Y, et al. Dual-path influence of risk perception on construction workers' safety participation and the moderating role of mindfulness. Journal of Construction Engineering and Management, 2025, 151(1): 04024194.

[7] Sitopu YB, Sitinjak KA, Marpaung FK. The influence of motivation, work discipline, and compensation on employee performance. Golden Ratio of Human Resource Management, 2021, 1(2): 72-83.

[8] Paais M, Pattiruhu JR. Effect of motivation, leadership, and organizational culture on satisfaction and employee performance. The Journal of Asian Finance, Economics and Business, 2020, 7(8): 577-588.

[9] Nguyen GN, Mani V, Kha M, et al. Supply chain social responsibility in labour-intensive industries: a practitioner's perspective. Production Planning & Control, 2023, 34(4): 371-390.

[10] Zhong Y, Li Y, Ding J, et al. Risk management: Exploring emerging Human Resource issues during the COVID-19 pandemic. Journal of Risk and Financial Management, 2021, 14(5): 228.

[11] Locke E A, Latham G P. What should we do about motivation theory? Six recommendations for the twenty-first century. Academy of Management Review, 2004, 29(3): 388-403.

[12] Al-Abbadi GMD, Agyekum-Mensah G. The effects of motivational factors on construction professionals productivity in Jordan. International Journal of Construction Management, 2022, 22(5): 820-831.

[13] Emmanuel N, Nwuzor J. Employee and Organisational Performance: Employees Perception of Intrinsic and Extrinsic Rewards System. Applied Journal of Economics, Management and Social Sciences, 2021, 2(1): 26-32.

[14] Bussin MH, Serumaga-Zake P, Mohamed-Padayachee K. A total rewards framework for the attraction of Generation Y employees born 1981–2000 in South Africa. SA Journal of Human Resource Management, 2019, 17(1): 1-14.

[15] RAMYA S, VANITHAMANI M. THE POWER OF EMPLOYEE RECOGNITION: BUILDING A CULTURE OF APPRECIATION IN THE WORKPLACE. JOURNAL OF TECHNICAL EDUCATION, 2023, 109: 109.

[16] Ihemereze KC, Eyo-Udo NL, Egbokhaebho BA, et al. Impact Of Monetary Incentives on Employee Performance In The Nigerian Automotive Sector: A Case Study. International Journal of Advanced Economics, 2023, 5(7): 162-186.

[17] Ogunsanmi O. Enhancing motivation and productivity in construction firms: A case study of selected construction firms in Nigeria. Journal of Construction in Developing Countries, 2018, 23(2): 123-138.

[18] Atnafu A, Hailemariam T, Endehabtu BF, et al. Implementation Outcomes of Performance Based Non-Financial Incentive: using RE-AIM framework. Ethiopian Journal of Health Development, 2023, 37(1).

[19] Ihensekien OA, Joel AC. Abraham Maslow's Hierarchy of Needs and Frederick Herzberg's Two-Factor Motivation Theories: Implications for Organizational Performance. Romanian Economic Journal, 2023, 26(85).

[20] Acquah A, Nsiah TK, Antie ENA, et al. Literature review on theories of motivation. EPRA International Journal of Economic and Business Review, 2021, 9(5): 25-29.

[21] Maksum I. Integration of Needs into a Qur'an Perspective Using Maslow and Herzberg's Motivation Theory. Saudi Journao of Humanities Social Science, 2021, 6(9): 354-362.
[22] Rojas M, Méndez A, Watkins-Fassler K. The hierarchy of needs empirical examination of Maslow's theory and lessons for development. World Development, 2023, 165: 106185.
[23] Stephen J. Broadening the Path to Self-Actualization through Systematic Changes to Maslow's Hierarchy of Needs (Doctoral dissertation), 2023.
[24] Yousaf S. Dissection of Herzberg's Two-Factor Theory to Predict Job Satisfaction: Empirical Evidence from the Telecommunication Industry of Pakistan. Lahore Journal of Business, 2020, 8(2).
[25] Tezel MS. Herzberg's Two-Factor Theory, and knowledge workers' motivation and job satisfaction: a study on academicians at foundation universities (Master's thesis, Işık Üniversitesi), 2023.
[26] Sugathadasa R, De Silva, ML Thibbotuwawa, et al. Motivation factors of engineers in private sector construction industry. Journal of Applied Engineering Science, 2021, 19(3): 794-805.
[27] Lakshitha M, Sugathadasa PRS, Bandara KACP, et al. MOTIVATION FACTORS OF ENGINEERS IN PRIVATE SECTOR CONSTRUCTION INDUSTRY. Journal of Applied Engineering Science, 2021, 19(3): 794-805.
[28] Lee B, Lee C, Choi I, et al. Analyzing determinants of job satisfaction based on two-factor theory. Sustainability, 2020, 14(19): 12557.
[29] Ganesh R, Liu Y. Employee satisfaction in the sales department of the automobile industry in Beijing, China: an approach with Herzberg's two-factor theory. International Journal of Services, Economics and Management, 2022, 13(1): 57-77.
[30] Jeelani I, Gheisari M. Safety challenges of UAV integration in construction: Conceptual analysis and future research roadmap. Safety Science, 2021, 144: 105473.
[31] Prodanova J, Kocarev L. Is job performance conditioned by work-from-home demands and resources?. Technology in Society, 2021, 66: 101672.
[32] Kumari K, Barkat Ali S, Un Nisa Khan N, et al. Examining the role of motivation and reward in employees' job performance through the mediating effect of job satisfaction: An empirical evidence. International Journal of Organizational Leadership, 2021, 10(4): 401-420.
[33] Baqir M, Hussain S, Waseem R, et al. Impact of reward and recognition, supervisor support on employee engagement. American International Journal of Business and Management Studies, 2020, 2(3): 8-21.
[34] Manoharan K, Dissanayake P, Pathirana C, et al. Assessment of critical factors influencing the performance of labour in Sri Lankan construction industry. International Journal of Construction Management, 2023, 23(1): 144-155.
[35] Hamza M, Shahid S, Bin Hainin, et al. Construction labour productivity: a review of factors identified. International Journal of Construction Management, 2022, 22(3): 413-425.
[36] Chien GC, Mao I, Nergui E, et al. The effect of work motivation on employee performance: Empirical evidence from 4-star hotels in Mongolia. Journal of Human Resources in Hospitality & Tourism, 2020, 19(4): 473-495.
[37] Olanrewaju AL, Lee AHJ. Investigation of the poor-quality practices on building construction sites in Malaysia. Organization, Technology and Management in Construction: an International Journal, 2022, 14(1): 2583-2600.
[38] Turner CJ, Oyekan J, Stergioulas L, et al. Utilizing industry 4.0 on the construction site: Challenges and opportunities. IEEE Transactions on Industrial Informatics, 2020, 17(2): 746-756.
[39] Xing X, Zhong B, Luo H, et al. Effects of physical fatigue on the induction of mental fatigue of construction workers: A pilot study based on a neurophysiological approach. Automation in Construction, 2020, 120: 103381.
[40] Greco M, Grimaldi M, Locatelli G, et al. How does open innovation enhance productivity? An exploration in the construction ecosystem. Technological Forecasting and Social Change, 2021, 168: 120740.
[41] Sun M, Zhu F, Sun X. Influencing factors of construction professionals' burnout in China: a sequential mixed-method approach. Engineering, Construction and Architectural Management, 2020, 27(10): 3215-3233.
[42] Famakin IO, Aigbavboa C, Molusiwa R. Exploring challenges to implementing health and safety regulations in a developing economy. International Journal of Construction Management, 2023, 23(1): 89-97.
[43] Ebekozien A. Construction companies' compliance to personal protective equipment on junior staff in Nigeria: issues and solutions. International journal of building pathology and adaptation, 2022, 40(4): 481-498.
[44] Ayodele OA, Chang-Richards A, González V. Factors affecting workforce turnover in the construction sector: A systematic review. Journal of Construction Engineering and Management, 2020, 146(2): 03119010.
[45] Dodanwala TC, Santoso DS, Yukongdi V. Examining work role stressors, job satisfaction, job stress, and turnover intention of Sri Lanka's construction industry. International Journal of Construction Management, 2023, 23(15).
[46] Nardi PM. Doing survey research: A guide to quantitative methods. Routledge, 2018: 2583-2592.
[47] Lu Y. Motivation factors influencing construction workers' productivity: A comparative analysis. Construction Management and Economics, 2020, 38(5): 431-446.
[48] Choudhry R M. Factors affecting motivation of construction professionals: A comparative study of the UK and Pakistan. Journal of Construction Engineering and Management, 2017, 143(10).
[49] Odeh A M, Battaineh H T. Factors influencing workers' productivity in the building construction projects: The case of the Gaza Strip. Journal of Engineering, Project, and Production Management, 2016, 6(2): 69-82.
[50] Olanrewaju OI, Bello AO, Semiu MA, et al. Critical barriers to effective communication in the construction industry: evidence from Nigeria. International Journal of Construction Management, 2024: 1-19.

[51] Bakri S. Safety as a demotivating factor in construction projects: A case study of Yemen. Journal of Engineering, Design, and Technology, 2019, 17(2): 383-397.
[52] Odeyinka H A, Kaka A P. Influence of social factors on Productivity of building construction workers in Nigeria. Journal of Construction Engineering, Project, and Management, 2015, 5(2): 49-59.
[53] Zhang H. Motivating construction workers in China: The role of supervisory support, intrinsic motivation, and extrinsic motivation. International Journal of Project Management, 2016, 34(2): 311-318.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3009

5G ABNORMAL SIGNALING DETECTION BASED ON AUTO-ENCODER

YaDi Fu
School of Cyber Security, Beijing University of Posts and Telecommunications, Beijing 100876, China.
Corresponding Email: 2022111026@bupt.cn

Abstract: With the proliferation of 5G networks, increasing attention has been directed toward associated security risks. The N4 interface, serving as the interface between the user plane and control plane in 5G architecture, encompasses functionalities including session management and policy enforcement, and is susceptible to risks such as session hijacking. PFCP, as the application layer protocol of the N4 interface, can be effectively monitored through anomaly detection to identify abnormal behaviors within the N4 interface. Consequently, this paper proposes an autoencoder algorithm model implemented with Transformer neural networks. During the training process, the model learns the sequential characteristics of normal PFCP bidirectional flows. In the detection phase, data is processed through the encoder and decoder, and the model computes the Euclidean distance between the reconstructed data and the original data to derive an anomaly score. Additionally, this paper employs publicly available datasets to experimentally validate the efficacy of the algorithmic model in detecting PFCP traffic anomalies.
Keywords: 5G security; Signaling detection; N4 interface; PFCP

1 INTRODUCTION

With the evolution of mobile communication technologies, society has progressively transitioned into the 5G era. 5G represents the fifth generation of mobile communication technology, standardized by the 3GPP (3rd Generation Partnership Project) and endorsed by the ITU (International Telecommunications Union). In comparison to 4G networks, 5G mobile communication networks deliver enhanced speed and capacity for inter-device communications, while offering considerable flexibility through network slicing functionality that enables customization of network capabilities according to specific service scenarios. The scope of 5G extends beyond traditional human-to-human communications to encompass human-to-machine and machine-to-machine interactions, thereby significantly broadening its application domains. Critical services and vertical industries supported by 5G technology include industrial internet, vehicular networks, Internet of Things (IoT), and intelligent healthcare, all of which impose stringent requirements on network reliability and security[1]. The proliferation of terminal devices connecting to networks and the substantial integration of IoT devices simultaneously expands the attack surface and compounds the complexity of security management. As 5G networks progressively integrate with diverse societal sectors, their implications for national and social security become increasingly profound, resulting in heightened attention to 5G security vulnerabilities[2].

5G networks not only introduce novel technologies but also adopt new signaling protocols, such as PFCP protocols. The PFCP protocol functions as the application protocol for the N4 interface between Session Management Function (SMF) and User Plane Function (UPF). In wireless communication networks, signaling protocols facilitate fundamental network management and mobility management functionalities, specifically encompassing user authentication, authorization, billing, terminal state transitions, and terminal handovers. In an era of exponentially increasing mobile communication users, attacks targeting signaling protocols have proliferated, including the prevalent signaling storm attacks, which consume substantial network bandwidth resources, compromise network equipment processing capabilities, and in severe cases, precipitate network disruptions. Additionally, attackers exploit protocol vulnerabilities to execute session hijacking, eavesdropping, and other malicious activities; these security risks similarly extend to 5G's novel signaling protocols.

Signaling traffic in 5G networks, as well as in broader mobile communication networks, serves the critical function of information transmission. Consequently, the analysis and detection of signaling traffic enable the identification of certain behaviors within 5G networks. In high-volume traffic environments, anomalous traffic detection is extensively employed in security domains such as intrusion detection and attack identification. The detection of anomalous signaling traffic to identify security threats constitutes a common and efficacious security approach in mobile communication networks.

This paper proposes an autoencoder-based algorithm for detecting abnormal PFCP signaling traffic. This algorithm can train a model capable of detecting anomalous PFCP signaling behaviors even when only normal datasets are available. The performance of the model is validated using the public PFCP intrusion dataset provided by George. To address the issue of incomplete normal PFCP signaling data in this dataset, this paper establishes a 5G simulation environment based on OAI to simulate a more comprehensive and diverse range of normal PFCP signaling traffic.

2 RELATED WORK

With the advancement of 5G networks, the entire telecommunications industry is endeavoring to address the security challenges inherent in 5G architecture, technology, and services. The security architecture defined by the 3GPP committee encompasses three security layers and six security domains. The three security layers comprise the Transport Layer, Home Service Layer, and Application Layer. The six security domains consist of Network Access Security, Network Domain Security, User Domain Security, Application Domain Security, Service-Based Architecture (SBA) Domain Security, and Security Visibility and Configurability.

In February 2020, China's IMT-2020 (5G) Promotion Group compiled and published the "5G Security Report," which systematically analyzed 5G security challenges from two perspectives: key technologies and typical scenarios. The report elucidated that new 5G technologies introduce novel security challenges, necessitating the refinement of security measures in accordance with the specific characteristics of various 5G vertical domains.

Regarding the specifics of 5G signaling security, Hu[3] conducted an analysis of the HTTP/2 signaling protocol utilized between core network elements, identifying potential attacks facilitated by HTTP/2, including stream multiplexing attacks and header compression attacks. Inspired by DDoS attacks, George[4] investigated PFCP protocol-based attacks under the assumption that attackers had obtained N4 interface access. Their research encompassed unauthorized PFCP session deletion requests, unauthorized PFCP session modification requests, and unauthorized PFCP session establishment flooding attacks.

Numerous security solutions predicated on 5G traffic detection have emerged. Radivilova T[5] synthesized and experimentally evaluated existing anomalous traffic detection methodologies, conducting tests on authentic data sets with numerical characteristics approximating 5G traffic, subsequently comparing experimental outcomes and analyzing their distinctive features and appropriate application scenarios. LORENZO[6] proposed an adaptive deep learning-based anomaly detection system for 5G networks after comprehensive consideration of 5G network architecture. Pacherkar[7] introduced a security framework featuring a traceable graph for malicious flow detection, primarily targeting three attacks: SMS storm attacks, PFCP attacks, and network slicing attacks. Robert[8] simulated PFCP signaling attacks to generate normal and anomalous PFCP protocol datasets, subsequently employing LSTM neural networks for the detection of anomalous PFCP signaling.

3 METHOD

3.1 Overview

As illustrated in Figure 1, the algorithmic framework comprises three core modules: the PFCP Traffic Parsing Module, the PFCP Traffic Aggregation Module, and the Transformer-AE Model.

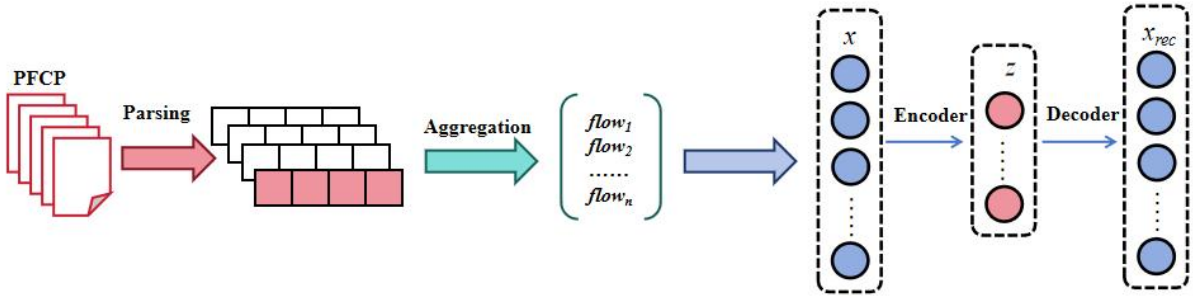


Figure 1 Overview of Anomaly Detection Framework

3.2 PFCP Traffic Parsing Module

This module is responsible for multi-level parsing of PFCP traffic. At the data link layer (MAC layer), timestamp information is parsed, establishing the foundation for subsequent temporal information learning. At the network layer (IP layer), source IP address and destination IP address parameters are extracted. At the transport layer, the source port and destination port fields are parsed. The packet parsing results from the network and transport layers will be utilized for subsequent traffic aggregation processing. At the application layer, the system parses PFCP signaling information and extracts PFCP message type fields. The final output format can be represented as: packet = [$feature_1$, $feature_2$, ..., $feature_n$], where n denotes the number of parsed fields for each packet.

3.3 Traffic Aggregation Module

In this phase, the input is packets = [$packet_1$, $packet_2$, ..., $packet_m$], i.e., m packets, with each packet containing n parsed fields. Based on source IP address, destination IP address, source port, and destination port parameters, the system aggregates packets into multiple bidirectional flows. Each flow is represented as flow = [$packet_1$, $packet_2$, ..., $packet_N$], where $N \leq 100$. The final output of this module is a collection of multiple flows, represented as flows = [$flow_1$, $flow_2$, ..., $flow_M$].

3.3 Transformer AE Module

In the model implementation process, both the encoder and decoder employ Transformer neural network architecture. As an emerging neural network structure, the Transformer[9] encompasses both encoder and decoder components. The encoder consists of a stack of identical layers, with each layer containing two sub-layers: a multi-head attention mechanism and a feed-forward network. Residual connections are applied around each sub-layer, followed by layer normalization. The decoder's structure is similar to the encoder's, also comprising multiple stacked layers, but each layer contains three sub-layers, namely, an additional multi-head attention mechanism that operates on the encoder's output is inserted between the encoder's two sub-layers.

In the Transformer AE[10] algorithm, the first layer of the encoder is a linear layer responsible for mapping the original input dimensions to the model's internal dimensions, ensuring input data compatibility with Transformer processing dimensions. Given that the Transformer neural network structure inherently lacks positional information, sequential positional encoding is necessary. This research employs sinusoidal and cosine functions to generate fixed positional encodings. After positional encoding, the system applies a dual-layer Transformer encoding structure to encode the data. The decoder, conversely, utilizes a dual-layer Transformer decoding structure to generate data from the input latent space.

During the training process, Mean Squared Error (MSE) is adopted as the loss function. In the anomaly detection phase, the model calculates an anomaly score for the input data, The formula is expressed as: $\mathrm{A} = \| D(E(x)) - x \|_2$

For the anomaly scores output by the Transformer AE model, this research employs a validation set method, optimizing evaluation metrics within the validation set to determine a threshold. In practical anomaly detection processes, when the anomaly score output by Transformer AE exceeds this threshold, it is classified as anomalous; conversely, it is classified as normal.

4 EXPERIMENTS

4.1 DataSet

4.1.1 Introduction of PFCP intrusion dataset

This paper utilizes the publicly available PFCP intrusion dataset provided by George for validation. The dataset includes PFCP session establishment DoS attacks, PFCP session deletion DoS attacks, and PFCP session modification flooding attacks. The aim of PFCP Session Establishment DoS Attack is to exhaust the resources of the UPF by inundating it with genuine Session Establishment Requests and Heartbeat Requests. The goal of PFCP Session Deletion DoS Attack is to disconnect a specific UE from the DN. The purpose of PFCP session modification flooding attacks is to attempt to alter session flows through a large volume of packets, thereby achieving the attack objective. There are two specific methods for this type of attack. The one is to invalidate packet handling rules for a specific session, leading to the disassociation of a targeted UE from the DN. The other one is to utilise the DUPL flag in the Apply Action field to compel the UPF to replicate rules for the session, generating multiple paths for the same data from a single source.

4.1.2 Normal PFCP signaling dataset

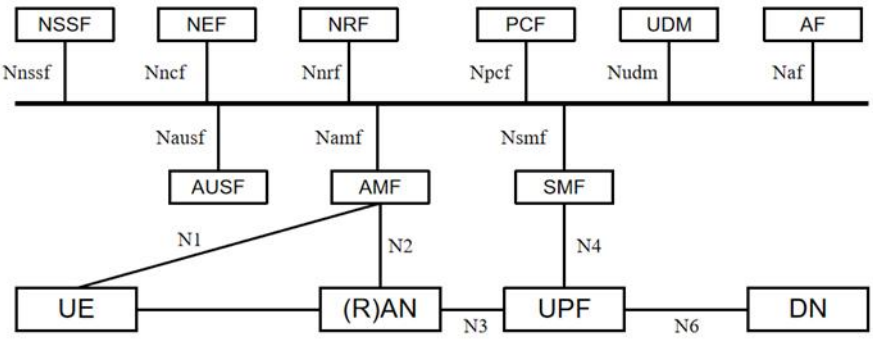


Figure 2 Overview of 5G Network Architecture

The 5G core network environment utilizes OAI simulation. The fundamental architecture of the 5G core network is illustrated in Figure 2. The network elements are constructed based on OAI network element packages. Multiple virtual machines are established in a Linux environment with appropriate network design and configuration. Each virtual machine functions as a distinct network element, interconnected within the server-provided network. The UE and base station are simulated using EXFO. These components collectively constitute the 5G core network, enabling the implementation of relevant communication functionalities.

The interface between the SMF (Session Management Function) and UPF (User Plane Function) network elements is designated as the N4 interface. At the application layer, the N4 interface employs the PFCP (Packet Forwarding Control Protocol). The primary functions of PFCP include data forwarding control, session management, separation of control

and user planes, QoS management, and others. Within the PFCP protocol stack, UDP serves as the transport layer protocol, while PFCP operates at the application layer. The PFCP protocol encompasses various signaling procedures, such as heartbeat procedures and session management procedures.

Following the establishment of the aforementioned simulated 5G core network environment, normal PFCP signaling traffic datasets were collected by capturing PFCP traffic on the N4 interface while executing normal activities in the UE access core network, including UE registration, PDU SESSION establishment, PDU SESSION termination, and UE deregistration. Through traffic parsing, aggregation, and data normalization processing, data with a length of 50 and a dimension of 4 can be obtained. The final dataset sizes available for training and testing are shown in Table 1.

Table 1 The Size of dataset

Type	Size
Normal	18431
PFCP session establishment DoS attack	7701
PFCP session deletion DoS attack	3432
PFCP session modification flooding attack	5375
All abnormal PFCP	16508

4.2 Result And Analysis

We used 60% of the normal data to train the model. 10% of normal data and 10% of abnormal data were used for threshold determination, with the remaining normal and abnormal data serving as the test set. First, this paper conducted experiments on datasets combining each type of anomaly with the normal test set, yielding the results shown in Table 2.

Table 2 Performance of anomaly detection methods

Data Type	Accuracy	Precision	Recall	F1-score
PFCP session establishment DoS attack	0.9930	0.9981	09893	0.9937
PFCP session deletion DoS attack	0.9937	0.9974	0.9851	0.9912
PFCP session modification flooding attack	0.9943	0.9985	0.9892	0.9939
All abnormal PFCP	0.9902	0.9981	0.9884	0.9932

To validate the advantages of the model, this paper selected three algorithms for comparison. The first applies Fourier transformation to the data, followed by clustering using the k-means[11] method. Training similarly used the normal dataset, resulting in multiple normal data cluster centers. When determining anomalies, the Euclidean distance between the data and the nearest cluster center is calculated; the greater the distance, the higher the degree of abnormality. The second replaces the Transformer neural network with an LSTM neural network, namely LSTM-AE[10], with data processing and training methods identical to those of the algorithm model in this paper. The third is a stacked encoder, with data processing and training methods identical to those of the algorithm model in this paper. The final results are shown in Table 3.

Table 3 Performance of anomaly detection methods

Algorithm	Accuracy	Precision	Recall	F1-score
K-means	0.9453	0.9755	09488	0.9619
LSTM-AE	0.9818	0.9949	0.9801	0.9875
SAE	0.9741	0.9885	0.9759	0.9821
Transformer AE	0.9902	0.9981	0.9884	0.9932

The experimental results indicate that during the detection process of each type of PFCP anomalous traffic, the accuracy consistently exceeded 99%, demonstrating overall excellent performance. In comparative analyses with baseline algorithms, the Transformer AE continued to exhibit superior detection capabilities. The K-means algorithm demonstrated the least effective detection performance, attributable to the limited capacity of machine learning models to effectively learn from sequential data. Conversely, the LSTM-AE algorithm, which maintained the autoencoder structure while replacing the Transformer network with LSTM neural networks, yielded relatively favorable outcomes. The results suggest that while LSTM retains considerable advantages in learning temporal data, Transformers possess

enhanced dependency capture capabilities. Consequently, autoencoder models implemented with Transformer architecture demonstrate superior performance compared to those implemented with LSTM neural networks.

5 CONCLUSION

This paper primarily investigates PFCP anomalous signaling detection and proposes an autoencoder algorithm model based on Transformer architecture. Through learning the bidirectional flow sequence characteristics of normal PFCP traffic, we have implemented an anomaly detection model. We constructed a 5G core network simulation environment using OAI and generated a rich corpus of normal data. The model was trained using the generated data and subsequently employed to detect anomalies in public PFCP intrusion datasets, achieving favorable accuracy rates. Comparative experiments were conducted to further validate the algorithmic advantages of our approach.

COMPETING INTERESTS

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

REFERENCES

[1] Kim, H. 5g core network security issues and attack classification from network protocol perspective. Internet Serv. Inf. Secur., 2020, 10(2): 1-15.
[2] Eric Ruzomberka, David J, Love, Christopher G, Brinton, et al. Challenges and Opportunities for Beyond-5G Wireless Security. IEEE Secur. Priv., 2023, 21(5): 55-66.
[3] Xinxin Hu, Caixia Liu, Shuxin Liu, et al. Signalling Security Analysis: Is HTTP/2 Secure in 5G Core Network?. 10th International Conference on Wireless Communications and Signal Processing, IEEE, 2018: 1-6.
[4] George Amponis, Panagiotis, Radoglou-Grammatikis, Thomas Lagkas, et al. Threatening the 5G core via PFCP DoS attacks: the case of blocking UAV communications. EURASIP J. Wirel. Commun. Netw., 2022, 2022(1): 124-150.
[5] Radivilova T, Kirichenko L, Lemeshko O, et al. Analysis of Anomaly Detection and Identification Methods in 5G Traffic. 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), IEEE, 2021: 1108-1113.
[6] Maimo, L F, Gomez, Gómez, Ángel Luis Perales, Clemente, Félix J, García, et, al. A self-adaptive deep learning-based system for anomaly detection in 5g networks. IEEE Access, 2018, 6, 7700-7712. DOI: 10.1109/ACCESS.2018.2803446.
[7] Harsh Sanjay Pacherkar, Guanhua Yan. PROV5GC: Hardening 5G Core Network Security with Attack Detection and Attribution Based on Provenance Graphs. Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks, ACM, 2024, 254-264.
[8] Robert Pell, Mohammad Shojafar, Sotiris Moschoyiannis. LSTM-Based Anomaly Detection of PFCP Signaling Attacks in 5G Networks. IEEE Consumer Electron. Mag., 2025, 14(1): 56-64.
[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you nee. arXiv preprint arxXiv: 1706.03762, 2017.
[10] Yingfei Xu, Yong Tang, Qiang Yang. Deep Learning for IoT Intrusion Detection based on LSTMs-AE[C]. AIAM2020: 2nd International Conference on Artificial Intelligence and Advanced Manufacture, ACM, 2020, 64-68.
[11] Gousiya Begum, S, Zahoor Ul Huq, A P, Siva Kumar. Fuzzy K-Means with M-KMP: a security framework in pyspark environment for intrusion detection. Multim. Tools Appl., 2024, 80(30): 73841-73863.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3010

AUTOMATED CYBERSECURITY INCIDENT RESPONSE: A REINFORCEMENT LEARNING APPROACH

MingJie Zhao, Rui Chen*
Beijing University of Posts and Telecommunications, Beijing 100000, China.
Corresponding Author: Rui Chen, Email: rchen272@bupt.edu.cn

Abstract: Cybersecurity incident response is critical for defending digital infrastructures from evolving cyber threats. Traditional manual systems and rule-based automation methods cannot efficiently cope with dynamic and sophisticated attacks. This paper explores the application of reinforcement learning (RL) to automate cybersecurity incident response. By modeling the response process as an RL problem, where an agent learns from interactions with the environment, the proposed system aims to enhance detection accuracy, minimize response times, and reduce false positives. Experimental results demonstrate the system's ability to mitigate threats effectively, showing that RL can significantly improve the efficiency and scalability of cybersecurity defenses. This approach leverages machine learning to automate decisions in real-time, adapting to evolving threats and optimizing incident response strategies. The integration of RL in incident response has the potential to dramatically reduce human error, improve system adaptability, and scale efficiently in complex, high-volume environments.
Keywords: Cybersecurity; Incident response; Reinforcement learning; Automation; Cyber threats; Machine learning; AI

1 INTRODUCTION

In the contemporary digital era, cybersecurity has become an urgent priority for individuals, businesses, and governments. As organizations increasingly depend on digital platforms to store sensitive information and conduct transactions, cyber threats have evolved to become more sophisticated, diverse, and dangerous. Ransomware, phishing attacks, and Distributed Denial-of-Service (DDoS) attacks are just a few examples of the growing cyberattack landscape that affects various sectors globally. These attacks, if left unchecked, can lead to severe financial loss, data breaches, reputational damage, and even national security threats[1].

Traditional cybersecurity methods, such as signature-based detection systems, firewalls, and antivirus software, have served as the foundation of digital security for decades. However, these methods are increasingly failing to keep up with the dynamic and evolving nature of modern cyberattacks. Signature-based systems, for instance, are effective only against known threats and are often inadequate for detecting new, unseen, or polymorphic attacks. Rule-based systems, though effective for predefined scenarios, lack the adaptability to address novel threats in real time[2]. Moreover, the increasing complexity of attacks requires more than just automated responses; it demands intelligent, adaptive decision-making systems that can optimize their response strategies as new threats emerge.

Incident response, which is a critical component of cybersecurity, involves identifying, analyzing, and mitigating security breaches as quickly and effectively as possible. Traditional incident response often depends on manual intervention, human judgment, and predefined response strategies. However, as cyber threats grow in volume and complexity, human intervention becomes slower, more error-prone, and increasingly inadequate to keep up with the pace of modern attacks. Furthermore, rule-based systems and predefined responses can be rigid and fail to detect or properly respond to new types of attacks. Thus, there is a growing need for automated, scalable, and real-time incident response systems that can handle the growing number of cybersecurity incidents without compromising performance.

RL, a subfield of machine learning, provides a promising solution to these challenges. Unlike traditional rule-based systems, RL enables systems to learn from experience, make decisions autonomously, and adapt to new threats over time. RL models are based on a feedback loop, where an agent interacts with an environment, takes actions, and receives feedback in the form of rewards or penalties. Over time, the agent learns to make optimal decisions that maximize cumulative rewards. In the context of cybersecurity incident response, this means that the RL agent can continually learn from past interactions, improving its decision-making and adapting to new types of cyber threats [3].

This paper explores the application of RL to automate cybersecurity incident response. We propose a system architecture that leverages RL for real-time threat detection and mitigation. Our system continuously learns from feedback and adapts to evolving attack patterns, offering a more dynamic and effective solution than traditional methods[4]. Through experiments and simulations, we demonstrate that RL-based incident response systems can outperform rule-based systems in key areas such as detection accuracy, response time, and false positive rates. This paper also discusses the advantages of RL in providing a scalable, adaptive, and efficient solution for modern cybersecurity challenges.

2 LITERATURE REVIEW

Cybersecurity is a critical domain that requires constant innovation due to the ever-evolving nature of cyber threats. Traditional approaches to cybersecurity largely rely on manual intervention or rule-based systems, which focus on predefined attack signatures or patterns. These systems work well for known threats but struggle when faced with new, previously unseen attack types[5]. Over the years, several machine learning techniques have been explored to address the limitations of traditional methods, and they have shown great promise in enhancing the efficiency and effectiveness of cybersecurity systems.

In particular, machine learning (ML) has become an essential tool in the cybersecurity field due to its ability to detect new, unknown attacks that may not fit predefined patterns. Various ML techniques, such as supervised learning, unsupervised learning, and deep learning, have been applied in areas like intrusion detection, malware analysis, and anomaly detection. Supervised learning, for example, requires labeled data for training, where the system is taught to identify specific features of known threats[6]. It has been successfully applied to malware classification and phishing detection, where the model learns from labeled examples of malicious and benign behavior.

However, supervised learning models face limitations, particularly when it comes to detecting unknown threats. Since these models rely on previously labeled examples, they are incapable of identifying new, unseen threats unless retrained with new labeled data. This dependency on labeled data makes supervised learning less practical in dynamic environments like cybersecurity, where new threats emerge continuously and require the system to adapt in real time. Unsupervised learning, on the other hand, does not require labeled data and can detect anomalies or deviations from normal behavior, making it useful for identifying novel attacks. However, unsupervised learning models often result in a higher rate of false positives, which can undermine the overall effectiveness of the system.

Deep learning, which is a subset of machine learning that uses artificial neural networks to learn from large amounts of data, has gained attention in the cybersecurity domain. Deep learning models have been applied to tasks such as intrusion detection and malware classification[7]. These models are capable of identifying complex patterns in large datasets and are particularly effective at handling unstructured data, such as network traffic or raw log files. However, deep learning models require substantial amounts of training data and computational power, which can be a limiting factor in resource-constrained environments.

RL, a form of machine learning in which an agent learns to make decisions by interacting with its environment, has emerged as a powerful tool in dynamic and sequential decision-making tasks. Unlike supervised and unsupervised learning, RL does not rely on labeled data or predefined patterns. Instead, the agent learns by taking actions and receiving feedback in the form of rewards or penalties[8]. This makes RL particularly well-suited for environments like cybersecurity, where attack patterns are constantly evolving, and predefined rules or signatures are insufficient.

RL has been applied to various cybersecurity tasks, such as intrusion detection, network defense, and malware analysis. In intrusion detection systems (IDS), RL agents have been trained to classify network traffic as benign or malicious based on patterns learned from past interactions. Similarly, RL has been used to optimize strategies for network defense, where the agent learns to block malicious traffic, isolate infected systems, or adjust security configurations in response to ongoing attacks. In malware analysis, RL has been applied to identify malicious behavior by observing how software interacts with the system and learning to distinguish between benign and malicious activities.

Despite the promising results in these applications, the full-scale application of RL to automate complete incident response systems—encompassing detection, decision-making, and mitigation—remains an area of active research. While RL offers the potential to significantly improve the efficiency and adaptability of cybersecurity systems, there are several challenges that must be addressed before RL can be widely implemented in real-world cybersecurity environments. One of the primary challenges is the need for high-quality training data. RL agents rely on feedback from interactions with the environment to learn optimal decision-making strategies. If the environment is not accurately modeled or the data is biased, the RL agent may learn suboptimal or even harmful strategies.

Another challenge is the computational complexity of RL models. Training RL agents requires significant computational resources, particularly when working with large-scale datasets or complex environments. For instance, training a deep RL agent to handle real-time cybersecurity incidents requires massive computational power, which can be a limiting factor in environments where resources are constrained. Additionally, the time it takes to train an RL model can be a barrier, as cyberattacks evolve rapidly and require systems to adapt almost instantaneously.

Furthermore, RL-based systems are susceptible to adversarial attacks. Adversarial attacks aim to manipulate the learning process of RL agents by introducing misleading or malicious feedback[9]. These attacks can potentially compromise the security of the system, allowing attackers to bypass detection or manipulate the system's decision-making process. To address this vulnerability, it is essential to develop techniques to secure RL systems against adversarial manipulation and ensure their robustness in the face of such threats.

Despite these challenges, the application of RL in cybersecurity incident response has the potential to revolutionize the field by enabling systems to continuously learn, adapt, and improve over time. By automating decision-making and optimizing responses to emerging threats, RL-based systems can reduce response times, improve detection accuracy, and scale more effectively to handle large volumes of cybersecurity events. As RL research continues to advance, it is expected that RL will play an increasingly important role in the future of automated cybersecurity.

3 METHODOLOGY

3.1 System Architecture

The proposed system architecture for the RL-based automated incident response is designed to address the dynamic and evolving nature of cybersecurity threats. Traditional rule-based systems, while effective against known threats, lack the flexibility and adaptability to handle novel or complex attacks. This is where RL offers significant advantages by providing the system with the ability to learn from experience and adapt its responses to new and unforeseen threats. The architecture of the RL-based incident response system is composed of several key components, which are crucial for the efficient and accurate identification, analysis, and mitigation of cybersecurity incidents.

The first component of the system is data collection, which is critical for providing the RL agent with the information it needs to make informed decisions. This data is gathered from multiple sources, such as network traffic, system logs, and external threat intelligence feeds. Network traffic data includes information on packet flow, source/destination IP addresses, port numbers, and timestamps, which can provide crucial indicators of potential cyber threats. System logs capture detailed information about system events, user actions, and application behavior, which can also help identify suspicious activities that may indicate an ongoing attack. External threat intelligence feeds provide valuable information on known attack patterns, emerging threats, and vulnerabilities, which can further assist in identifying potential risks.

Once the data is collected, it is preprocessed and analyzed to extract relevant features that can provide insights into the state of the system and potential threats. Preprocessing involves filtering out noise from the data, normalizing the data, and identifying patterns or anomalies that are indicative of malicious activities. This is often achieved using machine learning algorithms such as clustering or anomaly detection techniques, which are designed to recognize deviations from normal behavior. For example, if a particular IP address is sending an unusually high volume of requests to a server, the system may flag this as a potential denial-of-service (DoS) attack. Similarly, if there is suspicious behavior or unauthorized access attempts detected in system logs, it may indicate a potential breach.

After the data has been processed and potential threats have been identified, the next step involves the RL agent, which is the heart of the automated incident response system. The RL agent is responsible for analyzing the current state of the system and deciding on the most appropriate action to mitigate the detected threat. The decision-making process is based on an evaluation of the current system status, which includes factors such as the nature and severity of the threat, the potential impact on the system, and the available resources to respond to the attack. The RL agent learns to take optimal actions through interactions with the environment, where it continuously improves its decision-making based on feedback received from past actions.

In an RL framework, the agent makes decisions through a process known as policy learning. The agent's policy is a strategy that determines the best action to take in each state. The policy is refined over time as the agent interacts with the environment and receives feedback. The feedback comes in the form of rewards and penalties, which help the agent learn which actions are effective in mitigating threats and which are not. For example, if the RL agent successfully blocks a malicious attack, it receives a positive reward, whereas if it fails to detect or mitigate the attack in time, it receives a penalty. This process of trial and error allows the agent to learn from its mistakes and continuously improve its policy.

Once the RL agent determines the optimal action to take, the system moves to the final stage, which is response orchestration. This component is responsible for executing the mitigation actions decided by the RL agent in real-time. The system can carry out various actions, such as blocking malicious network traffic, isolating infected systems from the network, or alerting administrators for manual intervention if necessary. For example, if the RL agent detects a ransomware attack, the system might isolate the infected host and block any outgoing traffic to prevent further spread of the malware. In the case of a DDoS attack, the system may reroute traffic or block malicious IP addresses to minimize the impact of the attack.

The entire architecture is designed to operate in real-time, ensuring that the system can respond to cybersecurity threats as quickly as possible. The RL agent is continuously learning from new data and adapting to emerging threats, allowing the system to handle a wide range of attack scenarios. Moreover, the system is highly scalable, enabling it to manage large volumes of security events without compromising performance as in Figure 1.

Figure 1 System Architecture of RL-Based Automated Incident Response

3.2 Reinforcement Learning Model Design

The RL model used in the system is based on the Markov Decision Process (MDP) framework, which provides a mathematical model for decision-making under uncertainty. In this framework, the state represents the current conditions of the system, such as active threats and network activity. The actions correspond to the possible responses that the RL agent can take, such as blocking traffic or isolating a system. The reward function provides feedback on the effectiveness of the actions taken. Positive rewards are given for successful mitigation, while penalties are assigned for failures.

To train the RL model, we use Q-learning, a model-free reinforcement learning algorithm. Q-learning allows the agent to update its decision-making policy based on the rewards or penalties it receives from the environment. The agent learns to maximize cumulative rewards by selecting optimal actions in each state.

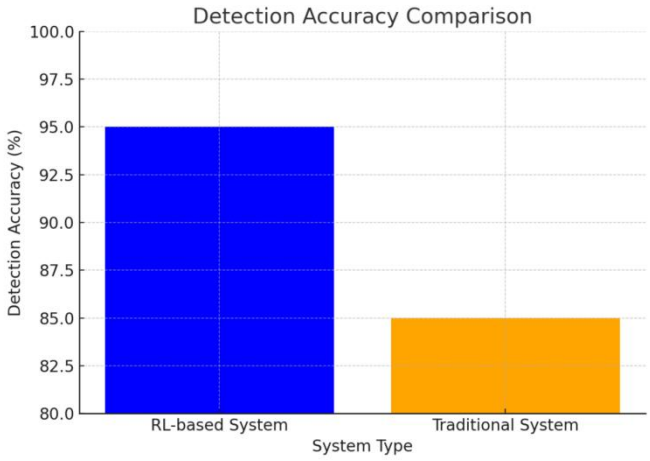
Q-learning involves updating Q-values, which represent the expected cumulative reward for taking an action in a given state. The agent refines its policy over time by adjusting these Q-values. The goal of the RL model is to continuously improve its decision-making and respond to new threats in real-time, providing an adaptive solution for automated incident response.

4 RESULTS AND DISCUSSION

4.1 Experimental Results

The RL-based system was tested in a simulated environment and compared with traditional rule-based systems. The results indicated that the RL-based system outperformed traditional methods. Specifically, the RL system detected 95% of threats, while traditional systems detected only 85%. Additionally, the RL system responded 30% faster than rule-based systems, and it had a false positive rate of 3%, compared to 8% for traditional systems, as in Figure 2.

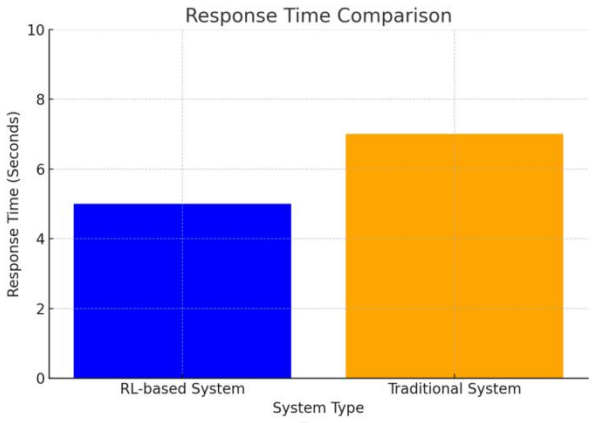
Figure 2 Detection Accuracy Comparison



4.2 Analysis of Results

The RL-based system demonstrated the ability to learn from feedback and adapt to new threats. It consistently improved its decision-making, reducing false positives and optimizing response times over time as in Figure 3.

Figure 3 Response Time Comparison



4.3 Limitations and Challenges

Despite the positive results, several challenges remain. One significant challenge is data dependency. The quality of the training data is crucial for the system's effectiveness, as limited or biased data can affect the RL agent's learning process. Computational complexity is another challenge, as RL models require substantial computational resources, especially for training, which can be a barrier to large-scale deployment. Finally, the RL system may be vulnerable to adversarial attacks that manipulate its learning process.

5 CONCLUSION

This study demonstrates that RL can be highly effective in automating cybersecurity incident response. Traditional rule-based systems, which rely on predefined rules and signatures, often fail to keep up with dynamic and sophisticated cyber threats. In contrast, the RL-based system used in this study showed substantial improvements across key metrics, such as detection accuracy, response time, and false positive rate. Specifically, the RL system achieved a higher detection rate by accurately identifying a wider range of both known and novel threats. Additionally, it outperformed traditional methods in response time, reducing the time needed to mitigate attacks. Another key advantage of the RL system is its ability to reduce false positives, ensuring that security resources are not wasted on false alarms.

The RL agent continuously learns from its feedback, allowing it to adapt to new and evolving cyber threats. This adaptability makes the RL-based system an efficient and scalable solution for real-time incident response. Looking ahead, future research should focus on improving the scalability of RL-based systems to handle large-scale environments. Further work is also needed to enhance the robustness of these models against adversarial attacks that could exploit vulnerabilities in the learning process. Additionally, integrating RL with other AI techniques, such as deep learning, could further improve threat classification and mitigation capabilities.

COMPETING INTERESTS

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

REFERENCES

[1] Alturkistani H, El-Affendi M A. Optimizing cybersecurity incident response decisions using deep reinforcement learning. International Journal of Electrical and Computer Engineering, 2022, 12(6): 6768.

[2] Dunsin D, Ghanem M C, Ouazzane K, et al. Reinforcement learning for an efficient and effective malware investigation during cyber Incident response. arXiv preprint arXiv:2408.01999, 2024.

[3] Ren S, Jin J, Niu G, Liu Y. ARCS: Adaptive Reinforcement Learning Framework for Automated Cybersecurity Incident Response Strategy Optimization. Applied Sciences, 2025, 15(2): 951.

[4] Naseer A, Naseer H, Ahmad A, et al. Moving towards agile cybersecurity incident response: A case study exploring the enabling role of big data analytics-embedded dynamic capabilities. Computers & Security, 2023, 135: 103525.

[5] Manda J K. Cybersecurity Automation in Telecom: Implementing Automation Tools and Technologies to Enhance Cybersecurity Incident Response and Threat Detection in Telecom Operations. Advances in Computer Sciences, 2021, 4(1).

[6] Hassan S K, Ibrahim A. The role of artificial intelligence in cyber security and incident response. International Journal for Electronic Crime Investigation, 2023, 7(2).

[7] Lee Z, Wu Y C, Wang X. Automated Machine Learning in Waste Classification: A Revolutionary Approach to Efficiency and Accuracy. In Proceedings of the 2023 12th International Conference on Computing and Pattern Recognition, 2023: 299-303.

[8] Alturkistani H, El-Affendi M A. Optimizing cybersecurity incident response decisions using deep reinforcement learning. International Journal of Electrical and Computer Engineering, 2022, 12(6): 6768.

[9] Li X, Wang X, Chen X, et al. Unlabeled data selection for active learning in image classification. Scientific Reports, 2024, 14(1): 424.

[10] Liang Y, Wang X, Wu Y C, et al. A study on blockchain sandwich attack strategies based on mechanism design game theory. Electronics, 2023, 12(21): 4417.

[11] Schlette D, Caselli M, Pernul G. A comparative study on cyber threat intelligence: The security incident response perspective. IEEE Communications Surveys & Tutorials, 2021, 23(4): 2525-2556.

[12] Mouratidis H, Islam S, Santos-Olmo A, et al. Modelling language for cyber security incident handling for critical infrastructures. Computers & Security, 2023, 128: 103139.

[13] Oriola O, Adeyemo A B, Papadaki M, et al. A collaborative approach for national cybersecurity incident management. Information & Computer Security, 2021, 29(3): 457-484.

[14] He Y, Zamani E D, Lloyd S, et al. Agile incident response (AIR): Improving the incident response process in healthcare. International Journal of Information Management, 2022, 62: 102435.

[15] Liu Y, Wu Y C, Fu H, et al. Digital intervention in improving the outcomes of mental health among LGBTQ+ youth: a systematic review. Frontiers in psychology, 2023, 14: 1242928.

[16] Wang X, Wu Y C, Ma Z. Blockchain in the courtroom: exploring its evidentiary significance and procedural implications in US judicial processes. Frontiers in Blockchain, 2024, 7: 1306058.
[17] Wang X, Wu Y C, Zhou M, et al. Beyond surveillance: privacy, ethics, and regulations in face recognition technology. Frontiers in big data, 2024, 7: 1337465.
[18] Guo H, Ma Z, Chen X, et al. Generating artistic portraits from face photos with feature disentanglement and reconstruction. Electronics, 2024, 13(5), 955.
[19] Andrade R O, Cordova D, Ortiz-Garcés I, et al. A comprehensive study about cybersecurity incident response capabilities in Ecuador. In Innovation and Research: A Driving Force for Socio-Econo-Technological Development 1st. Springer International Publishing, 2021: 281-29.
[20] Fauziyah F, Wang Z, Joy G. Knowledge Management Strategy for Handling Cyber Attacks in E-Commerce with Computer Security Incident Response Team (CSIRT). Journal of Information Security, 2022, 13(4): 294-311.
[21] Ahmad A, Maynard S B, Desouza K C, et al. How can organizations develop situation awareness for incident response: A case study of management practice. Computers & Security, 2021, 101: 102122.
[22] van der Kleij R, Schraagen J M, Cadet B, et al. Developing decision support for cybersecurity threat and incident managers. Computers & Security, 2022, 113: 102535.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3012

LEGAL ATTRIBUTES OF COMMERCIAL DATA AND APPROACHES TO INTELLECTUAL PROPERTY PROTECTION

Zhaoxia Deng
Guangdong University of Foreign Studies, Guangzhou 510515, Guangdong, China.
Corresponding Email: Angelina_dzx@163.com

Abstract: As a core production factor in the digital economy, the legal attributes and protection pathways of commercial data have become a research hotspot in the field of legal studies. This paper begins by examining the legal nature of commercial data and its attributes under intellectual property rights, establishing the legitimacy foundation for intellectual property protection of commercial data through labor property theory, utilitarian theory, and transaction cost theory. Besides, it proposes a "limited rights + fair use" governance model to address current legal dilemmas in data rights allocation. The study further recommends approaches such as strengthening technical governance and collaborative regulation, promoting international rule harmonization, to balance innovation incentives with public interests. Ultimately, these measures aim to achieve efficient circulation and value realization of commercial data.
Keywords: Commercial data; Legal attributes; Intellectual property protection; Fair use; Technical governance

1 INTRODUCTION

In the era of digital economy, data, as a new type of production factor, has become the core driving force for social transformation and economic growth. Commercial data, due to its unique economic value and strategic significance, plays an irreplaceable role in industrial innovation, market competition, and social governance. However, as the value of data resources has increasingly emerged, issues such as vague legal attributes, disputes over rights ownership, and divergences in protection paths have gradually come to the fore. The traditional legal framework faces structural challenges in addressing data rights confirmation, circulation, and benefit distribution: on the one hand, characteristics such as the non-exclusivity and replicability of data make it difficult to fully fit into existing rights systems such as property rights or intellectual property rights; on the other hand, excessive protection may inhibit data sharing and technological innovation, while insufficient protection is prone to triggering "tragedy of the commons" or "free-riding" behaviors, leading to market failures. Against this backdrop, clarifying the legal attributes of commercial data and exploring an adaptive intellectual property protection approach has become an urgent proposition to be solved in legal research. Starting from the legal attributes of commercial data, this paper systematically demonstrates the justificatory foundation for its intellectual property protection by integrating theories of legal philosophy and law and economics. Aiming at the dilemmas of "over-generalization of rights" and "circulation blockages" in current empowerment practices, it proposes corresponding countermeasures and suggestions, with the aim of providing theoretical references for constructing a data governance framework that balances innovation incentives and public interests.

2 ANALYSIS OF THE LEGAL ATTRIBUTES OF COMMERCIAL DATA

As a new type of production factor in the digital economy era, commercial data exhibits unique attributes in physical, economic, and legal dimensions. At the physical level, data exists in the form of binary code and has three technical characteristics: non-exclusivity (the same data can be held by multiple subjects simultaneously), replicability (marginal cost approaches zero), and non-consumability (use does not lead to value depreciation). At the economic level, data resources demonstrate dynamic value (value-added effects through processing), application scenario dependency (value varies with the application environment), and compound rights and interests (interweaving interests of multiple subjects). At the legal level, there are normative challenges such as the virtualization of object form (lack of physical boundaries of traditional property rights objects), ambiguity in rights ownership (conflicts between the rights and interests of data producers, processors, and users), and hierarchical protection needs (differences in protection intensity between raw data and derivative data)[1].

There is a theoretical confrontation in current academic circles regarding whether commercial data can constitute the object of intellectual property rights. From the perspective of traditional theoretical frameworks, opponents put forward three main objections: first, the lack of an originality requirement, as raw data, as a record of objective facts, does not meet the originality standard of copyright law; second, the impossibility of exclusive control, as the technical characteristics of data circulation inherently conflict with the exclusivity of intellectual property rights; third, a misalignment in the protection system, arguing that existing systems such as trade secrets and anti-unfair competition laws are sufficient to regulate data rights and interests[2]. Proponents, however, argue that derivative data products, formed through algorithmic processing, contain original content, and their nature as intellectual achievements is homogeneous with the objects protected by intellectual property rights. Based on the practices of digital economic development, they emphasize three aspects of protectability for data products: derivative data reconstructed through algorithms possess the attributes of intellectual achievements (such as user behavior analysis models); databases formed

through processing massive data meet the requirements for compiled works (such as product information databases on e-commerce platforms); and commercial intelligence formed by specific data combinations conforms to the criteria for trade secrets (such as precision marketing databases)[3].

This paper argues that commercial data refers to a collection of derivative data with market value, formed through large-scale extraction and processing of raw data via intellectual labor, and filtered and intelligently pushed for sales through algorithmic technology. Raw data, lacking creative input and independent economic value, is not suitable for direct rights assignment; by contrast, derivative data products, through human intellectual activities, form new information configurations that meet the core requirements for intellectual property protection.

3 POSITIONING THE ATTRIBUTE OF INTELLECTUAL PROPERTY RIGHTS FOR COMMERCIAL DATA

Although derivative data can become the object of intellectual property protection, the intellectual property rights in commercial data still differ significantly from traditional intellectual property rights. This section discusses the differences mainly from three aspects: the object of rights, the content, and the effect of rights.

3.1 The Dynamic Nature of Rights Objects

The objects of traditional intellectual property rights exhibit distinct static characteristics. For example, the expressions protected by copyright law—works—must have a fixed form at the time of creation (such as literary, musical, or artistic works), and their rights boundaries are defined directly by the author's creative act. Technical solutions, as objects protected by patent law, must be disclosed in a specification and form definite claims, with the scope of rights limited to the authorized text and subject to strict restrictions on subsequent modifications. For trademarks, the distinctiveness of the mark must comply with legal requirements at the time of registration, and the scope of rights is limited to the approved mark and the goods or services for which it is registered.

In contrast to the objects of traditional intellectual property rights, commercial data, as the object of data intellectual property rights, demonstrates significant dynamic characteristics. First, data aggregation is fluid. Derivative data products process massive raw data in real time through algorithms, and their content evolves dynamically as data sources update and models optimize[4]. For example, a traffic flow prediction model must continuously incorporate real-time traffic data to maintain accuracy, causing the boundary of the rights object to change continuously. Second, the value generation of data products is iterative. The value of data products is not solidified in a single instance; rather, it is formed through multiple rounds of refinement to create a "value-added data chain". For instance, user profiles evolve from basic labels (such as age and gender) to in-depth insights (such as consumption preferences and risk predictions) through multiple algorithmic iterations, with the rights object increasing in value dynamically. In addition, the application scenarios of data products are uncertain. Since the functions of data products are deeply tied to their application scenarios, they may lose economic value when detached from specific contexts. For example, health code data used during pandemic prevention and control may transform into data assets for urban governance research after the pandemic ends, marking a qualitative change in the rights object as the scenario shifts.

3.2 The Composite Nature of Rights Content

Traditional intellectual property rights exhibit linear characteristics in their content. For instance， copyright centers on the "right of reproduction,"extending to derivative rights such as distribution and communication through information networks, forming a bundle of rights focused on controlling utilization methods. Patent rights revolve around the "exclusive right to implement," encompassing specific entitlements like manufacturing, using, and selling, with clear and closed boundaries. Trademark rights prioritize the "exclusive right to use," supplemented by protections against consumer confusion and cross-class safeguards for well-known trademarks, all confined by designated product categories.

By contrast, the content of intellectual property rights in commercial data exhibits networked characteristics, specifically manifested as follows: (1) the multidimensional nature of affirmative rights. For example, based on the right to control, data processors have the authority to store and access data products (such as the classified and graded data management provisions in the *Data Security Law*); based on the right to develop, data processors may enhance data quality through algorithm optimization (such as the original labor involved in data cleaning and labeling); based on the right to benefit, relevant right holders are entitled to economic returns from data product transactions (such as licensing revenue-sharing mechanisms in data exchanges); (2) the defensive nature of negative rights. Data owners possess immunity to resist improper intervention by public authorities (such as commercial data exemption clauses in government data openness); they also hold the right to defense, allowing them to claim "fair use" in data scraping disputes (such as the "non-material substitution" defense in the *Dianping v. Baidu* case); (3) the interactive nature of interdependent rights. Interactivity is reflected in the coordination of conflicts between data intellectual property rights and individual information rights, such as the separation of data ownership and rights after anonymization processing; it also manifests in the convergence between data intellectual property rights and antitrust law, such as the determination of market dominance caused by data concentration [5].

3.3 Limitations on the Effect of Rights

Traditional intellectual property rights have strong exclusivity in their effect. For example, under copyright law, reproducing or distributing others' works without permission constitutes infringement, with exceptions strictly limited to statutory conditions—for instance, fair use must satisfy the "three-step test." Patent law grants right holders the authority to prohibit others from implementing identical or similar patented technologies, and independent development does not exempt one from infringement liability. Trademark law prohibits confusing use, with territorial protection strictly defining the boundaries of rights.

Different from traditional intellectual property rights, the effect of intellectual property rights in commercial data exhibits limited exclusivity. First, the scope of commercial data intellectual property rights is narrowed. Such protection excludes raw data and public data from rights assignment—for example, government-open data is not subject to data intellectual property rights. Additionally, the protection of derivative data products must meet an "innovative" threshold; the EU *Data Act*, for instance, requires data products to have "significant added value." Second, the exercise of commercial data intellectual property rights is restricted. During public crises (e.g., pandemic prevention and control) or for necessary technological innovation (e.g., AI training data), administrative authorities may compulsorily open commercial data usage rights; research institutions may use data for non-commercial purposes without authorization, as seen in the open-sharing principles of the *Regulations on Scientific Data Management*, which expand the concept of fair use in copyright law [6]. Furthermore, there are triggers for the invalidation of commercial data intellectual property rights. When data products lose economic value due to technological iteration, their rights automatically terminate—for example, outdated navigation data is no longer protected. When data subjects withdraw their consent, the rights to products derived from such data also become invalid.

In summary, the intellectual property rights in commercial data represent an extension of the intellectual property system in the digital era. In essence, they constitute a new type of information property rights, possessing both private and public attributes, and requiring institutional innovation to achieve the dual goals of incentivizing innovation and promoting circulation. Given that data continuously aggregates and evolves during its flow, the boundaries of derivative data products are uncertain, making the static object paradigm of traditional intellectual property rights difficult to adapt to the dynamic characteristics of data. It is necessary to respond to the needs of data mobility through flexible rights allocation, such as dynamic registration systems and scenario-based rights definition. Furthermore, the intellectual property rights in commercial data include not only affirmative rights (such as the right to use and license) but also negative defensive rights (such as immunity). The traditional "bundle of rights" model of intellectual property is insufficient to accommodate the complexity of data rights, necessitating the introduction of rights network theory to achieve dynamic balance of interests among multiple subjects. Finally, these rights must break through the "absolute control" logic of traditional intellectual property and adopt mechanisms of rights weakening and fair use to achieve Pareto optimality between data mobility and innovation incentives.

4 JUSTIFICATIONS FOR THE INTELLECTUAL PROPERTY PROTECTION OF COMMERCIAL DATA

4.1 Labor Theory of Property: Labor Input in Data Value

Locke's labor theory of property provides a moral-philosophical justificatory foundation for data rights assignment. The core proposition of this theory is that "labor removes natural objects from the common state and grants laborers exclusive rights." In the data domain, data processing labor exhibits significant heterogeneity, which can be divided into foundational labor and creative labor. Foundational labor includes low-skill repetitive tasks such as data collection, cleaning, and annotation. For example, data annotators classifying and labeling images, while time-consuming, lack originality; their labor value is mainly reflected in the "usability" of data resources. Creative labor, by contrast, involves high-intellectual activities such as algorithm design, model training, and knowledge discovery. For instance, using deep learning algorithms to extract disease diagnostic features from medical images—such labor creates new increments of knowledge, aligning with the core premise of Locke's theory that "labor creates value" [7].

Locke's labor theory of property provides a moral basis for data rights assignment. Data processors transform raw data into economically valuable derivative products through labor inputs such as data cleaning and algorithmic modeling. For example, e-commerce platforms generating consumer preference reports through user behavior data analysis—this process incorporates technical, financial, and intellectual labor, conforming to the justificatory logic of "labor creates property rights."

4.2 Utilitarian Theory: Incentivizing Innovation and Enhancing Efficiency

Bentham's utilitarian principle, which advocates for "the greatest happiness for the greatest number," manifests in data rights assignment as a need to balance innovation incentives and public welfare. First, granting data processors exclusive rights can motivate enterprises to increase investment in data R&D and promote data-driven technological innovation. However, excessive reinforcement of exclusive rights may also stifle subsequent innovation [8]. According to the patent thicket effect, if strong exclusivity is granted to data sets, downstream developers must obtain licenses at each layer, leading to a surge in transaction costs. For example, the "data silo" phenomenon in the medical data field forces AI drug development companies to spend years negotiating data usage rights. Additionally, over-strengthening exclusive rights carries the risk of inhibiting innovation—the EU *Database Directive*, which granted database creators a 15-year "sui generis" right, was criticized for "hindering data reuse," ultimately prompting the EU to shift toward an "open data strategy."

Thus, utilitarianism requires that rights allocation reserve institutional space for "maximizing overall social utility." Compulsory licensing mechanisms and data openness dividends reflect the balance between innovation incentives and public interests. When the marginal social benefit of data use exceeds the marginal cost of exclusive rights, the boundaries of rights protection should be transcended to grant the public certain open access rights [9].

4.3 Transaction Cost Theory: Internalization of Externalities

Coase's theorem posits that clear property rights definitions can reduce transaction costs. Applying this theorem requires reconstructing the efficiency boundaries of rights allocation in light of the unique characteristics of data as a factor of production. Data transaction costs include search costs, verification costs, and enforcement costs. Data transactions may also generate dual effects of positive and negative externalities: the former is exemplified by the open-source data platform Kaggle, which promotes global AI competition innovation through shared datasets and has incubated over 200,000 machine learning models; the latter, such as unanonymized geographic location data potentially used for crime prediction, triggering privacy disputes [1].

However, the prevalent information asymmetry and "Arrow's paradox" (disclosure leads to loss of value) in current data transactions cause market failures. Therefore, data intellectual property rights reduce uncertainties in data circulation by clarifying rights boundaries, internalizing economic externalities. For example, commercial data processors, assured of rights protection, are more willing to participate in data transactions, thereby promoting the optimal allocation of data resources.

5 APPROACHES TO THE INTELLECTUAL PROPERTY PROTECTION OF COMMERCIAL DATA

Despite the justificatory foundation for protecting data intellectual property rights, significant challenges remain. First is the ambiguity of rights boundaries. Pntellectual property protection for commercial data currently face dual risks of "overprotection" and "underprotection" : on the one hand, excessive emphasis on exclusivity may hinder data sharing and create "data silos"; on the other hand, unclear protection standards for derivative data in existing systems lead to heavy reliance on general clauses of the *Anti-Unfair Competition Law* in judicial practice, resulting in inconsistent adjudication standards[2]. Second is the complexity of interest conflicts. Data rights involve the interests of multiple stakeholders, posing significant challenges to balance: conflicts exist between private rights and public welfare, as data monopolies may threaten national security and public interests [10]; conflicts also arise between domestic laws and international rules, as divergent data governance models across countries impede cross-border data flows. Finally, there is the impact of technological iteration. Technologies such as blockchain and artificial intelligence (AI) challenge traditional intellectual property systems—for example, distributed ledger technology weakens the effectiveness of centralized registration systems, and existing regulations lack clear provisions on the originality of AI-generated data. In response to these legal challenges, this paper attempts to explore the intellectual property protection approaches for commercial data from the following aspects.

5.1 Building a Governance Model of "Limited Rights + Fair Use"

Given the significant challenges in protecting commercial data intellectual property rights, a governance model of "limited rights + fair use" can be constructed based on the rights attributes of commercial data intellectual property. The data industry chain involves multiple stakeholders: data contributors (individuals/enterprises), data processors (algorithm developers), and data users (consumers/third parties). Based on the theory of interest balancing, rights allocation should follow the "contribution principle," granting right holders limited rights [11].

On the one hand, in line with the value-added law of the data value chain, a hierarchical protection system of "raw data—basic data products—value-added data products" can be established: (1) Raw data (e.g., original temperature records collected by sensors), lacking creative labor input, should be regulated by laws such as the *Data Security Law* and *Personal Information Protection Law*; (2) Basic data products (e.g., cleaned structured datasets), although involving preliminary processing, have limited creativity and can be protected by property rights rules to safeguard the legitimate rights of data controllers; (3) Value-added data products (e.g., market demand prediction models formed through machine learning), which embody substantial intellectual creation and meet the dual standards of "original expression" and "manifest commercial value," should be included in the scope of intellectual property objects. This classification not only aligns with the "three-rights separation reform" direction proposed in the *Twenty Guidelines on Data* (i.e., separating data resource ownership, data processing and use rights, and data product operation rights) but also balances the dual needs of data circulation and rights protection.

Second, the limitation of rights can also be implemented through the protection term of data intellectual property rights. In general, the protection term for data intellectual property rights should be shorter than that of traditional intellectual property to promote data circulation. This paper proposes adopting a "statutory minimum protection period + dynamic adjustment" mechanism, namely: a basic protection period of 3–5 years for data intellectual property rights to ensure data processors recoup their innovation costs; for derivative data that is continuously updated and has significant social value, an extension may be applied for, with a maximum period not exceeding 10 years.

Additionally, fair use mechanisms or compulsory licensing systems should be applied to limit data intellectual property rights. For example, exceptions such as non-commercial use and scientific research can be established to ensure the

realization of data's public attributes. When public interests require it, administrative authorities may grant compulsory data usage licenses but must provide reasonable compensation.

5.2 Strengthening Technological Governance and Collaborative Regulation

First, the immutability feature of blockchain technology provides a new technical path for data rights attribution. By constructing a distributed ledger architecture, full-lifecycle behaviors such as data production, circulation, and processing are fixed in the form of hash values, forming a tripartite evidence chain that includes timestamps, operating subjects, and data fingerprints. This mechanism relies on asymmetric encryption algorithms (such as RSA, Elliptic Curve Cryptography) to digitally sign and authenticate data operation behaviors, ensuring that each data interaction can be traced and verified through the chain structure. In judicial practice, blockchain evidence storage that complies with the *Provisions of the Supreme People's Court on Several Issues Concerning the Trial of Cases by Internet Courts* has the effect of electronic evidence, effectively solving the problem of reversed burden of proof in data rights disputes. However, it should be noted that there is an inherent conflict between the storage efficiency of public blockchains and privacy protection, which requires realizing permission separation between regulatory nodes and business nodes through an alliance chain architecture [12].

Moreover, a collaborative regulatory system of "government-platform-third-party certification institutions" can be constructed. First, enable the docking of cross-platform data regulatory interfaces through interoperability standards. Second, use regulatory sandbox mechanisms to conduct risk stress tests on new data technologies. Third, introduce third-party audit institutions to carry out interpretability certification for algorithmic models.

5.3 Promoting International Rule Coordination

The international community currently faces a dilemma of "Regime Competition" in the protection of commercial data intellectual property rights: the WIPO *Treaty on the Protection of Databases* has been long dormant due to disagreements among member states; the TRIPS Agreement does not clarify the protection boundaries for data products; and while Chapter 18 of the CPTPP establishes the principle of free data flow, it lacks specific protection standards. To resolve this impasse, a "functionalism-oriented" governance framework should be constructed. In the first tier, relying on the WTO e-commerce negotiation mechanism to establish "technology-neutral" standards for determining the originality of data products. In the second tier, through the RCEP Agreement, establish non-discriminatory protection principles for derivative data products, and set up a dispute settlement panel for data intellectual property within the RCEP framework to build a specialized arbitration mechanism that includes a technical investigator system. In the third tier**,** exploring the convergence between domestic laws and international rules. Leverage the Guangdong-Hong Kong-Macao Greater Bay Area's cross-border data flow pilots and the RCEP Agreement to construct a unified protection framework for RCEP commercial data intellectual property rights, achieving collaborative governance of commercial data intellectual property within the RCEP region.

Resolving the dilemma of international rule coordination requires achieving two-way interaction between the "legalization of technical standards" and the "technicalization of legal rules." In the rights confirmation stage, under the RCEP framework, promoting the mutual recognition mechanism for data intellectual property registration procedures, incorporating elements such as blockchain evidence timestamps and data feature hash values into a unified certification system. In the rights exercise stage**,** translating the algorithmic interpretability requirements of GDPR Article 22 into enforceable technical standards to ensure that the use of data products complies with the transparency requirements of China's *Personal Information Protection Law* Article 24. In the rights protection stage, relying on the WIPO Lex database to establish a cross-border data intellectual property infringement characteristics library, using AI technology for real-time monitoring of cross-border infringement.

Additionally, the protection of commercial data intellectual property faces an inherent conflict between "legal territorialism" and "data borderlessness." Domestically, China needs to construct an "Embedded Compliance" framework. At the legislative level, through judicial interpretations of Article 127 of the *Civil Code* (data rights and interests clause), clarifying the "usufructuary rights of intellectual property" nature of derivative data products. At the judicial level**,** drawing inspiration from California's *Consumer Privacy Act* Section 1798.150 to establish a punitive compensation system for data intellectual property infringement (proposing a base amount of 1–3 times the actual losses of the right holder). At the enforcement level**,** relying on the National Data Administration to build a data intellectual property record-filing platform [2].

6 CONCLUSION

The legal attributes of commercial data and its intellectual property protection approaches need to break free from traditional paradigms and seek dynamic balance between incentivizing innovation and promoting circulation. By adopting a governance model of "limited rights + fair use," strengthening technological governance and collaborative regulation, and promoting international rule coordination, a data property rights system can be built that balances stability and adaptability. Future research should focus on the standardization of commercial data rights confirmation, judicial practices for rights relief, and the localized adaptation of international rules to provide a solid institutional foundation for the development of the digital economy.

COMPETING INTERESTS

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

REFERENCES

[1] Jia L P. Justification of Rights and Elaboration of Rules for Data Intellectual Property. Law and Social Development, 2024(4): 207–208, 213.
[2] Gao L. Dilemmas and Solutions in the Empowerment of Data Intellectual Property Rights. Jianghai Academic Journal, 2025(1): 185–186, 191.
[3] Liu X. Normative Approach and Institutional Framework for the Registration of Data Intellectual Property Rights. Journal of Huazhong University of Science and Technology (Social Sciences Edition), 2025(2): 92.
[4] Lv B B. An Intellectual Property Approach to the Establishment of Data Property Rights. Law and Business Review, 2025(1): 90.
[5] Rainer Kulms. Data Sharing and Data Protection. Romanian Review of Private Law, 2022(1): 151.
[6] Yanan Wang, Shanshan Cui. Deepening Digital Intellectual Property Rules and Digital International Patenting. Economics Letters, 2025, 251: 12.
[7] Zhang X B. On Data Property Rights as a New Type of Property Right. Social Sciences in China, 2023(4): 155–156.
[8] Feng X Q. Theoretical Explanation and Construction of Data Property Rights and Their Legal Regulation. Tribune of Political Science and Law, 2021(4): 83–84.
[9] Yang L H, Wang S T. Study on the Justification of Protecting Intellectual Property Rights in Commercial Data. Wuling Journal, 2024(1): 103–104.
[10] Wu H D. Empowerment of Data Property Rights: From Data Exclusive Rights to Data Use Rights. Law and Business Review, 2024(3): 10.
[11] Cui G B. Foundational Theory of Limited Exclusivity in Big Data. China Legal Science, 2019(5): 9.
[12] Eugenia Pedro, Joao Leitao, Helena Alves. Screening and Enhancing Intellectual Capital Consistency: A Scoping Review of Systematised Literature Reviews. Journal of Innovation & Knowledge, 2025, 10(2): 10.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3013

DATA-DRIVEN DECISION-MAKING SYSTEM FOR INJECTION MOLDING PRODUCTION AND MAINTENANCE

DuAng Chen, XinJie Zhou, Fan Xin, ShiWei Xu*
School of Artificial Intelligence, Wenzhou Polytechnic, Wenzhou 325035, Zhejiang, China.
Corresponding Author: ShiWei Xu, Email: swxu@wzpt.edu.cn

Abstract: A data-driven injection molding production and maintenance decision-making system is designed to address the issues of low efficiency and poor real-time performance in traditional data collection models, meeting the modern industrial needs for high reliability and intelligence. The system adopts a three-layer architecture, including data collection, edge computing, and maintenance decision-making layers. It achieves real-time collection and processing of multi-source heterogeneous data to assess equipment health status dynamically and predict failures. The data collection layer integrates sensor, PLC, and visual device data; the edge computing layer processes key parameters through lightweight models to reduce cloud-side pressure; the maintenance decision-making layer predicts the remaining life of the equipment using the Weibull distribution model and optimizes maintenance strategies. The system proposes a quantitative evaluation index for the health of the injection molding machine and utilizes a weighted fusion algorithm for accurate maintenance decisions, significantly reducing operational costs and improving production efficiency, providing a feasible technical solution for intelligent manufacturing.
Keywords: Injection molding production; Edge computing; Equipment health; Maintenance decision-making

1 INTRODUCTION

To achieve the goal of becoming a manufacturing powerhouse as part of "Made in China 2025", technologies such as Industry 4.0, Internet of Things (IoT), and intelligent manufacturing have rapidly developed. Data collection technology, as a bridge between the physical and digital worlds, provides foundational support for equipment condition monitoring, failure prediction, and maintenance decision-making. However, traditional data collection methods often face issues such as diverse device protocols, reliance on manual integration, low efficiency, high costs, and delayed responses, making it difficult to meet the high reliability and real-time demands of modern industries.

In this context, the operation and maintenance decision-making system, derived from data collection, has emerged. This system builds a closed-loop framework from data to decision-making through real-time collection, edge computing, and intelligent analysis of multi-source heterogeneous data, realizing dynamic assessment of equipment health, early failure prediction, and accurate maintenance strategy recommendations. For example, in an injection molding factory, by monitoring parameters such as vibration and temperature in real-time and combining predictive maintenance algorithms, failure handling can shift from "passive response" to "active prevention", significantly reducing operational costs. This contributes to the large-scale application of data in modern industrial settings and holds significant practical and applied value.

Current technologies primarily focus on overall production line forecasting, resulting in bulky systems with high on-site deployment costs. Many features remain impractical for factory applications, while the integration between real-time data and decision-making loops remains inadequate. Point-to-point equipment maintenance still faces challenges of being undetectable and unpredictable.Taking injection molding machines as an example, existing maintenance systems fail to effectively accommodate the diverse production needs across different factories. Significant disparities exist in equipment intelligence levels, while data monitoring devices remain relatively outdated. Equipment failures consequently cause severe production losses. Given the substantial market demand, there's an urgent need to enhance information-based management and intelligent control of injection molding equipment. This includes real-time monitoring of production data and predictive maintenance scheduling - specifically, establishing a mapping model between multi-source heterogeneous data and equipment health status, validating lightweight deployment of LSTM algorithms on edge computing nodes, and ultimately elevating intelligent operational standards.

2 SYSTEM DESIGN

The system is designed with a three-layer architecture, consisting of the data collection layer, edge computing layer, and maintenance decision-making layer. The data collection layer is responsible for integrating raw data and transmitting it to the edge computing layer. Data collection is distributed across various production processes to enable comprehensive monitoring of the entire production flow. The edge computing layer deploys edge computing devices at key nodes to process data streams. It handles high-feedback real-time parameters, reducing latency and cloud-side computational pressure, while leaving traditional data measurement and control tasks to be processed locally. The maintenance decision-making layer receives standardized data forwarded from the edge computing layer. It provides essential maintenance services, including overall data monitoring, individual equipment health assessment, and an open

application programming interface (API) to support the development of customized maintenance services. The overall architecture is shown in Figure 1.

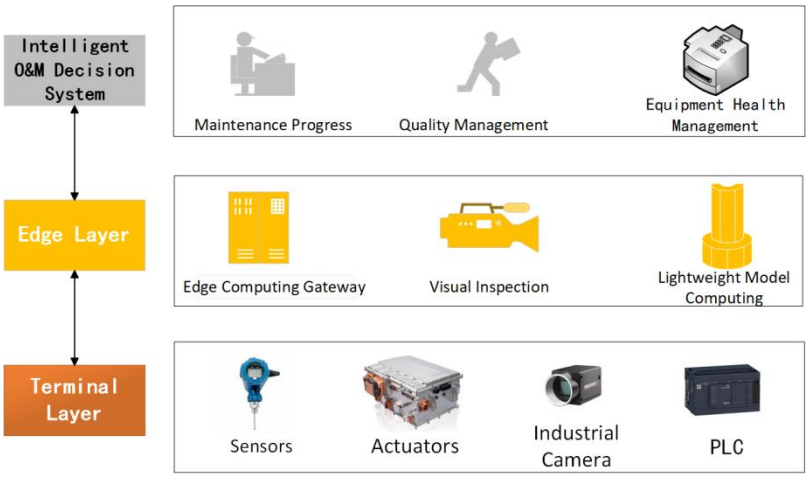


Figure 1 Architecture Design Diagram

The system needs to integrate a large volume of data, with a wide range of sizes, involving multiple process optimizations and complex algorithms. Maintaining high reliability, efficiency, and ease of deployment places high demands on the system design and the related technologies.

3 DATA COLLECTION

3.1 Data Source Analysis

Data collection needs to address the synchronization, integrity, and noise interference of multi-source heterogeneous data. In the injection molding process, a variety of data types need to be reported, covering multiple dimensions such as equipment, processes, and the environment[1]. These include equipment operational data such as real-time monitoring data on voltage, current, temperature, pressure, vibration, etc., as well as status data like on/off states, fault codes, and alarm signals, which are collected from IoT sensors, PLCs (Programmable Logic Controllers), and SCADA (Supervisory Control and Data Acquisition) systems. Production process data includes parameters such as temperature settings, pressure thresholds, and formula ratios, as well as quality control data like product size, weight, and production batch numbers. Environmental and energy consumption data include environmental data like temperature, humidity, and atmospheric pressure, as well as energy usage data such as water and electricity consumption, and emissions data including exhaust, wastewater, and noise levels. Additionally, for product quality inspection, defect detection images and industrial visual data such as barcodes/QR codes are also collected.

3.2 Data Collection Scheme Design

Based on the types of equipment involved in the injection molding process, a targeted data collection plan is designed, as shown in Figure 2. The specific plan is as follows:
Sensor-based Collection: Wireless sensors such as LoRa and Zigbee are used to transmit vibration and temperature data, suitable for mobile devices or long-distance scenarios. Wired sensors with 4-20mA/RS-485 interfaces are used for high-reliability requirements (e.g., high-pressure equipment).
Industrial Equipment Direct Collection: Data is directly collected from PLC/CNC devices by adapting industrial protocols such as OPC UA, Modbus, and Profinet.
Visual Image Collection: Industrial line scan cameras are used to capture equipment monitoring data, and industrial array cameras are used for defect detection in products.
Edge Gateway Collection: Edge computing gateways are deployed to integrate node data, enabling data packaging and uploading, along with local preprocessing of the data.

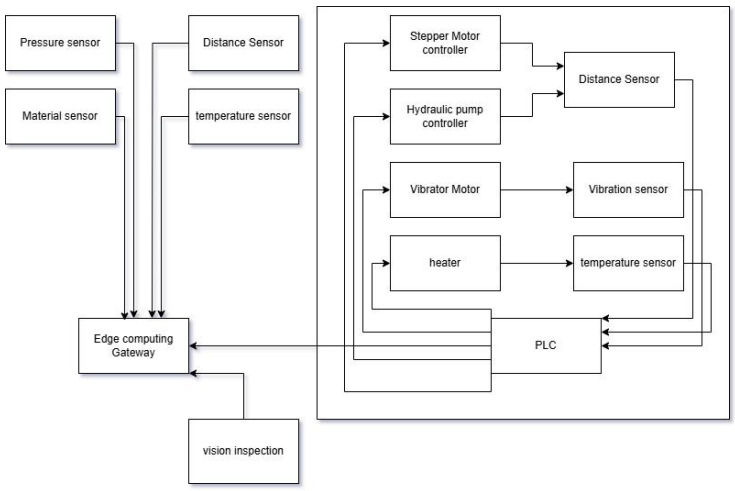


Figure 2 Data Collection Scheme Design Diagram

3.3 Data Collection Technology Implementation

The new sensor network, as an essential component of the modern Internet of Things (IoT), reflects the development trend of sensors in factory environmental monitoring through higher sensitivity, broader applicability, and lower environmental impact. The complementary use of wired and wireless sensors demonstrates broader application prospects. Wired sensors, with their high reliability, frequency, and precision in data collection, hold an irreplaceable position in collecting critical data, while wireless sensors strike the optimal balance between expanding the scope of collection and reducing collection costs.

The data collection module of this system is designed based on advanced Wireless Sensor Networks solutions (WSNs) in the industry[2], incorporating a wired sensor application environment model. This provides a reliable overall collection framework, ensuring data accuracy and transmissibility. Additionally, it integrates automatic control nodes such as PLCs, microcontrollers, and industrial PCs, combining the data collection capabilities of these controllers to offer a more flexible data node configuration for the collection framework.

3.4 Edge Computing Design

With the development of the Internet of Things (IoT), the number of edge devices has rapidly increased, and the amount of data generated has reached the zettabyte (ZB) level[3]. The centralized big data processing model, based on cloud computing, is no longer efficient enough to handle the data generated by edge devices, which has led to the emergence of edge computing technology. Compared to traditional data transmission gateways, edge computing gateways are more expensive, but the technological advantages they bring are significant. By reasonably designing edge nodes for device deployment, it is possible to balance the allocation of computing resources between the edge and the cloud, while also enhancing the speed of massive data processing. Therefore, we have introduced edge node design evaluation standards, using node reliability and performance requirements as criteria to ensure efficient data processing on edge devices.

For special assessments based on the characteristics of the injection molding environment, continuous operation tests were conducted at 85℃, and accelerated aging experiments were used to estimate whether the equipment's mean time between failures exceeds the 50,000-hour industrial requirement. Temperature and humidity range tests (-10℃ to 85℃ for thermal cycling) were performed to verify the stability of components under extreme environmental conditions. Basic performance evaluations, including LINPACK for floating-point operations, Dhrystone for integer operations, and quantifying computational power, were conducted to meet the real-time processing demands of equipment health algorithms. Memory bandwidth was evaluated using the Stream testing tool, and local storage IOPS were tested using the Fio tool for simulating random read/write operations to ensure efficient processing of high-frequency sensor data. Network throughput was measured using the iPerf3 tool, and end-to-end latency was tested with precise timestamp ping tests.

3.5 Edge Defect Detection

Existing industrial camera systems typically record image data and then transmit it to a cloud computing center after direct or simple image processing. However, as video data continues to grow in volume, applications in the injection molding field require defect detection systems to provide real-time and efficient image data processing. To address this, an edge computing model integrates computationally capable hardware units into the existing defect detection system hardware platform, supported by corresponding software technologies, creating a new type of defect detection system with edge computing capabilities[4].

In the edge computing model, computation typically occurs near the data source, i.e., image data processing takes place at the edge where the data is collected. To achieve this, a preprocessing function module based on intelligent algorithms is used to perform partial or full computational tasks on the real-time collected image data, while ensuring data

reliability. This enables timely responses to applications with high real-time requirements, while also reducing the computational and bandwidth load on the cloud computing center.

4 OPERATION AND MAINTENANCE DECISION MAKING

4.1 Operation and Maintenance Decision Making Design Approach

In injection molding production, continuous monitoring and assessment of the injection molding machine's health status is crucial for ensuring the safe operation of the factory. By establishing a quantitative relationship between monitoring data and equipment status, this system introduces the concept of injection molding machine health. It also incorporates machine health into the maintenance decision-making process, aiming to avoid both excessive maintenance and insufficient maintenance.

4.2 Injection Molding Machine Health Assessment

The core concept of injection molding machine health is to accurately calculate the equipment's operating status using data, with identifying key parameters being crucial. By referring to relevant Chinese national standards, data layering and linear normalization are applied to convert raw values such as vibration, temperature, and pressure into a range of [0, 1]. The general normalization method is as follows:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x represents the raw data.

A lightweight model is deployed at the edge layer to calculate the feature health (single-parameter level), and the weighted fusion algorithm used is expressed mathematically as follows:

$$\theta = w_a \times \theta_a + w_b \times \theta_b + w_c \times \theta_c + w_d \times \theta_d \tag{2}$$

Where w represents the weight, the sum of w equals 1, and θ represents the normalized key parameter value[5].

The processed data is forwarded to the cloud for subsequent subsystem health fusion calculations and global health calculations (at the equipment level)[6].

4.3 Predictive Maintenance Decision Design

The system predicts the failure downtime through equipment failure prediction, allowing for proactive maintenance scheduling to avoid production interruptions, reduce downtime, and lower maintenance costs. At the same time, combined with the injection molding machine health assessment, it provides optimal maintenance recommendations.

This system uses a Weibull distribution-based degradation model to fit the equipment performance degradation curve[7]. The Weibull distribution is a commonly used life distribution model that effectively describes the equipment failure process. Its probability density function is as follows:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{3}$$

Where β is the shape parameter and η is the scale parameter.

The parameters of the Weibull distribution are estimated using the maximum likelihood estimation method. Given a set of equipment failure time data t1, t2,..., tn, the estimated values for the shape parameter β and the scale parameter η are:

$$\hat{\beta} = \frac{n}{\sum_{i=1}^{n}\left(\ln\frac{t_i}{\bar{t}}\right)^2}\sum_{i=1}^{n}\left(\ln\frac{t_i}{\bar{t}}\right)\left(\frac{t_i}{\bar{t}}\right)^{\hat{\beta}} \tag{4}$$

$$\hat{\eta} = t^{-}\left(\Gamma\left(1+\frac{1}{\hat{\beta}}\right)\right)^{-1} \tag{5}$$

Where t^{-} is the sample mean, and Γ is the Gamma function. 3. Remaining useful life prediction.

Based on the estimated Weibull distribution parameters, the remaining useful life (RUL) of the equipment can be predicted. Given the current time t0, the remaining useful life of the equipment is:

$$RUL(t_0) = \eta\left(-\ln(1 - F(t_0))\right)^{1/\beta} - t_0 \tag{6}$$

F (t0) represents the cumulative failure probability of the equipment at time t0[8].

5 CONCLUSION

This paper presents a data-driven injection molding production operation and maintenance decision system, which integrates multi-source heterogeneous data collection, edge computing, and intelligent analysis technologies. It enables dynamic monitoring of equipment health status and predictive maintenance, effectively addressing the inefficiencies and delayed responses inherent in traditional operation and maintenance models. The system adopts a three-layer architecture design, combining lightweight models and Weibull distribution algorithms, which significantly enhance the real-time processing of data and the accuracy of maintenance decisions. This solution is not only suitable for large intelligent factories but also adaptable to the production environments of small and medium-sized enterprises, offering high flexibility and scalability. In the future, with the further development of the industrial Internet of Things and artificial intelligence technologies, the system can further enhance its intelligence level through optimization algorithms

and strengthened edge computing capabilities, providing strong support for the digital transformation of the manufacturing industry.

COMPETING INTERESTS

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

REFERENCES

[1] Benhua Z, Xuguang W, Peilong G, et al. Application of field data acquisition in intelligent manufacturing system. Manufacturing Technology & Mechine Tool, 2016, 2016(6): 33-39.
[2] Qian Z, Wang Y. Internet of things-oriented wireless sensor networks review.Journal of Electronics and Information Technology , 2013, 35(1): 215-227.
[3] Shi W, Sun H, Cao J, et al. Edge computing-an emerging computing model for the internet of everything era. Journal of computer research and development, 2017, 54(5): 907-924.
[4] Shi W, Zhang X, Wang Y F, et al. Edge computing: state-of-the-art and future directions. Journal of Computer Research and Development, 2019, 56(1): 69-89.
[5] LEI Yaguo, CHEN Wu, LI Naipeng, LIN Jing.A Relevance Vector Machine Prediction Method Based on Adaptive Multi-kernel Combination and Its Application to Remaining Useful Life Prediction of Machinery. DOI:10.3901/JME.2016.01.087
[6] Deng C, Sun Y Z, Li R, et al. Hidden markov model based on the heavy-duty CNC health state estimate. Computer Integrated Manufacturing System, 2013, 19(03).
[7] ZHAO Shenkun, JIANG Chao, LONG Xiangyun.Remaining Useful Life Estimation of Mechanical Systems Based on the Data-driven Method and Bayesian Theory. DOI: 10.3901/JME.2018.12.115
[8] Yu X H, Zhang L B, Wang C H, et al. Reliability life analysis of the equipment based on new Weibull distribution parameter estimation method. Journal of mechanical strength, 2007, 29(6): 932-936.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3014

EEG-BASED EMOTION RECOGNITION USING ENSEMBLE LEARNING MODEL

YunYi Wang
Zhixin high school, Guangzhou 510060, Guangdong, China.
Corresponding Email: usk19231899@outlook.com

Abstract: Throughout history, emotion recognition has played a pivotal role across various fields. Due to technology limitation, emotions were traditionally evaluated through interviews or questionnaires, methods prone to subjectivity even among psychologists. Hence, development of automated emotion recognition system is necessary. In recent years, prodigious process has been made in this area, with EEG-based emotion recognition becoming increasingly popular. However, the models that used to classify the EEG-based emotion data are still not powerful enough. In this paper, we propose an efficient model for emotion classification, utilizing EEG data from DEAP dataset. The neural signals are first decomposed into the Gamma, Alpha, Beta, and Theta bands according to their respective frequencies, and preprocessed using Welch method. A hybrid model, combining Long Short-Term Memory (LSTM), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA), is employed for classification. This hybrid model distinguishes among three emotional states: positive, negative, and neutral. Compared to existing work, our ensemble learning model achieves a higher accuracy, performs with 95.99% and 95.68% for arousal and valence. Furthermore, our method successfully mitigates the weaknesses of these individual base models, bringing in a more robust emotion recognition framework.
Keywords: EEG; Emotion recognition; Ensemble model; Machine learning; Long-Short Term Memory

1 INTRODUCTION

Emotions play a pivotal role in shaping human behavior and cognition, largely governed by the brain neural activities. Thus, understanding the connection between emotions and brain signals has gained increasing attention[1-4].This research area has found applications in various fields, such as human-computer interaction (HCI) systems[5-9], and medical domains, including the diagnosis of conditions like autism spectrum disorder (ASD) and depression[10-14]. Among the many approaches to studying emotions, EEG-based emotion recognition has emerged as a focal point for researchers. Investigating this relationship is vital for advancements in Artificial Intelligence and Ambient Intelligence[15-16]. The rapid evolution of brain-device interface technology highlights the importance of this area[17-21].

A crucial step in EEG-based emotion recognition is feature extraction. Yoon and Chung[22] employed Fast Fourier Transform (FFT) to extract features from EEG signals in the DEAP dataset. While FFT is useful for frequency analysis, it may overlook important time-domain information. Bhardwaj et al.[23] used Independent Component Analysis (ICA) to decompose EEG signals into five bands, but ICA can be sensitive to noise, which may affect signal accuracy. Alhagy et al.[24] focused on specific frequency bands (theta, alpha, beta) for feature extraction, but limiting the analysis to only a few bands might omit valuable data from other frequencies. Bazgir et al.[25] applied the Average Mean Reference (AMR) method for noise reduction, though it could also remove some signal information along with the noise. Atkinson and Campos[26] used Maximum-Relevance (mRMR) for feature selection, enhancing relevant features but potentially excluding others that might still be useful. Liu et al.[27] extracted features from multiple domains, offering comprehensive analysis but adding computational complexity. Naser and Saha[28] employed advanced techniques like DT-CWPT and SVD for feature selection, though these methods may be computationally intensive. Bhagwat and Paithane[29] utilized Wavelet Transform, which is effective but can be limited by the choice of wavelet function.

Once features are extracted, classification plays a crucial role in emotion recognition. In their work, Bhardwaj et al.[23] used SVM and LDA classifiers on preprocessed data, achieving moderate accuracies. Liu et al.[27] opted for KNN and Random Forest classifiers, refining their models to achieve good results, though they required adjustments for broader applications. Naser and Saha[28] employed SVM as their classifier, which, while effective for feature selection, did not provide the complexity needed for higher accuracy. Alhagy et al.[24] employed a Long Short-Term Memory (LSTM) network with multiple layers, including Dropout and Dense layers, to classify the EEG data into arousal, valence, and liking classes. Despite their relatively high accuracies of around 85%, improvements were still necessary for better generalization. Bazgir et al.[25] used SVM, KNN, and Artificial Neural Network (ANN) models, with SVM incorporating a Radial Basis Function (RBF) kernel and KNN using different nearest neighbors, resulting in accuracies exceeding 90%. However, further refinement was required in certain channels. Iyer et al.[30] introduced a hybrid approach combining Convolutional Neural Networks (CNNs) and LSTMs, achieving a remarkable 97.16% accuracy. Nevertheless, their model was computationally expensive, especially when applied to small datasets.

Considering the challenges in both feature extraction and classification, this paper proposes a novel method that combines Welch's method for feature extraction with an ensemble model for classification. Welch's method effectively captures frequency-domain information while reducing noise, making it well-suited for EEG signal analysis. Our

ensemble model ,combining LSTM, SVM, LDA, and KNN, is tested on the DEAP dataset, divided into four frequency bands: theta (4-8Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-64 Hz). This ensemble model, leverages the strengths of each classifier: LSTM excels at handling time-series data, while SVM and LDA are effective for binary classification tasks, and KNN is simple yet powerful for pattern recognition. The ensemble method improves robustness and generalization, making the model adaptable to both simple and complex datasets, providing a more robust approach to EEG-based emotion classification.

This paper is organized as follows: Section II introduces the dataset and the ensemble learning-based emotion recognition framework, along with the improvements made to these methods. Section III discusses the pre-processing, visualization, and decoding results using the DEAP dataset, followed by a comparison of our ensemble model with others. The conclusion is presented in Section IV (Figure 1). The whole process for creating the ensemble model can be divided into five parts, including data acquisition, data preprocessing, feature extraction, Ensemble Learning, and Emotion Prediction.

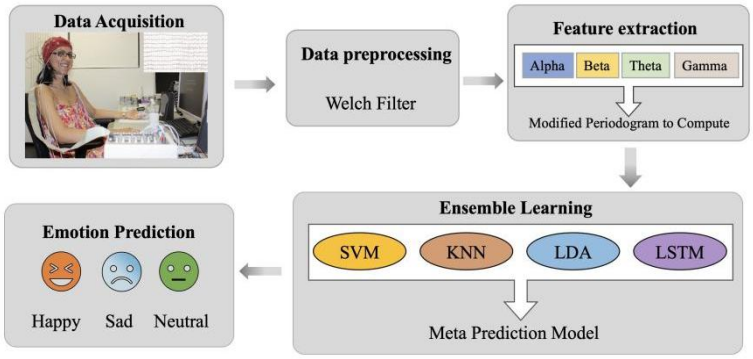


Figure 1 Structure of the Paper

2 METHODS

2.1 Dataset

We used data from the DEAP dataset, which was recorded when 32 participants watched 40 music videos. Each video lasted one minute. The participants were aged between 19 and 32,with 16 males and 16 females. The dataset recorded the emotional changes of each participant while watching these videos and rated the videos according to four aspects: arousal, valence, dominance, and liking. Each file contains data, including 40 trials and 40 channels, and a label array, which contains 4 trials and 4 subjective ratings. The sampling frequency was downsampled to 128Hz. The data was preprocessed according to these labels.

2.2 Ensemble Learning-Based Emotion Recognition Framework

We proposed an EEG-based emotion recognition architecture that employs ensemble learning model as illustrated in Figure 1. The EEG data were sourced from the DEAP dataset, where participants' signals were recorded and preprocessed while they watched short video clips. After preprocessing, features were extracted and periodograms were adjusted for further analysis. The extracted features were then used to train the LDA, SVM, LSTM, and KNN models. Based on the outputs from these models, emotion predictions were generated.

2.1.1 Data preprocessing and feature extraction

EEG signals have four different frequency bands: theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and gamma (30-64 Hz), which are commonly analyzed in the emotion recognition field. We use the Welch method to estimate the power spectrum features based on the preprocessed data.

In the Welch method, we first decompose the signal into several blocks based on their ranges and calculate the periodogram of each block. Then, we estimate the power spectra by averaging the periodograms of the blocks. The formula for the Welch method is shown below:

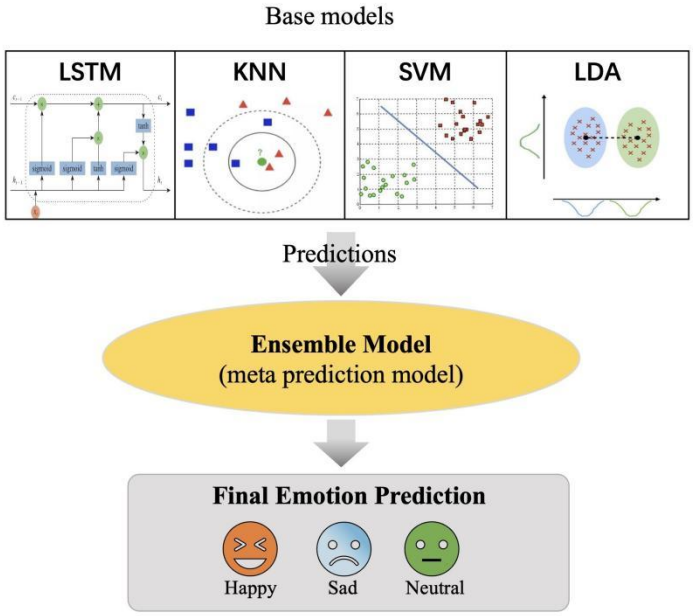
$$\hat{S}_x \triangleq \frac{1}{K}\sum_{m=0}^{k-1}\left[\frac{1}{L}|FFT(x_n)|^2\right] \quad (1)$$

We decompose the signal into n continuous blocks.x_n represents each block. K is the total number of blocks in the signal. And L represents the total number of points in these signals. Both the 32 channels in the DEAP dataset and the four EEG bands result in 128 power features. After the preprocess, our result in each band is plotted as an image that show in the result part.

2.1.2 Ensemble learning model

Machine learning has become a crucial component in the study of emotions due to its strong capability in model training and solving computational intelligence problems. Among the various techniques available, ensemble models stand out as they enhance model performance through different strategies. Three of the most widely used ensemble learning methods are Boosting, Bagging, and Stacking.

In this paper, we focus on Stacking, also known as stacked generalization, which is considered one of the most effective ensemble modeling techniques. Stacking improves predictions by combining the outputs of multiple base models—in our case, SVM, LSTM, KNN, and LDA—to generate a more accurate final prediction. All sub-models contribute equally, which is why the technique is referred to as "stacking." The diagram illustrates the architecture of our stacking model as shown in Figure 2 (The base models are LSTM, KNN, SVM, and LDA, by putting the labels of these models in to the Ensemble Model, we successfully increasing the accuracy of emotion recognition). We utilize LSTM, SVM, LDA, and KNN as base models, whose predictions are then combined to form the ensemble, which ultimately produces



the final prediction.

Figure 2 Structure of the Ensemble Model

2.3 Classification Algorithms

2.3.1 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a technique that uses statistics and machine learning to find the linear combination of features that best separates two or more classes of objects. The resulting combination is often used for dimensionality reduction. As a classic linear learning method, LDA projects of the same kind of samples closed to each other and far from other classes in a given training set. The targets for using LDA are to minimize intra-class covariance(make similar projection points as close as possible) and maximize inter-class covariance(keep the Heterogeneous projection points as far as possible). For a given data set $D = \{(x_i, y_i)\}_{i=1}^{m}, y_i \in \{0,1\}$, the objective function is shown as:

$$J = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w} = \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T(\Sigma_0+\Sigma_1)w} \quad (2)$$

Where X_i, μ_i, Σ_i represent the collection of the samples that belong to $i \in \{0,1\}$, mean vector, and covariance matrix. If projecting the data onto the line w, the projection of the centers of these two kinds of samples are $w^T\mu_0$ and $w^T\mu_1$.

2.3.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) are generalized linear classifier used to infer a function or relationship from data through supervised learning. It can find the hyperplane that best separate the data into different classes and maximize the margin between the classes. this hyperplane is defined as follows:

$$\min_{w,b} \frac{1}{2}\|w\|^2, s.t., y_i(wx_i + b) \geq 1 \quad (3)$$

Where (x, y) represents the coordinates of these points, w, b denotes the vector and bias.

In this way, data were separated into two parts. SVMs are commonly applied in pattern recognition, classification and regression analysis. This supervised learning algorithm can successfully separate these two classes when a set of data that consist of two different classes is provided.

2.3.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet effective machine learning algorithm used to recognize and classify unknown objects. It excels at classifying objects with a large sample size. The result of KNN is determined by the class of one or more samples closest to the unknown samples. The ‘k’ in KNN represents the number of the closest neighbors considered when making a classification decision. however, it can be computationally intensive with large datasets, as it requires calculating the distance between the unknown sample and all the existing samples in the dataset.

In order to use KNN model for classifying data, we need to determine the distance between two instance points in a feature space. Given the dataset $x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^T$ and $x_j = \left(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)}\right)^T$, the formula of the distance L_p is defined as:

$$L_p(x_i, x_j) = \left(\sum_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}} \quad (4)$$

Where x_i and x_j represent the two points in feature space X.

2.3.4 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a method that is created in order to solve the long-term dependence problem in Recurrent Neural Network (RNN). In LSTM model, issue of vanishing/exploding can be overcome. LSTM commonly consists of input gates, forget gates, memory cells and output gates.

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \quad (5)$$

$$f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \quad (6)$$

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \quad (7)$$

Where i_t means the input gate is open or close,f_t represents the state of the forget gate, o_t represents the state of the forget gate. x_t is the value of the data, h_t represents the hidden state of the memory cell, W and b are weights and biases, respectively, and σ means the sigmoid function.

The sigmoid activation function in the input gates decide which information is important and in the output gates choose the information that should be output. The tanh activation function produces new information and process information for output in input and output shapes. Forget gates, only have sigmoid activation functions and decide which information should be removed.

We also need to compute the state of the memory cell. The formula of the cell's state c_t is

$$c_t = f_t c_{t-1} + i_t c_t^{'} \quad (8)$$

$$c_t^{'} = \tanh(W_c[x_t, h_{t-1}] + b_c) \quad (9)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (10)$$

Where $c_t^{'}$ is the status of candidate cell and h_t represents the hidden state of the cell.

3 RESULT

In this section, we will first visualize the DEAP dataset data in terms of valence and arousal, and explain the process of label generation. Next, we will use Welch's filter to calculate the power spectrum for each frequency band to facilitate further analysis. Finally, we will present and compare the classification results obtained with individual algorithms and the ensemble learning model, focusing on decoding accuracy.

First, we visualize the structure of the DEAP dataset. EEG and peripheral physiological signals were recorded from 32 participants as they watched 40 music videos each. Participants rated each video based on four subjective dimensions: arousal, valence, dominance, and liking. The data was downsampled to 128Hz, preprocessed, and stored in pickled Python formats. Each participant's file contains two arrays: a "data" array (40 trials x 40 channels x 8064 data points) and a "label" array (40 trials x 4 ratings: valence, arousal, dominance, and liking). We combined the data files into two new arrays, resulting in 880 trials for all 32 participants.

Valence represents the degree of positivity or negativity of an emotion, while arousal indicates the intensity of the emotional state. We plot the first 40 trials of one subject, as shown in Figure 3, where the x-axis represents trial numbers and the y-axis shows intensity. The plot reveals that the combinations of valence and arousal can be categorized into four emotional states: High Arousal Positive Valence (Excited, Happy), Low Arousal Positive Valence (Calm, Relaxed), High Arousal Negative Valence (Angry, Nervous), and Low Arousal Negative Valence (Sad, Bored). Based on these distributions, we generate box plots to visualize the four emotional classes under the conditions of valence and arousal as shown in Figure 3.

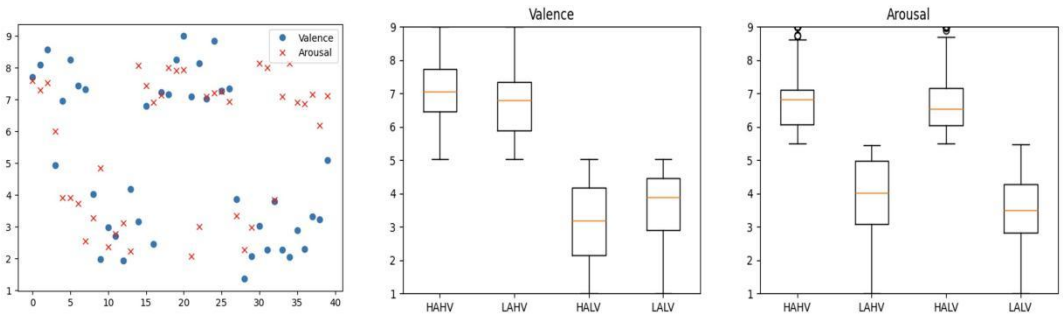


Figure 3 The result of Data Preprocessing

Next, we use Welch's method to extract the band power from each channel, as illustrated in Figure 4. The x-axis represents frequency, while the y-axis denotes power spectral density. The curve shows the distribution of power spectral density across different frequencies. The shaded areas—blue, yellow, green, and red—represent the theta, alpha,

beta, and gamma bands, respectively. We observe that the power spectral densities in the theta and alpha bands are generally higher compared to the beta and gamma bands. From the chart, it is evident that the trends in the theta and alpha bands are not straightforward, showing neither a consistent increase nor decrease, with notable fluctuations in density. In contrast, the data in the beta and gamma bands display a more discernible diminishing trend.

During preprocessing, we also generated topomaps for each frequency band at different time points. Each row corresponds to a specific band, while each column represents a different time stamp (Figure 4). The most noticeable changes occur in the Theta and Alpha bands, whereas the Beta and Gamma bands show less variation over time. This observation aligns with the results from Welch's periodogram, where we also see significant fluctuations in the Alpha and Theta bands, and more consistent patterns in the Beta and Gamma bands.

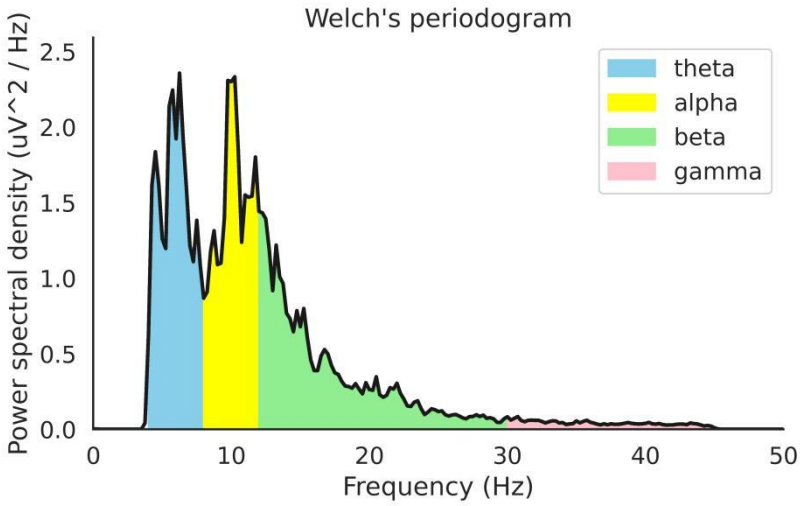


Figure 4 Frequency Distribution across Different Data Bands after Feature Extraction, Distinct Colors are used to Visualize Differences among Bands more Clearly

After the preprocessing, we used individual algorithms (SVM, KNN, LSTM, and LDA) to classify the preprocessed data. The data was shuffled five times with different initializations, with 70% used for training, 10% for evaluation, and the remaining 20% for testing. For SVM, a simple linear kernel was applied. In KNN, the number of neighbors was set to 5. For LDA, the SVD solver was employed. The LSTM model used the standard LSTM formulation, with logistic functions on the gates and hyperbolic tangents for activations. The input size was 10752*32, and the model included two LSTM layers with dropout (rate = 0.5) between them. The Adam optimizer was used, with a learning rate of 0.0001. The classification accuracies for valence and arousal using these methods are shown in the table. For SVM, the highest accuracy for arousal was 61.79%. LDA achieved an optimum accuracy of 60.16%. KNN reached 69.11% accuracy for valence. The LSTM classifier performed the best, with 83.66% accuracy for arousal in the beta band and 84.77% for valence in the theta band and the we also printed the training loss of LSTM (Figure 5). After evaluating all the single algorithm, we applied an ensemble model to further improve classification accuracy. The base models (SVM, KNN, LSTM, and LDA) first produced emotion predictions, which were then used as inputs for the ensemble model. Specifically, we combined the outputs of these four individual models into a new input vector, and fed it into a meta-model, which in this case was a logistic regression model. This meta-model integrated the predictions from the base models to generate the final output.

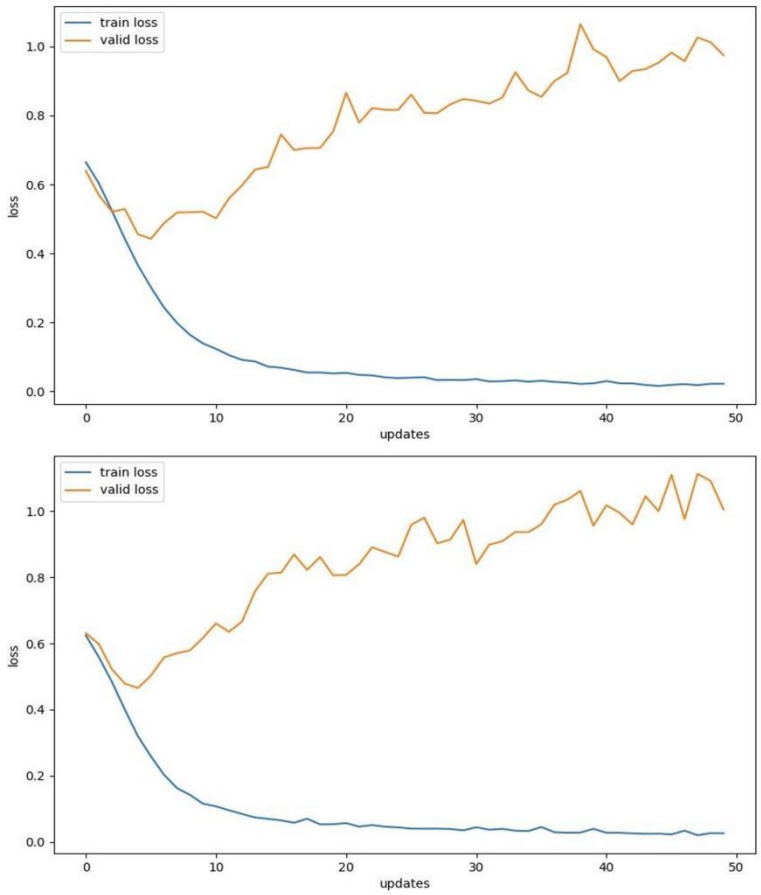


Figure 5 The Picture at the Top is the Train Loss of LSTM for Valence and the Picture at the Bottom is the Train Loss for Arousal

The performance of the ensemble model is shown below in Table 1. Compared to other researchers, our model achieves higher accuracy. The overall accuracy of SVM and LDA in Bhardwaj et al.[23]'s research are 74.13% and 66.50% and lower than the accuracies of our ensemble Model, which are 95.68% and 95.99%.Similarly, our results are also better than the accuracies in Nawaz et al.[7]'s work, which are 77.62% and 78.96% for valence and arousal. Compared to the individual models, the ensemble method also demonstrated superior accuracy. The accuracies of the ensemble model are fully deserved the best among these models since it achieved 95.68% accuracy of valence and 95.99% accuracy of arousal which are the highest accuracies for valence and arousal. By leveraging the strengths of each base model, the ensemble was able to better capture the complexity of the data, leading to improved classification results for both valence and arousal across all frequency bands. This highlights the effectiveness of ensemble learning in emotion recognition tasks, as it mitigates the limitations of single models and provides more robust predictions.

Table 1 The results of the Ensemble Model Compared to Other Base Models

Method	Valence	Arousal
LSTM	84.77	83.66
KNN	69.11	64.23
SVM	60.16	61.79
LDA	58.54	60.16
Ensemble Model	95.68	95.99

4 CONCLUSION

Emotion recognition has gained significant attention in recent years due to its applications in human-computer interaction and mental health analysis.In this study, we aimed to find a method that is both efficient and practical for emotion recognition using EEG signals. The results demonstrate the potential of EEG-based approaches for accurate emotion classification. We first divided the DEAP dataset into four frequency bands—Theta, Alpha, Beta, and Gamma—and used Welch's method for preprocessing. Then, we proposed a hybrid model that integrates SVM, KNN, LDA, and LSTM to classify emotions into positive, negative, and neutral categories. Among the base models, SVM outperformed LDA and KNN, with LSTM achieving the highest accuracy of 90%. Additionally, the average accuracy

of LSTM across all trials was noteworthy. The hybrid model further improved classification performance, outperforming individual classifiers and providing a more accurate and robust framework.

One of the key advantages of the hybrid model is its ability to mitigate the limitations of individual algorithms. For instance, SVM can be sensitive to irrelevant data, while LSTM and KNN face challenges related to data complexity. By combining the strengths of each model, the hybrid approach compensates for these weaknesses, resulting in more reliable predictions.

However, despite these promising results, there are several areas for improvement in future work. One potential enhancement involves diversifying the emotion stimuli. Currently, most datasets, including DEAP, rely on pictures or short videos to evoke emotions. Future studies could explore alternative stimuli, such as text or speech, which may elicit more nuanced emotional responses in certain contexts.

Another improvement lies in the feature extraction process. Although Welch's method proved effective for our data, more advanced and sophisticated techniques may lead to better results, particularly with complex datasets.

Finally, the biggest challenge remains applying this technology in real-world scenarios. While EEG-based emotion recognition has made significant strides, it is still in the experimental phase. Real-world applications face hurdles such as inefficient classifiers and noise from external factors, which complicate real-time analysis. Overcoming these obstacles is crucial for transitioning from experimental results to practical, real-world implementations of emotion recognition systems.

COMPETING INTERESTS

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

REFERENCES

[1] L Chen, X Mao, Y Xue, et al. Speech emotion recognition: Features and classification models. Digital Signal Processing, 2012, 22(6): 1154–1160.

[2] A Dzedzickis, A Kaklauskas, V Bucinskas. Human emotion recognition: Review of sensors and methods. Sensors, 2020, 20(3): 592.

[3] X Li, Y Zhang, P Tiwari, et al. EEG based emotion recognition: A tutorial and review. ACM Computing Surveys, 2022, 55(4): 1–57.

[4] S Sahu, R Kumar, H V Long, et al. Early-production stage prediction of movies success using K-fold hybrid deep ensemble learning model. Multimedia Tools and Applications, 2023, 82(3): 4031–4061.

[5] R Jenke, A Peer, M Buss. Feature extraction and selection for emotion recognition from EEG. IEEE Transactions on Affective Computing, 2014, 5(3): 327–339.

[6] N Thammasan, K Moriyama, K Fukui, et al. Continuous music-emotion recognition based on electroencephalogram. IEICE Transactions on Information and Systems, 2016, 99(4): 1234–1241.

[7] R Nawaz, K H Cheah, H Nisar, et al. Comparison of different feature extraction methods for EEG based emotion recognition. Biocybernetics and Biomedical Engineering, 2020, 40(3): 910–926.

[8] X Li, Y Zhang, P Tiwari, et al. EEG based emotion recognition: A tutorial and review. ACM Computing Surveys, 2022, 55(4): 1–57.

[9] R W Picard. Building HAL: Computers that sense, recognize, and respond to human emotion. Proceedings of SPIE – The International Society for Optical Engineering, 2001, 4299: 518–523.

[10] A Lartseva, T Dijkstra, C C Kan, et al. Processing of emotion words by patients with autism spectrum disorders: Evidence from reaction times and EEG. Journal of Autism and Developmental Disorders, 2014, 44: 2882–2894.

[11] A Kołakowska, A Landowska, M Szwoch, et al. Emotion recognition and its applications. Human-Computer Systems Interaction: Backgrounds and Applications 3, 2014: 51–62.

[12] J Á Bitsch, R Ramos, T Ix, et al. Psychologist in a pocket: Towards depression screening on mobile phones. Stud Health Technol Inform, 2015, 211: 153–159.

[13] R Ge, G Feng, X Jing, et al. ENACP: An ensemble learning model for identification of anticancer peptides. Frontiers in Genetics, 2020, 11: 760.

[14] M Mohy-Eddine, A Guezzaz, S Benkirane, et al. An ensemble learning based intrusion detection model for industrial IoT security. Big Data Mining and Analytics, 2023, 6(3): 273–287.

[15] R Dunne, T Morris, S Harper. A survey of ambient intelligence. ACM Computing Surveys (CSUR), 2021, 54(4): 1–27.

[16] U A Khan, Q Xu, Y Liu, et al. Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects. Multimedia Systems, 2024, 30(3): 115.

[17] M A Lebedev, M A L Nicolelis. Brain–machine interfaces: Past, present and future. Trends in Neurosciences, 2006, 29(9): 536–546.

[18] K S Rao, S G Koolagudi. Emotion recognition using speech features. Springer Science & Business Media, 2012.

[19] E H Houssein, A Hammad, A A Ali. Human emotion recognition from EEG-based brain–computer interface using machine learning: A comprehensive review. Neural Computing and Applications, 2022, 34(15): 12527–12557.

[20] Y Cai, X Li, J Li. Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. Sensors, 2023, 23(5): 2455.

[21] K Okoye, J T Nganji, J Escamilla, et al. Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education. Computers and Education: Artificial Intelligence, 2024, 6: 100205.
[22] H J Yoon, S Y Chung. EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm. Computers in Biology and Medicine, 2013, 43(12): 2230–2237.
[23] A Bhardwaj, A Gupta, P Jain, et al. Classification of human emotions from EEG signals using SVM and LDA classifiers. International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2015: 180–185.
[24] S Alhagry, A A Fahmy, R A El-Khoribi. Emotion recognition based on EEG using LSTM recurrent neural network. International Journal of Advanced Computer Science and Applications, 2017, 8(10).
[25] O Bazgir, Z Mohammadi, S A H Habibi. Emotion recognition with machine learning using EEG signals. National and 3rd International Iranian Conference on Biomedical Engineering (ICBME). IEEE, 2018: 1–5.
[26] J Atkinson, D Campos. Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers. Expert Systems with Applications, 2016, 47: 35–41.
[27] J Liu, H Meng, A Nandi, et al. Emotion detection from EEG recordings. International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2016: 1722–1727.
[28] D S Naser, G Saha. Recognition of emotions induced by music videos using DT-CWPT. Indian Conference on Medical Informatics and Telemedicine (ICMIT). IEEE, 2013: 53–57.
[29] A R Bhagwat, A N Paithane. Human disposition detection using EEG signals. International Conference on Computing, Analytics and Security Trends (CAST). IEEE, 2016: 366–370.
[30] A Iyer, S S Das, R Teotia, et al. CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. Multimedia Tools and Applications, 2023, 82(4): 4883–4896.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3015

BREAKING THE "WORD POVERTY" AND MAKING THE LANGUAGE LIVELY —— INVESTIGATION AND ANALYSIS ON THE CURRENT SITUATION OF COLLEGE STUDENTS' WORD POVERTY AND ITS INFLUENCING FACTORS

Xing Zeng
Department of Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Author: Xing Zeng, Email: 1446118780@qq.com

Abstract: In contemporary society, the phenomenon of lexical poverty is becoming increasingly prominent, as reported by the Chinese Youth Daily, which highlighted the lexical poverty issue among young people. Survey data indicates that over half of the interviewed youth feel their language expr ession abilities are gradually declining. Additionally, 47.1% of the interviewed youth believe they have a limited vocabulary and exhibit repetitive expressions. They often experience cognitive difficulties during speech, such as mental blockages, forgetting words while writing, and experiencing linguistic lapses. For university students, proficient language expression is one of their essential competencies. However, the research on lexical poverty among university students in China is currently limited, necessitating a deeper exploration of this phenomenon. This study employed a questionnaire survey method to investigate the lexical poverty status and influencing factors among university students in four r egions of Guangxi. A total of 925 questionnaire responses were collected, of which 772 were deemed valid after removing invalid responses. Basic descriptive analysis of the data provided insights into its distribution. Subse quently, differential analysis and multiple response analysis were conducted to explore the association between respondent characteristics and research questio ns. Finally, a structural equation model was constructed to explore the complex relationship between university students' understanding of lexical poverty and their behaviors. The research findings reveal that lexical poverty is widespread among university students, posing a significant challenge to their language expression abilities. The main reasons include inadequate reading habits, internet usage patterns, and limited face-to-face communication. Additionally, educational background and academic disciplines also influence the occurrence of lexical poverty. Students with higher educational levels experience r elatively fewer instances of lexical poverty compared to their counterparts, with students in STEM fields being more susceptible than those in humanities. Moreover, the frequency of lexical poverty increases significantly with the growing reliance on the internet, declining reading habits, intensified fragmented reading patterns, reduced face-to-face communication, and decreased participation in organized communication activities at universities.In light of these findings, it is recommended to strengthen guidance and regulation on internet information, conduct vocabulary education and language training activities, and actively encourage and support students to engage in debate, public speaking, writing, and other practical activities to alleviate lexical poverty among university students.
Keywords: University students; Lexical poverty; Differential analysis; Multiple response; Structural equation

1 INTRODUCTION

1.1 Research Background and Significance of the Topic

1.1.1 Research background

In recent years, with the rapid development of information technology and the popularity of social media, the phenomenon of word poverty among college students has aroused widespread concern. As the main users of information technology, college students frequently communicate on the Internet and social media. After the Internet entered the human world, the keyboard became a communication tool. Tapping on the keyboard and clicking on the mobile phone has become the normal state of most people's "writing". "Forgetting the words with a pen" has become a pain point for many people at present. In addition, many young people also feel that their language skills have declined[1]. In the Internet age, the way people get information is often fragmented and fast-food. Pictures can convey information more quickly and conveniently than words, and they are also more popular. Today, people prefer to gain knowledge by watching short videos rather than reading long articles. Young people who have grown up in the internet age have faced the problem of "poor words" over time[2].This high-frequency virtual communication mode has had a far-reaching impact on their daily life and shaped their communication methods and habits. However, some studies have found that despite their outstanding performance in the digital world, college students often encounter word poverty in written and oral expressions. This phenomenon is not only limited to the ability of language expression, but also may have a negative impact on college students' way of thinking and social communication. Therefore, it is of great significance to study the phenomenon of college students' poor words. By deeply understanding the causes and effects of word poverty, we can provide effective countermeasures and methods for improving college students' language

expression ability. At the same time, it also helps to promote the development of college students' thinking ability and improve their comprehensive quality.

1.1.2 Research purposes

In today's rapidly developing design society, we should keep more of our love for traditional culture and absorb more nutrients from tradition. Only in this way can we not be overwhelmed by western design theories and techniques, and we can get out of the dilemma of unyielding reason and poor words"[3].Ask for help, how to describe the scenery in this picture", "How to praise a person euphemistically" and "What are the words to express amazement" ... Nowadays, similar "help posts" are not uncommon on online platforms. Behind these "help posts", many young people are troubled by the lack of words and language[4]. The construction of national language ability is an important part of national development, and it is also one of the core propositions in the study of national language conditions. As the foundation of national language ability, citizens' language ability plays a key supporting role in the language environment and cultural construction of the whole country. There has been such a paragraph on the internet. In the face of beautiful scenery, many people can only sigh again and again, "How beautiful! It's beautiful! " Obviously, I have read many poems, and my emotions are stirring in my chest, but I can't express them when I talk about them. The only excuse is "poor words"[5]. In this process, the language ability of college students is very important, because they represent the main force of the future society, and the improvement of their language expression ability is directly related to the language of the whole country quality and international competitiveness.

This paper aims to deeply study the causes and influencing factors of college students' word poverty, so as to better understand the language habits and communication characteristics of this group. By analyzing the phenomenon of college students' word poverty, it can provide a basis for formulating more effective language education strategies. Such research not only helps to improve college students' language expression ability, but also promotes their all-round development in academic, professional and social fields. Only on the basis of comprehensive improvement of language ability can we better meet a richer and more diverse information age and realize the long-term development goal of the country.

1.1.3 Research significance

The ability to "speak" is first of all to have something to say, that is, to have the content of expression. Where does the content come from? Content has two sources, one is from life, and the other is from books. Only by guiding students to carefully observe and experience life, read widely, and form good study habits, the content will be like flowing water from the source, accumulating more and more[6]. In this book, through the analysis of the collected data, we will clearly understand the attitude of college students to the problem of word poverty and the actual impact of this phenomenon on their lives. Based on these analysis results, we put forward corresponding improvement measures and suggestions to help college students better cope with the problem of word poverty, improve their quality of life and learning efficiency, and then better adapt to college life and go to society. By improving college students' language ability, they will have a better chance to integrate into society, thus making positive contributions to China's economic development and social stability. This will also help to promote the progress of social exchanges and communication, build a more harmonious and inclusive social environment, and lay a solid foundation for the country's long-term development goals. Therefore, the significance of this study is not only to solve the problem of college students' word poverty itself, but also to make positive contributions to the progress of society and the country.

1.2 Research Status

Recently, I often hear the argument that it is difficult to write comments, especially the old problems that are often publicized in newspapers, which are even more difficult to write, because we can't find new arguments and new words, and there is quite a tendency to "run out of words"[7].Educational linguists across England and the USA have long critiqued deficit-based language ideologies in schools, yet since the early 2010s, these have enjoyed a marked resurgence in England's education policy as evident in discourses, funding, and pedagogical materials related to the so-called 'word gap'[8]. According to a recent survey conducted by the social survey center of China Youth Newspaper, 1333 young people were surveyed. The results show that more than half of the respondents (53.3%) feel that their language and writing skills have declined in recent years, while 47.1% of the respondents feel that their vocabulary is insufficient and their expression is single. At the same time, 43.2% of the respondents said that the frequency of writing decreased, and 41.5% of the respondents had encountered unclear expression and unsatisfactory words[9]. This trend is closely related to the development of the Internet era. Many young people are used to using buzzwords and expression packs to communicate on online platforms. However, in offline communication or formal occasions, they often forget words, and there is a certain gap between thinking and oral expression, so it is difficult to find an appropriate expression. In addition, in terms of written expression, words are often forgotten or ambiguous.

From "YYDS" to "Juejuezi", from "simply" to "too cool and spicy", has your expressive ability declined because of the popularity of online language? For example, I want to send a circle of friends with "culture", but I don't know where to start; When visiting relatives and friends, "speak a few words" and hold back for a long time; Academic reports, public speeches, and stumbling, lack of confidence.[10]. These online expressions often have unique expressions and concise forms, which can quickly convey information and emotions, so they are particularly popular among young people. However, over-reliance on network language and lack of normative guidance may lead to people's rigid thinking, making it difficult to express their language normally after leaving the network environment. Because network language

usually pays attention to concise and direct expression, which affects people's language expression ability to some extent, especially it is in the face of formal occasions or when deep communication is needed.

1.3 The Question Raised

As an increasingly common phenomenon, college students' word poverty has aroused widespread concern and discussion from all walks of life. Word poverty not only affects individuals, but also has a certain negative impact on the whole society. As the backbone of society, college students' lack of language ability may affect the overall language environment of society. This paper will focus on the following issues: 1. Understand the current situation of word poverty; 2. Reveal the embodiment of word poverty in different aspects; 3. Explore the effects of word poverty; 4. Analyze the causes of word poverty; 5. Put forward suggestions to improve the phenomenon of word poverty.

1.4 Research Methods

1.4.1 Literature research method

Collect, classify and screen the documents related to the problem through China HowNet database and WeChat official account, and make an in-depth analysis of the sorted documents by using appropriate methods and tools, so as to integrate the cognitive status of "word poverty" and the relevant measures to reduce it, and provide theoretical basis for this paper.

1.4.2 Questionnaire survey method

After the pre-investigation of the questionnaire is reasonable, this study distributes the questionnaire by distributing leaflets on the test paper online and offline, obtains college students' understanding and suggestions on "word poverty", and collects a lot of data to provide basis for subsequent analysis.

1.4.3 Data analysis method

Statistical analysis of the collected questionnaire data reveals the manifestations, causes and influences of college students' "word poverty", which is convenient for the development of this paper.

1.5 Research and Innovation

1.5.1 The topic is novel, filling the research blank

In recent years, the phenomenon of word poverty has gradually aroused widespread social concern. Although we can often hear the word "word poverty", it has not been discussed in depth, especially in academic circles. It is difficult for us to find literature on the phenomenon of word poverty on HowNet, and it is more reported in newspapers, especially for college students. Therefore, by studying college students' cognition of word poverty and putting forward measures to improve word poverty, we can appropriately fill the research gap in this neighborhood and provide some basic data and theoretical support for the follow-up research.

1.5.2 The uniqueness, concreteness and development of the research object

Previous studies may pay more attention to the general population or specific occupational groups. As a special group, college students' language expression ability, vocabulary and cognition of word poverty may be different from other groups. Therefore, it is unique to study the current situation of college students' cognition of word poverty. Only for college students, the scope is smaller and more specific, and the implementation of measures may be more effective. Moreover, as a new force in society, college students' language expression ability has a great influence on the future social development and national competitiveness.

2 THE INVESTIGATION PLAN AND IMPLEMENTATION

2.1 Design of Research Scheme

2.1.1 Purpose of investigation

This paper hopes to find out the current situation of college students' language expression, especially whether there is the so-called "word poverty" phenomenon, as well as the universality and seriousness of this phenomenon, and to explore the possible reasons leading to the phenomenon of word poverty among college students, so as to formulate measures to effectively deal with and improve the phenomenon of word poverty among college students. In the education system, students can be encouraged to participate in debates and speech contests, and more standardized language can be encouraged to be used in online communication.

The purpose of this paper is as follows:

(1) To understand whether contemporary college students have experienced word poverty and its frequency.

(2) Understand the forms, causes and effects of contemporary college students' word poverty, explain the importance of studying the phenomenon of word poverty, and provide analytical basis for the topic.

(3) To understand the attitudes and views of contemporary college students towards the emergence of word poverty and their efforts to do so.

(4) In order to reduce the word poverty of college students, methods and measures are formulated through investigation and analysis to help college students improve their language expression ability.

2.1.2 Object of investigation

Most people think that the phenomenon of poor words is generally people with low academic qualifications, while those with academic qualifications speak like water. In fact, there are many contemporary college students who are poor in words, and there are many forms of expression. Based on this background, we take college students at different stages as the survey objects, which involve junior college students, undergraduates and graduate students.

2.1.3 Survey content

(1) The survey object is limited to college students; (2) Basic information involves: gender, age, education background and subject category; (3) To investigate the degree of college students' understanding of word poverty and its manifestations; (4) Understand their frequency of word poverty and their emotional and psychological reactions; (5) Analyze the main reasons and influences of college students' word poverty; (6) Investigate college students' coping strategies and measures to reduce word poverty.

2.1.4 Investigation method

(1)Pre-investigation

Before the formal investigation, we distributed pre-questionnaires in the circle of friends, QQ space and private chat with friends, and collected a total of 95 questionnaires in two days. The reliability and validity of these questionnaires were tested by SPSS, aiming at modifying the questionnaires and ensuring the rationality of the questionnaires for subsequent development.

(2)random sampling

Simple random sampling can be divided into two types: sampling with replacement and sampling without replacement. For the sampling with return, each individual drawn is returned to the population, and each individual may be drawn more than once. For non-return sampling, it will not be returned after drawing, and each individual can only be drawn once. Select the key survey method to extract individual samples from each key unit, including Guilin, Nanning, Liuzhou and Yulin.

(3)major investigation

Key survey refers to an incomplete survey method that only some key units (here refers to regions) are selected for investigation, so as to obtain data. Although the key units only account for a small part of all the respondents, the number of survey marks accounts for a large proportion, so the data obtained from the survey of these key units can be used to reflect the basic trend of economic changes. Through the key units shown in the questionnaire, this paper selects college students from Guilin, Nanning, Liuzhou and Yulin in Guangxi as the research objects to study the related problems of college students' "word poverty".

2.1.5 Investigation scheme design

(1)Determination of sample frame

According to the results of our questionnaire, the sample frame is composed of college students in Guilin, Nanning, Liuzhou and Yulin, where the questionnaire groups are concentrated.

(2)Determination of sample size

Because the survey of college students' status is located in many cities, in order to reduce the consumption of human, material and financial resources, this survey is conducted by simple random sampling. Then, from the sample size formula (1) in simple random sampling

$$n = \frac{Z_{\alpha/2}^{\ 2}}{d^2} S^2 \qquad (1)$$

In equation 2.1, n is the sample size. Under the guarantee of 95% confidence level, $Z_{\alpha/2} = 1.96$, d =0.05 is absolutely wrong, and S^2 stands for population variance. Through the questionnaire collected this time, it is easy to know our sample. The difference s^2 is 0.6019, so is the sample variance s^2 instead of population variance, and the sample size calculated by formula (1) is 925.

3 THE QUESTIONNAIRE DESIGN

By consulting and reading a large number of relevant data and documents in the early stage, a questionnaire is designed for the investigation and analysis of the current situation of college students' word poverty and its countermeasures. The topic content distribution of the questionnaire is shown in the following table 1.

Questionnair Section	Explain
A. Basic information	Some basic information of college students' interviewees, such as gender, age, education and subject category (questions 1-4)
B. The current situation of word poverty	College students' understanding of word poverty, the frequency of word poverty, the emotional distress caused by word poverty to college students, and their views on word poverty (questions 5-8)
C. the trouble of poor words	The influence of poor vocabulary on college students' study, life and future development (questions 10-12)
D. Influencing factors of word poverty	What do college students think is the source of word poverty (question 13)

E. Countermeasures for word poverty	The countermeasures of college students' respondents to the phenomenon of word poverty, such as personal aspects, social aspects, education system and network environment (questions 14-18)

Table 1 Five Sections of the Questionnaire

Because some interviewees have the frequency of "never", in order to prevent logical errors in the questions answered by interviewees, this questionnaire has set a skip question in question 6.

4 PRE-INVESTIGATION

4.1 the Implementation Process of Pre-Investigation

Under normal circumstances, in order to ensure the overall reliability of the questionnaire and the smooth progress of the formal investigation, in the formal adjustment before the investigation, it is necessary to conduct a small-scale pre-investigation. According to the analysis of the results of the pre-investigation, mistakes in the questionnaire can be found in time and revised appropriately, and the options with repeated meanings can be deleted. If necessary, the questions can be supplemented appropriately. The pre-survey selected 95 college students from some colleges and universities in Guangxi. For the college students who are close to each other, paper questionnaires were used, while the college students who are far away switched to online questionnaires issued by the Questionnaires platform.

4.2 Reliability Test

Our team will use Cronbach reliability coefficient to measure the internal consistency coefficient of each item in the questionnaire. The specific calculation formula of Cronbach reliability coefficient is:

$$\alpha = \frac{\mathrm{K}}{K-1}\left(1-\frac{\sum S_i^{\,2}}{S_x^{\,2}}\right) \qquad (2)$$

Where is the reliability coefficient, which usually ranges from [0,1], and the closer it is to 1, the more the questionnaire is tested. The greater the reliability of the quantity, the closer x it is to 0, the smaller the reliability. K is the number of test i items in the scale, $\mathrm{S}_i^{\,2}$ Represents the variation of scores of all subjects on question I, and $\mathrm{S}_x^{\,2}$ is the variance of total s cores of all subjects.

The reliability of five questions in this questionnaire (question 5, question 10, question 11, question 13 and question 14) is tested and analyzed by SPSS software, and the Cronbach coefficient of each questionnaire is shown in the following table 2:

Table 2 Cronbach Coefficient Table of Pre-questionnaire Scale

Item	Cronbach coefficient	number of terms	Reliability evaluation
Question 5	0.825	five	better
Question 10	0.925	eight	good
Question 11	0.905	five	good
Question 13	0.833	six	better
Question 14	0.802	five	better

As can be seen from the above table, the Cronbach coefficient of each scale item in the questionnaire is greater than 0.80, and the Cronbach coefficient of the whole scale is 0.895, so the reliability of the questionnaire design is considered to be good.

4.3 Validity Test

There are many questions in this questionnaire, so structural validity is chosen to test the data collected in the pre-survey. The specific KMO and Bartlett test coefficients are shown in Table 3 below.

Table 3 KMO and Bartlett Test Coefficient Table of Pre-questionnaire Scale

Item	KMO coefficient	Approximate chi-square	significance
Question 5	0.725	199.267	0.000
Question 10	0.943	1029.123	0.000
Question 11	0.899	651.798	0.000
Question 13	0.765	221.199	0.000
Question 14	0.742	167.496	0.000

Generally speaking, the KMO value is between [0,1]. When the KMO value is closer to 1, it means that the correlation between variables is stronger, and vice versa. According to the results of the above coefficient table, it is easy to know that the correlation coefficients between items under each scale are all greater than 0.7, and some of them are as high as

0.8-0.9, which shows that the questionnaire design is effective.

4.4 The revision of the topic

Through the pre-survey, the most authentic feedback from the surrounding college students who participated in the pre-survey questionnaire was collected, and the overall content, typesetting and topic distribution of the questionnaire were modified and optimized according to their feedback. The specific contents of the questionnaire are as follows:
【Question 1】Redundancy in expressing the degree of scale questions-changing the number 1-5 from "never" to "always" to the number 1-4, which are never, sometimes, often and always respectively.
【Question 2】What do you think are the causes of word poverty? In question 13, the first item "Internet usage habits (such as social media)" and the second item "Internet language and expression pack usage" are similar, resulting in redundancy, so they are merged into one item and changed to "Internet usage (such as network language and expression pack usage)".

5 FORMAL INVESTIGATION

With the completion of the pre-investigation, the collection results of 95 questionnaires distributed will be integrated and all of them will be passed the test of reliability and validity, and made a reasonable revision to the questions of the questionnaire. Therefore, according to the formula for calculating the sample size of simple random sampling in market research and considering the coverage of survey information, the final sample size required for formal survey is 925.

5.1 the Specific Distribution and Recycling of Questionnaires

Table 4 Distribution Table of Questionnaires in Guangxi

city	Number of questionnaires issued (copies)	Proportion (%)
Guilin	155	34.27
Nanning	189	24.97
Liuzhou	241	21.19
Qinzhou	187	11.35

As shown in table 4, in the formal investigation, we actually sent 925 questionnaires, and 772 valid questionnaires were recovered, with an effective recovery rate of 83.46%.

5.2 Reliability Test

According to the reliability test of 95 questionnaires collected in the pre-investigation, the actual situation of the formal investigation is analyzed. The reliability of 925 questionnaires was tested again. By analyzing five scales, it is easy to know that the Cronbach coefficient of each scale is greater than 0.8, and it is considered that the internal reliability of the questionnaire is ideal. Cronbach coefficient of each scale is shown in Table 5:

Table 5 Cronbach Coefficient Table of Formal Questionnaire

Item	Cronbach coefficient	number of terms	Reliability evaluation
Question 5	0.868	five	better
Question 10	0.972	eight	good
Question 11	0.949	five	good
Question 13	0.889	six	better
Question 14	0.873	five	better

5.3 Validity Test

Referring to the steps of pre-investigation, the validity of 925 questionnaires was tested again, and each quantity was obtained. The KMO coefficients of the questions are all greater than 0.8, and the P values are all 0.000, which shows that it is very suitable for factor analysis, so the questionnaire structure is well designed. The specific KMO and Bartlett test coefficients under each scale are shown in Table 6 below:

Table 6 KMO and Bartlett test coefficients of formal questionnaire

Item	KMO coefficient	Approximate chi-square	significance
Question 5	0.872	2024.818	0.000
Question 10	0.976	16348.894	0.000
Question 11	0.931	7219.097	0.000
Question 13	0.910	2699.033	0.000
Question 14	0.878	2095.517	0.000

5.4 Data Processing

925 questionnaires were imported into Excel through the platform of Questionnaires, downloaded according to the option serial number, and the data were sorted in Excel. For the processing of abnormal data and missing values in this questionnaire, the specific steps are as follows:

5.4.1 Abnormal data processing

(1)It takes an unusual time to fill in the questionnaire.
A total of 18 questions were designed in the questionnaire, and the time spent by college students to fill in the questionnaire was unified. Calculation and analysis, descriptive statistics as shown in table 7:

Table 7 Descriptive Statistics of Respondents' Time-consuming Questionnaire Filling

Statistic	result	unit	Statistic	result	unit
Sample size	925	share	mode	160	second
average value	137.1	second	minimum value	49	second
Standard deviation	50.574	second	maximum	1038	second
median	133	second	Decimal quantile	95	second

It is calculated that the average time for each student to fill in the questionnaire is 189.09 seconds, which can be seen from the above table 7. It is known that the average time for respondents to fill in the questionnaire is 137.1 seconds, and the time for most respondents to fill in the questionnaire is controlled between 86.526 seconds and 187.674 seconds (137.1+50.574). The long decile for college students to fill in the questionnaire is 95 seconds, and the difference between the modes and their median is not big, so we have reason to believe that students should take no less than 100 seconds for student respondents to fill in the questionnaire rationally and carefully, so we set the questionnaires that take less than 100 seconds to fill in the questionnaire are regarded as invalid questionnaires and are rejected.
(2)Logical conflict
Deal with the questionnaires that take a long time to fill in and have logical conflicts, export the data to Excel and use it to eliminate them. Therefore, through the overall screening of questionnaires, 772 valid questionnaires were finally recovered from 925 questionnaires, and the effective rate of questionnaire recovery reached 83.46%.
(3)Treatment of missing values
This questionnaire is mainly distributed through the quiz star network platform, and each question is set as a mandatory question. If you don't answer, you can't continue to answer the next question or submit the questionnaire. This questionnaire is set with jump questions. According to the default export rules of the test paper star platform, the option value of the questions that college students jump (unnecessary to fill in) is marked as -3. On this basis, the marked value in the jump questions is checked by using the Data-SelectCases method in SPSS software, and it is found that the missing value of this questionnaire does not exist.

5.4.2 Investigate the overall situation

The basic situation of the sample is: the number of college students in the survey sample is 925, among which the abnormal value is 153 students, and the actual number of effective surveys is 772. Among the 772 college students surveyed, male college students account for 46.63% and female college students account for 53.37%, and the ratio of male to female is relatively balanced. 31.22% are junior college students, 40.15% are undergraduate students and 28.63% are graduate students. The largest proportion is undergraduate; The liberal arts account for 35.36%, the science accounts for 32.25%, the engineering accounts for 32.38%, and the others account for a very small proportion, which can be ignored; Guilin accounts for 20.08%, Nanning for 24.43%, Liuzhou for 31.22% and Yulin for 24.22%. The specific distribution of basic information is shown in Table 8:

Table 8 Distribution of Basic Information of College Students

	classify	Response analysis		Percentage of cases
		Respose number	Response percentage	
Gender	Man	360	11.66%	46.63%
	Woman	412	13.34%	53.37%
Academic Degree	College for professional training	241	7.80%	31.22%
	Undergraduate course	310	10.04%	40.15%
	Postgraduate	221	7.16%	28.63%
Subject category	Liberal arts	273	8.84%	35.36%
	Science	249	8.06%	32.25%
	Field of engineering	250	8.10%	32.38%
Location	Guilin	155	5.02%	20.08%
	Nanning	187	6.06%	24.43%
	Liuzhou	241	7.80%	31.22%
	Yulin	189	6.12%	24.22%
Total		3088	100%	100%

6 DATA ANALYSIS OF SURVEY RESULTS

6.1 Descriptive Statistical Analysis

6.1.1 Sex distribution

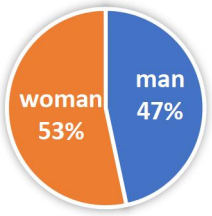


Figure 1 Gender Distribution Map

As shown in Figure 1 and 2, among the college students surveyed, the number of boys is 360, accounting for 47%, and the number of girls is 412, accounting for 53%. The ratio of male to female is about 1:1.

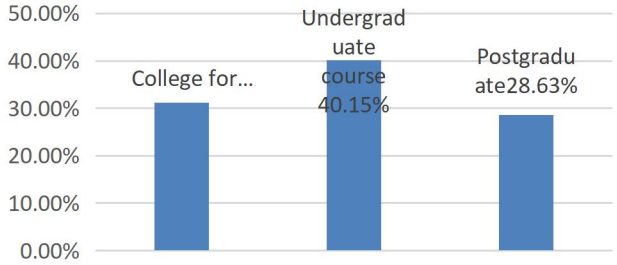


Figure 2 Education Distribution Map

6.1.2 Educational background distribution

Among the college students surveyed, there are 241 junior college students, accounting for 31.22%, 310 undergraduate students, accounting for 40.15%, and 221 graduate students, accounting for 28.63%. Among the people who participated in the questionnaire survey, the number of undergraduates was the largest, and the number of postgraduates was the least.

6.1.3 Regional distribution

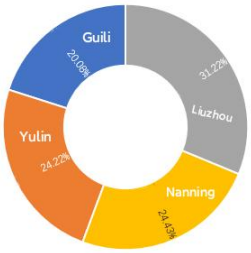


Figure 3 Regional Distribution Map

As shown in Figure 3, the proportion of college students in Guilin is 20.08%, accounting for the least; The proportion of college students in Yulin and Nanning is 24.22% and 24.43% respectively. The proportion of college students in Liuzhou is 31.22%, accounting for the largest proportion.

6.1.4 Subject category distribution

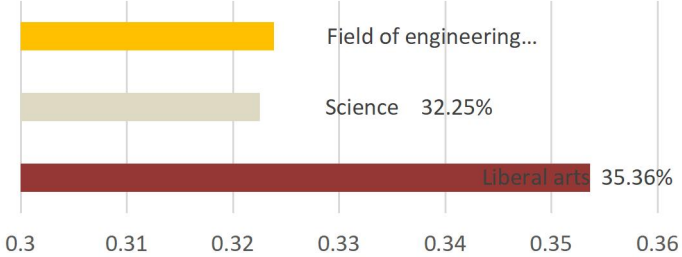


Figure 4 Distribution Map of Discipline Categories

In figure 4, the proportion of liberal arts students, science students and engineering students is 35%, 32% and 33% respectively, and the proportion of liberal arts students, science students and engineering students is about 1:1:1. Since only three people fill in other disciplines, we will not discuss them here.

6.1.5 Distribution of college students' understanding of word poverty

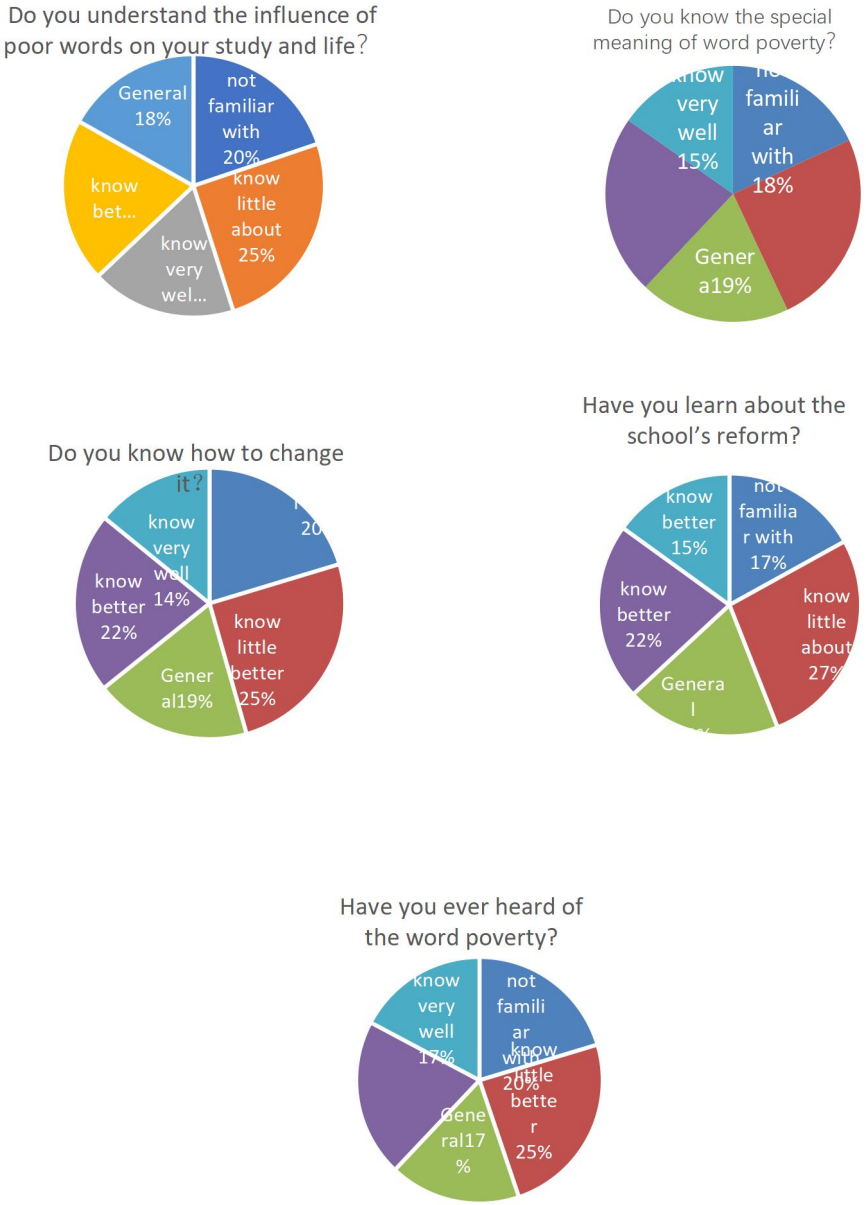


Figure 5 Distribution Map of Word Poverty Understanding

As can be seen from the distribution maps of understanding in Figure 5, 20% and 54% of college students are ignorant and relatively familiar with the word "word poverty"; 18% did not know the specific meaning of word poverty and 57% did. The impact of poor understanding of words on study and life and the relative understanding are 20% and 55% respectively; 20% and 55% did not know how to improve their own word poverty and relatively understood it, respectively; I didn't know the measures given by the school to improve the phenomenon of word poverty, and the relative understanding was 17% and 56% respectively. On the whole, we can see that college students still have a certain understanding of word poverty and pay more attention to the phenomenon of word poverty.

6.1.6 Distribution of the frequency of college students' word poverty

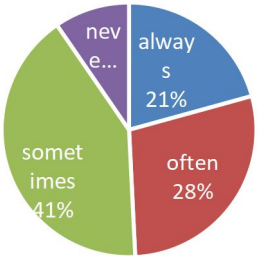


Figure 6 Distribution Diagram of Word Limited Frequency

In figure 6, among the college students surveyed, 10% never have word poverty, accounting for the least; There are 41% college students who sometimes have word poverty, accounting for the most; There are 28% college students who often have word poverty, and 21% college students always have word poverty. It can be seen that among college students, most people will have the phenomenon of word poverty.

6.1.7 Distribution of college students' poor expressions

The vertical axis in Figure 7 is from top to bottom, which is used for psychological stress, expressing complex feelings, discussing unfamiliar topics, using professional terms, communicating with strangers, answering questions quickly, academic writing and academic speech. And green means always, yellow means often, red means sometimes, and blue means never. As can be easily seen from Figure7 below, word poverty always occurs in academic writing, followed by discussing unfamiliar topics; The most common form of expression is academic speech, followed by psychological pressure; It can be seen that the manifestations of word poverty include all the situations listed in the table.

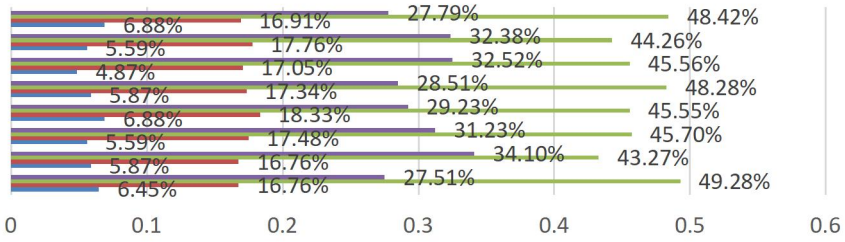


Figure 7 Distribution Map of Word-limited Expression Forms

6.1.8 The distribution of the causes of word poverty

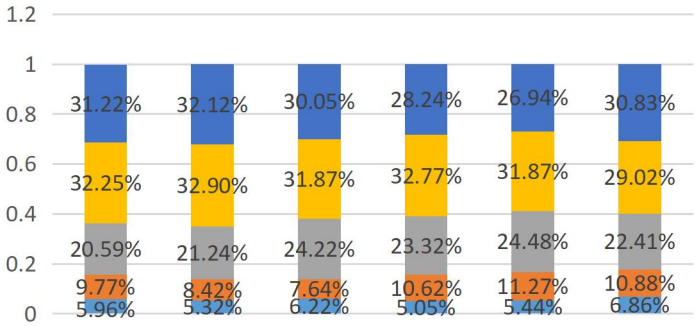


Figure 8 Distribution of Causes of Word Shortage

The horizontal axis in Figure 8 shows, from left to right, network usage habits, less reading, fragmented information browsing, face-to-face communication, less participation in exchange activities organized by schools, and great psychological pressure. Dark blue, yellow, gray, red and blue represent great influence, great influence, medium influence, slight influence and no influence in turn. As can be seen from Figure 8, the causes of college students' word poverty are internet use habits, less reading, fragmented browsing of information, less face-to-face communication, less participation in exchange activities organized by schools and great psychological pressure, among which the main reasons are less reading, less internet use habits and less face-to-face communication.

6.1.9 The influence distribution caused by word poverty

(1)The influence of poor words on college students' academic and social life.

The horizontal axis in Figure 9 represents no impact, slight impact, moderate impact, great impact and great impact from left to right. The colors in the picture from light to dark represent academic performance, class discussion participation, social activities, self-expression ability and personal self-confidence. As can be seen from Figure 9 below,

most college students feel that poor words will affect their academic performance, class participation, social activities, self-expression ability and personal confidence. Among the great influences, college students think that the greatest influence is academic performance, followed by social activities; Among the non-influences, college students think that the least influential is the participation in class discussion.

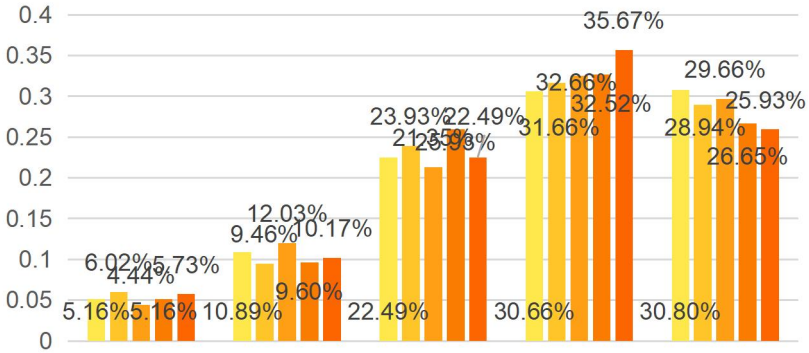


Figure 9 Distribution Diagram of the Influence of Word Poverty on Academic Social Aspects

(2) The influence of word poverty on college students' personal development

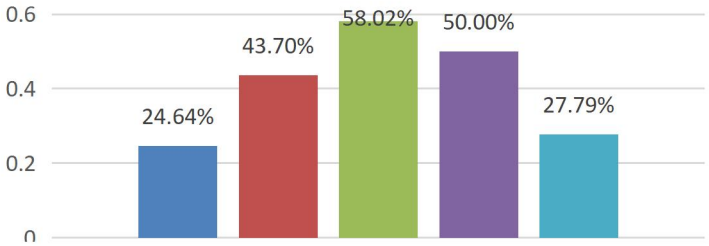


Figure 10 Distribution Map of the Influence of Word Poverty on Personal Development

From left to right, the horizontal axis in Figure 10 represents in turn that career choices are restricted, academic performance is affected, social activities are reduced, personal interests are changed, and self-evaluation is triggered. Talking about the influence of poor words on personal development, as shown in Figure 10 below, 58.02% of college students think that poor words have the greatest influence on personal development, which is to reduce my social activities, and secondly to change my personal interests, but also to career choice, academic performance and self-re-evaluation.

6.1.10 Distribution of measures to reduce word poverty

(1)Personal measures to reduce word poverty

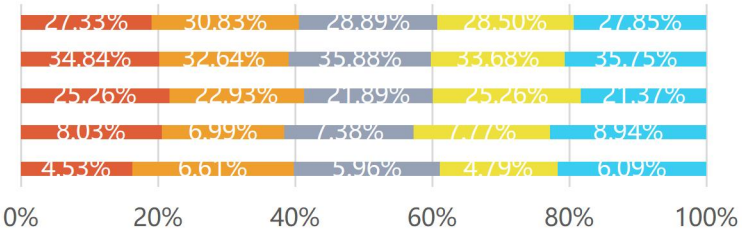


Figure 11 Distribution of Personal Measures to Improve Word Poverty

Red, orange, gray, yellow and blue in the horizntal axis of Figure 11 represent regular oral training, expanding reading range, learning and using new vocabulary, participating in public speeches and debates, and conducting emotional and stress training. As can be seen from Figure 11, among the methods that college students think can effectively reduce word poverty, expanding reading range is the most effective, followed by learning and using new vocabulary. But generally speaking, regular participation in oral expression training, public speaking and emotional management are also good ways to reduce word poverty. If a person often suffers from word poverty, he can improve it through these methods.

(2)Measures taken by the education system to reduce word poverty

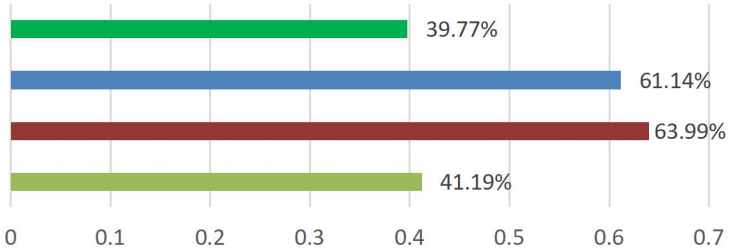


Figure12 Distribution Map of Measures Taken by the Education System to Improve Word Poverty

The green, blue, red and orange in the vertical axis of Figure 12 represent the introduction of multicultural education, the provision of psychological consultation and counseling, the encouragement of students to participate in debates and speeches, and the increase of language expression related courses.In view of the phenomenon of students' poor vocabulary, it is necessary for the education system to give some solutions. As can be seen from Figure 12, schools can introduce multicultural education, provide psychological counseling and counseling services, encourage and support students to participate in debates and speech contests, and increase language expression-related courses. The proportion of participating in debates and speech contests is 63.99%, providing psychological counseling and counseling services is 61.14%, increasing language expression-related courses is 41.19%, and introducing multiculturalism is 39.77%.
(3)Measures to reduce word poverty in network environment

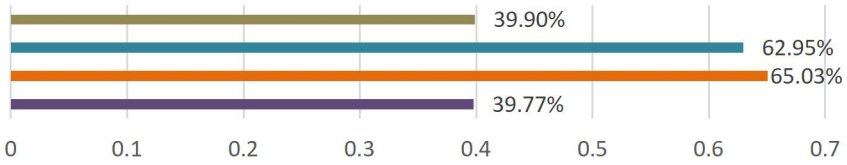


Figure 13 Distribution Map of Measures for Improving Word Poverty in Network Environment

Gray, blue-green, red and green in the vertical axis of Figure 13 represent encouraging the use of standardized language in network communication, increasing education on the differences between network and actual language, providing training on network communication skills, and encouraging practical communication in turn. Relying too much on network language and habits will also lead to poor words. As can be seen from Figure 13, among the network measures that college students think can effectively reduce word poverty, 39.9% are encouraged to use more standardized language in network communication, 62.95% are educated about the differences between network language and actual language, 65.03% are provided with network communication skills training, and 39.77% are encouraged to communicate in real life. Among them, the most recommended is to provide network communication skills training.

6.2 Difference Analysis

6.2.1 An analysis of the difference in the frequency of college students' word poverty

(1)An analysis of the differences in the frequency of word poverty a mong college students of different sexes
In order to explore the gender difference in the frequency of college students' word poverty, the data were tested by independent sample t, and the test results are shown in the following table 9.

Table 9 Difference Analysis Table of the Frequency of Word Poverty in Different Sexes

	N (sample size)	M (average)	SD (standard deviation)	t	P
man	360	2.4	0.92	0.102	0.919
woman	412	2.39	0.92		

From the above table, P=0.919>0.05, that is, there is no statistical difference in the frequency of word poverty among co llege students of different sexes, and it is considered that the frequency of word poverty among male and female student s is almost the same.
(2)An analysis of the differences in the frequency of college students' word poverty in different regions
In order to explore the regional differences in the frequency of college students' word poverty, the data were analyzed

by one-way ANOVA, and the test results are shown in the following table 10.

Table 10 Difference Analysis Table of the Frequency of Word Poverty in Different Regions

	N (sample size)	M (average)	SD (standard deviation)	F	P
Guilin	155	2.37	0.92	0.51	0.676
Yulin	187	2.41	0.97		
Liuzhou	241	2.44	0.93		
Nanning	189	2.34	0.86		

From Table 10, P=0.676>0.05, that is, there is no statistical difference in the frequency of word poverty among college students in different regions, and it is considered that there is little difference in the frequency of word poverty among college students in different regions.

(3)An analysis of the differences in the frequency of word poverty among college students with different academic quali fications

In order to explore the difference of the frequency of college students' word poverty in academic dimension, the data we re analyzed by one-way ANOVA, and the test results are shown in the following table 11.

Table 11 Difference Analysis Table of the Frequency of Word Poverty in Different Education

	N(sample size)	M(average)	SD(standard deviation)	F	P
College for professional training	241	2.12	0.96	21.74	0.000
undergraduate course	310	2.41	0.89		
postgraduate	221	2.67	0.83		

From the above table 11, it can be seen that P=0.000<0.05, that is, there is a statistical difference in the frequency of word poverty among college students with different academic qualifications, and it is considered that college students have more frequent word poverty.

(4)Analysis on the difference of the frequency of word poverty among college students in different disciplines

In order to explore the differences in the frequency of word poverty among college students in different subject categories in the dimension of subject categories, the data were analyzed by one-way variance analysis. Among them, because the number of questionnaires collected in other subject categories, such as sports and arts, was very small, it was not analyzed. The specific test results are shown in Table 12 below.

Table 12 Analysis on the Difference of Word-poor Frequency in Different Subjects

	N (sample size)	M (average)	SD (standard deviation)	F	P
liberal arts	273	2.8	0.86	48.39	0.00
science	249	2.26	0.87		
field of engineering	250	2.09	0.87		

From the above table 12, we can see that P=0.000<0.05, that is, there is a statistical difference in the frequency of word poverty among college students in different disciplines, and it is considered that there is no statistical difference between science and engineering, but there is a difference between science and liberal arts.

6.2.2 An analysis of the differences of college students' understanding of the phenomenon of word poverty

Because college students' understanding of word poverty is a scale question, five dimensions are used to measure their understanding. In order to explore the differences of college students' understanding of word poverty in gender, region, education and subject category, pearson chi-square test is carried out on the data respectively, and the specific test results are shown in the following tables.

(1) An analysis of the differences in the understanding of word poverty among college students of different sexes.

Table 13 Difference Analysis Table of Different Genders' Understanding of Word Poverty

Dimension	Degree of understanding	Gender		X^2	P
		Man	Woman		
Have you ever heard of the word poor?	not familiar with	73	84	0.052	1
	know little about	89	100		

	General understanding	62	71		
	Know better	75	85		
	Know very well	61	72		
Do you know the specific meaning of word poverty ?	not familiar with	70	70	4.406	0.354
	know little about	96	96		
	General understanding	66	81		
	Know better	71	104		
	Know very well	57	61		
Do you understand the influence of poor words on your study life ?	not familiar with	70	83	3.982	0.408
	know little about	84	111		
	General understanding	70	68		
	Know better	80	76		
	Know very well	56	74		
Have you ever learned how to improve your word poverty?	not familiar with	78	79	2.313	0.678
	know little about	93	102		
	General understanding	64	80		
	Know better	80	87		
	Know very well	45	64		
Have you ever learned about the measures given by the school to improve the phenomenon of word poverty?	not familiar with	70	61	5.546	0.236
	know little about	88	121		
	General understanding	69	78		
	Know better	74	95		
	Know very well	59	57		

From the above table 13, we can see that there is no statistical difference in the understanding of word poverty among college students of different sexes in the five dimensions, that is, there is no difference in the understanding of word poverty among college students of different sexes.

(2) An analysis of the differences of college students' understanding of word poverty in different regions

Table 14 Difference Analysis Table of Understanding Degree of Word Poverty in Different Regions

Dimension	Degree of understanding	Location				X^2	P
		Guilin	Yulin	Liuzhou	Nanning		
Have you ever heard of the word poor?	not familiar with	25	42	58	32	41.901	0.00***
	know little about	30	48	67	44		
	General understanding	22	22	34	55		
	Know better	37	43	44	36		
	Know very well	41	32	38	22		
Do you know the specific meaning of	not familiar with	27	40	44	29	34.721	0.001**
	know little about	37	60	52	43		

word poverty ?	General understanding	23	24	42	58		
	Know better	42	38	56	39		
	Know very well	26	25	47	20		
Do you understand the influence of poor words on your study life ?	not familiar with	39	36	44	34	22.026	0.037**
	know little about	39	52	56	48		
	General understanding	25	28	37	48		
	Know better	28	33	55	40		
	Know very well	24	38	49	19		
Have you ever learned how to improve your word poverty?	not familiar with	33	41	48	35	31.058	0.002***
	know little about	44	53	55	43		
	General understanding	26	26	39	53		
	Know better	21	42	61	43		
	Know very well	31	25	38	15		
Have you ever learned about the measures given by the school to improve the phenomenon of word poverty?	not familiar with	22	41	42	26	13.45	0.337
	know little about	41	47	56	65		
	General understanding	33	31	47	36		
	Know better	32	40	58	39		
	Know very well	27	28	38	23		

From the above table 14, we can see that in four of the five dimensions to measure the understanding of word poverty, the test P value of college students' understanding of word poverty in different regions is less than 0.05, but there is still one dimension with the test P value of college students' understanding of word poverty in different regions greater than 0.05. It can be considered that there are some statistical differences in the understanding of word poverty among college students in different regions, that is, there are some differences in the understanding of word poverty among college students in different regions, among which Yulin and Liuzhou have the highest rate of ignorance.

(3) An analysis of the differences in the understanding of word poverty among college students with different academic qualifications

Table 15 Difference analysis table of understanding degree of word poverty with different educational background

Dimension	Degree of understanding	Degree			X^2	P
		college for professional training	undergraduate course	postgraduate		
Have you ever heard of the word poor?	not familiar with	6	6	3	14.097	0.079*
	know little about	9	4	6		
	General understanding	3	12	5		
	Know better	3	15	5		
	Know very well	4	11	8		
Do you know the specific meaning of word poverty ?	not familiar with	4	7	2	12.533	0.129
	know little about	11	10	4		

	General understanding	3	13	5		
	Know better	3	11	7		
	Know very well	4	7	9		
Do you understand the influence of poor words on your study life ?	not familiar with	8	10	4	9.384	0.311
	know little about	9	10	6		
	General understanding	2	12	4		
	Know better	4	11	7		
	Know very well	2	5	6		
Have you ever learned how to improve your word poverty?	not familiar with	11	8	3	16.203	0.040**
	know little about	8	16	7		
	General understanding	2	7	7		
	Know better	1	5	6		
	Know very well	3	12	4		
Have you ever learned about the measures given by the school to improve the phenomenon of word poverty?	not familiar with	3	5	2	5.881	0.661
	know little about	9	12	7		
	General understanding	4	13	8		
	Know better	4	14	5		
	Know very well	5	4	5		

From the above table 15, it can be seen that the test P values of college students with different academic qualifications are almost all greater than 0.05, so there is no statistical difference in general, that is, there is no difference in the understanding of college students with different academic qualifications.

(4)An analysis of the differences of college students' understanding of the phenomenon of word poverty in different disciplines

Table 16 Difference analysis table of different disciplines' understanding of word poverty

Dimension	Degree of understanding	Subject			X^2	P
		liberal arts	Science	field of engineering		
Have you ever heard of the word poor?	not familiar with	3	2	10	47.522	0.000***
	know little about	0	6	13		
	General understanding	3	14	3		
	Know better	12	6	5		
	Know very well	15	5	3		
Do you know the specific meaning of word poverty ?	not familiar with	3	1	9	46.147	0.000***
	know little about	2	9	14		
	General understanding	3	15	3		
	Know better	12	3	6		
	Know very well	13	5	2		

Do you understand the influence of poor words on your study life ?	not familiar with	5	2	15	70.712	0.000***
	know little about	0	11	14		
	General understanding	2	14	2		
	Know better	18	2	2		
	Know very well	8	4	1		
Have you ever learned how to improve your word poverty?	not familiar with	5	4	13	70.23	0.000***
	know little about	2	13	16		
	General understanding	1	14	1		
	Know better	9	1	2		
	Know very well	16	1	2		
Have you ever learned about the measures given by the school to improve the phenomenon of word poverty?	not familiar with	1	1	8	58.579	0.000***
	know little about	3	9	16		
	General understanding	4	17	4		
	Know better	12	5	6		
	Know very well	13	1	0		

From the Above table16, it can be seen that the test P value of college students' understanding of word poverty in different disciplines is less than 0.05, so it is considered that there is a statistical difference in general, that is, there are differences in the understanding of word poverty among college students in different disciplines, of which engineering has the lowest understanding of word poverty and liberal arts has a relatively high understanding.

6.2.3 An analysis of the differences of college students' attitudes towards the phenomenon of word poverty in writing and communication

(1)An analysis of the differences in attitudes of college students of different sexes towards the phenomenon of word poverty

Table 17 Analysis Table of Differences between Different Genders' Attitudes towards Word Poverty

Gender	Think it's normal, don't worry too much.	Will affect others' evaluation of their professional ability.	Feeling frustrated, I hope there are ways to improve.	Try to find other ways to supplement the lack of language expression.	other	Total
man	116	111	46	52	0	325
woman	130	119	64	60	0	373
total	246	230	110	112	0	698

As can be seen from Table 17, the data is subject to independent sample t-test. The results show that P=0.376>0.05, and there is no statistical difference in the attitudes of college students of different sexes towards the phenomenon of word poverty, which means that the attitudes of college students of different sexes towards the phenomenon of word poverty are basically the same.

(3) An analysis of the differences of college students' attitudes towards word poverty in different regions

Table 18 Analysis table of differences in attitudes towards word poverty in different regions

City	Think it's normal, don't worry too much.	Will affect others' evaluation of their professional ability.	Feeling frustrated, I hope there are ways to improve.	Try to find other ways to supplement the lack of language expression.	other	Total
Guilin	47	47	30	15	0	139
Yulin	62	51	22	25	0	160
Liuzhou	80	77	26	40	0	223
Nanning	57	55	32	32	0	176
Total	246	230	110	112	0	698

As can be seen from Table 18, the data is analyzed by one-way ANOVA. The result shows that P=0.028<0.05, and the attitudes of college students in different regions towards the phenomenon of word poverty are statistically different, that

is to say, college students in different regions have different attitudes towards the phenomenon of word poverty, and the frequency of college students in Liuzhou is higher.

(4) An analysis of the differences in attitudes of college students with different academic qualifications towards the phenomenon of word poverty

Table 19 Analysis table of differences in attitudes towards word poverty with different academic qualifications

academic degre	Think it's normal, don't worry too much.	Will affect others' evaluation of their professional ability.	Feeling frustrated, I hope there are ways to improve.	Try to find other ways to supplement the lack of language expression.	other	Total
college for professional training	84	71	34	34	0	223
undergraduate course	100	92	37	49	0	278
postgraduate	62	67	39	29	0	197
Total	246	230	110	112	0	698

As can be seen from Table 19, the data is analyzed by one-way ANOVA. P=0.786>0.05. There is no statistical difference in the attitudes of college students with different academic qualifications towards the phenomenon of word poverty, which means that the attitudes of college students with different academic qualifications towards the phenomenon of word poverty are almost the same.

(5) An analysis of the differences of college students' attitudes towards word poverty in different disciplines.

Table 20 Analysis table of differences in attitudes of different disciplines towards word poverty

subject category	Think it's normal, don't worry too much.	Will affect others' evaluation of their professional ability.	Feeling frustrated, I hope there are ways to improve.	Try to find other ways to supplement the lack of language expression.	other	Total
liberal arts	69	74	38	42	0	223
science	89	76	38	32	0	235
field of engineering	88	80	34	38	0	240
Total	246	230	110	112	0	698

As can be seen from Table 20, the data is analyzed by one-way ANOVA. P=0.000<0.05. There are statistical differences in the attitudes of college students in different disciplines towards the phenomenon of word poverty, that is to say, college students in different disciplines have different attitudes towards the phenomenon of word poverty, among which science and engineering students think that word poverty is a normal phenomenon more frequently than liberal arts students.

6.3 Multiple response analysis

In the questionnaire, the views on the measures to improve college students' vocabulary poverty are investigated in the form of multiple-choice questions from two aspects: educational means and network influencing factors. Therefore, when using SPSS to analyze the multiple-choice questions in the questionnaire, this paper adopts the method of frequency and crosstab in multiple response analysis.

Table 21 College Students' Views on Measures to Improve Word Poverty

Measures to improve word poverty	n	Response rate
Add courses related to language expression	318	9.96%
Encourage and support students to participate in debates and speech contests.	494	15.47%
Provide psychological counseling and counseling services to help students overcome anxiety.	472	14.78%
Introduce multicultural education to enhance students' linguistic diversity and cultural understanding.	307	9.61%
Restrict the use of social media and encourage communication in real life.	307	9.61%

Provide network communication skills training, including how to effectively express and understand network language.	502	15.72%
Increase education on the differences between online language and actual language.	486	15.22%
Encourage the use of more standardized language in network communication.	308	9.64%
Total	3194	100%

As can be seen from the above table 21, the goodness-of-fit test is significant $\chi 2 = 161.056$, $p=0.000<0.05$, which means that the selection ratio of each item is obviously different, and the differences can be specifically compared by response rate or penetration rate.

At the same time, the data shows that the proportion of encouraging and supporting students to participate in debates and speech contests is 15.47%, which is higher than other measures. This reflects a positive advocacy and practice, that is, to improve language expression ability and self-confidence by participating in practical activities. Providing psychological counseling and counseling services to help students overcome anxiety is 14.78%, which is also an aspect worthy of attention. In addition, the proportion of providing network communication skills training and encouraging the use of more standardized language in network communication is 15.72% and 9.64% respectively. This shows that the school attaches importance to adapting to the information age, realizes the difference between online language and actual language, and the necessity of cultivating students' good online communication habits.

In a word, improving college students' word poverty requires many efforts and measures. In addition to strengthening classroom teaching and practical activities, we should also pay attention to students' mental health, cultural literacy and network communication ability.

Cross-table analysis will be made between the views on measures to improve college students' word poverty and their basic information (gender, education level and subject category), and the proportion of college students' views on measures to improve word poverty with different basic conditions will be obtained.

First of all, the cross table of different academic qualifications and measures to improve word poverty is shown in the following table 22.

Table 22 Cross-table of views on measures to improve word poverty with different academic qualifications

Measures to improve word poverty	academic degree			Total （n=772）
	college for professional training（n=241）	undergraduate course（n=310）	postgraduate（n=221）	
Add courses related to language expression	47.30%	38.06%	38.91%	41.19%
Encourage and support students to participate in debates and speech contests.	65.98%	61.61%	65.16%	63.99%
Provide psychological counseling and counseling services to help students overcome anxiety.	56.43%	63.23%	63.35%	61.14%
Introduce multicultural education to enhance students' linguistic diversity and cultural understanding.	36.51%	42.58%	39.37%	39.77%
Restrict the use of social media and encourage communication in real life.	43.98%	37.10%	38.91%	39.77%
Provide network communication skills training, including how to effectively express and understand network language.	60.58%	65.16%	69.68%	65.03%
Increase education on the differences between online language and actual language.	64.73%	61.94%	62.44%	62.95%
Encourage the use of more standardized language in network communication.	40.25%	35.81%	45.25%	39.90%
Chi-square test： x 2=11.410 p=0.654				

Secondly, the cross table of different genders and measures to improve word poverty is shown in the following table 23.

Table 23 Cross Table of Gender and Measures to Improve Word Poverty

Measures to improve word poverty	Gender		Total（n=772）
	Female（n=412）	Male（n=360）	
Add courses related to language expression	41.50%	40.83%	41.19%
Encourage and support students to participate in debates and speech contests	63.83%	64.17%	63.99%
Provide psychological counseling and counseling services to help students overcome anxiety	60.68%	61.67%	61.14%
Introduce multicultural education to enhance students' linguistic diversity and cultural understanding	42.96%	36.11%	39.77%
Restrict the use of social media and encourage communication in real life	39.56%	40.00%	39.77%
Provide network communication skills training, including how to effectively express and Understand network language	64.81%	65.28%	65.03%
Increase education on the differences between online language and actual language	59.95%	66.39%	62.95%
Encourage the use of more standardized language in network communication	40.29%	39.44%	39.90%

Finally, the cross table of different subject categories and measures to improve word poverty is shown in the following table 24.

Table 24 Cross table of different subject and measures to improve word poverty

Measures to improve word poverty	Subject			Total（n=772）
	field of engineering（n=250）	liberal arts（n=273）	science（n=249）	
Add courses related to language expression	41.60%	41.39%	40.56%	41.19%
Encourage and support students to participate in debates and speech contests	66.00%	62.64%	63.45%	63.99%
Provide psychological counseling and counseling services to help students overcome anxiety	62.00%	63.74%	57.43%	61.14%
Introduce multicultural education to enhance students' linguistic diversity and cultural understanding	34.40%	41.39%	43.37%	39.77%
Restrict the use of social media and encourage communication in real life	45.60%	38.10%	35.74%	39.77%
Provide network communication skills training, including how to effectively express and Understand network language	64.00%	64.84%	66.27%	65.03%
Increase education on the differences between online language and actual language	59.20%	62.64%	67.07%	62.95%
Encourage the use of more standardized language in network communication	36.40%	41.03%	42.17%	39.90%

7 COGNITION AND INFLUENCE ANALYSIS OF "WORD POVERTY" BASED ON STRUCTURAL EQUATION MODELl

In this survey, the reliability of 772 valid data collected was tested, and the representative core factors were extracted by exploratory factor analysis, so as to build a structural equation model. Considering the relationship between the representative factors, the cognitive degree of college students in four places in Guangxi (Guilin, Nanning, Liuzhou and Yulin) and the influence degree caused by word poverty were analyzed.

7.1 Brief Introduction of Structural Equation Model

7.1.1 Concept of structural equation model

Structural equation model (SEM) is a method to establish, estimate and test causality model. The model contains both observable obvious variables and potential variables that cannot be directly observed. Structural equation model can replace multiple regression, path analysis, factor analysis, covariance analysis and other methods to clearly analyze the role of individual indicators on the population and the relationship between individual indicators. Compared with traditional analysis methods, structural equation model can explain as many variations of variables as possible while understanding the covariant relationship between variables. When you want to study the causal relationship between complex variables, it is most appropriate to use structural equation model.

The construction of structural equation model is usually divided into four main parts, including model hypothesis, reliability and validity test, model establishment and model revision. Generally speaking, the establishment of structural equation model first needs to determine the design of external latent variable factors and internal latent variable factors. If the established structural equation model has a good fitting effect, that is, it passes the evaluation criteria of model fitting, there is no need to modify the model.

7.1.2 Basic form of structural equation model

Structural equation model usually has two definitions: ① structural equation model = factor analysis+path analysis; ② Structural equation model = measurement model+structural model, and there is no difference in essence.

(1)Measurement model

The measurement model is confirmatory factor analysis (CFA), which describes the relationship between the observed variable and the latent variable, and measures the measurement effect of the explicit variable (that is, the measuring tool) on the latent variable (the magnitude and significance of the load of the observed variable on the latent variable). When doing validity analysis, what we usually expect is to find that the observed variables are significantly loaded on theoretically related latent variables, but not on unrelated latent variables by confirmatory factor analysis. The specific model expression is:

$$\begin{cases} x = \Lambda_x + \delta \\ y = \Lambda_y + \varepsilon \end{cases} \quad (3)$$

Where x is a vector composed of exogenous indicators, y is a vector composed of endogenous indicators, Λ_x and Λ_y are factor load matrices, ε and δ are error terms.

(2)structural model

$$\eta = B\eta + \Gamma\xi + \zeta \quad (4)$$

Where B is the coefficient matrix of sum for η and η, which also becomes the path coefficient; Γ represents the coefficient matrix between η and η, and ζ represents the error term.

7.2 the establishment of structural equation model

7.2.1 factorial analysis

Before establishing the structural equation model, it is necessary to use SPSSAU online software to conduct exploratory factor analysis on all the valid data to be investigated, so as to explore the influencing factors of college students on the consequences of word poverty, find out the potential variables in the theory, reduce the number of questions, and get better variables.

First of all, KMO test and Bartlett spherical test are needed. The KMO value in the test results is 0.928, the Bartlett spherical test value is 25,548.947, and the P value is less than 0.05. The output four factors explain that the cumulative percentage of 24 variables is 80.01%, so it can be considered that there is a certain correlation between the items. The specific KMO and Bartlett spherical test analysis results are shown in the following table 25:

Table 25 KMO and Bartlett Spherical Test Analysis Results

project	numerical value

KMO sampling suitability quantity	0.928
Approximate chi-square of Bartlett spherical test	25548.947
freedom	276
significance	0.000

7.2.2 Setting of observation variables

According to the rotated factor load matrix, four factors are defined, which are: the understanding degree of word poverty, the form of word poverty, the consequences caused by word poverty and the reasons for word poverty. These four factors are latent variables, and their corresponding observed variables are shown in the following table 25:

Table 26 System Indicators of Structural Equation Model

Latent variable	Observed variable	Item
Understanding degree of word poverty	Have heard of the word poor	A1
	Understand the specific meaning of word poverty	A2
	Understand the influence of poor words on study and life	A3
	Understand the ways to improve word poverty	A4
	Understand the school's measures to improve word povert	A5
The form of expression of word poverty	Academic speech	B1
	Academic writing	B2
	Answer questions quickly	B3
	Communicate with strangers	B4
	Use technical terms to express opinions	B5
	Discuss unfamiliar topics	B6
	Express complex feelings or opinions	B7
	has psychological pressure	B8
The consequences of poor words	Decline in academic performance	C1
	Talking about the low degree of discussion	C2
	Low participation in social activities	C3
	Decline in self-expression ability	C4
	Personal self-confidence is frustrated	C5
The causes of the formation of word poverty	Internet usage habits	D1
	Less reading, weakened expression ability	D2
	fragmented browsing information, lacking systematic thinking	D3
	Less face-to-face communication	D4
	Participation in school organization and exchange activities is less	D5
	under great psychological pressure	D6

7.2.3 Reliability and validity test of data

After the statistics of dimensionality reduction, the total number of variables finally included can be explained by four common factors, and then the reliability of these four common factors is tested by SPSSAU, and the results are as follows Table 27:

Table 27 Reliability Test Table of Latent Variables

aspect	Cronbach's α	number of terms
Understanding degree of word poverty	0.763	5
The form of expression of word poverty	0.992	8
The consequences of poor words	0.979	5
The causes of the formation of word poverty	0.887	6

According to Table 26, Cronbach's Alpha coefficient values of all latent variables are above 0.70, which indicates that the latent variables have high reliability and can be included in the path diagram of structural equation model.

7.2.4 Model setting

After setting the relationship between latent variables, the online software SPSSAU is used to set up the causality path diagram as required, and the specific results are as follows figure 14:

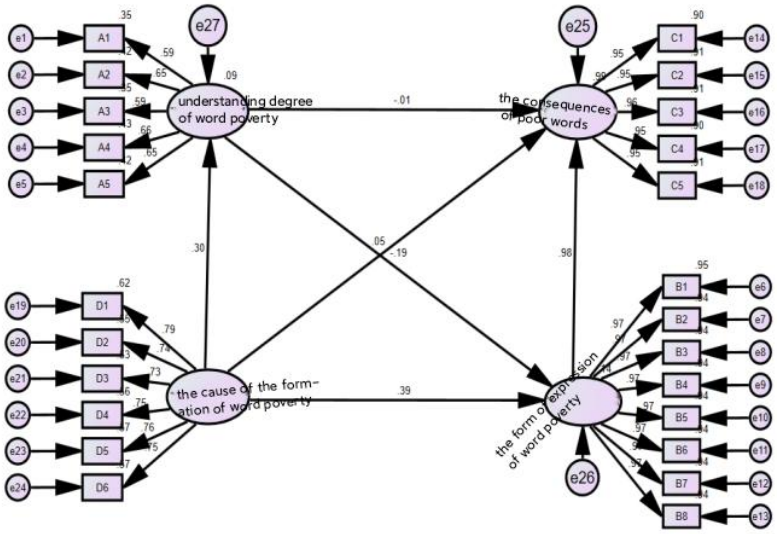


Figure 14 Road Map of Research Theory Model

7.2.5 Fitting degree analysis and hypothesis test of the model

(1)Fitting degree analysis of the model

Estimating the set model parameters with sample data is the primary task to test the structural equation model, but it is the goal of parameter estimation to generate another covariance matrix of the observed variables, so as to make the covariance matrix of the observed variables as close as possible. When the two covariance matrices are very close, we think that the model fits the data.

In this paper, the specific results of fitting index calculated by SPSSAU online software are as follows:

Table 28 Fitting index of structural equation model

Fitting index	CMIN /DF	GFI	RMSEA	RMR	CFI	NFI	NNFI
numerical value	1.223	0.969	0.017	0.046	0.996	0.988	0.998

According to Table 28, the value of chi-square/degree of freedom, that is, CMIN/DF, is 1.225, which is less than the criterion 3, so the data of this model is well fitted. The GFI value is 0.969, which is greater than the criterion of 0.9, indicating that the fitting degree is good. The RMSEA value is 0.017, which is less than the criterion of 0.1, indicating that the fitting degree is good. RMR value is 0.046, which is less than the criterion of 0.05. The CFI value is 0.996, which is greater than the criterion of 0.9, indicating that the fitting degree is good. The NFI value is 0.988, which is greater than the criterion of 0.9, indicating that the model fits well. The value of NNFI is 0.998, which is greater than the criterion of 0.9, indicating that the model has a good fitting degree. To sum up, most of the evaluation indexes of this model have reached the standard.

(2)Hypothesis test of model

In this paper, the test results of each path obtained by AMOS26.0 software are shown in the following table 29:

Table 29 Test Results of Path Hypothesis

null hypothesis	path			P value	result
H1	Understanding degree of word poverty	<—	The causes of the formation of word poverty	***	found
H2	The form of expression of word poverty	<—	The causes of the formation of word poverty	***	found
H3	The form of expression of word poverty	<—	Understanding degree of word poverty	***	found
H4	The consequences of poor words	<—	The form of expression of word poverty	***	found
H5	The consequences of poor words	<—	Understanding degree of word poverty	0.232	false
H6	The consequences of poor words	<—	The causes of the formation of word poverty	***	found
H7	Understand the school's measures to improve word povert.	<—	Understanding degree of word poverty	***	found

(The following 22 paths are omitted, and the results are all valid.)

According to the above table, we can find that there is a significant relationship between the degree of understanding of word poverty and its manifestation, the causes of word poverty are all significantly related to the degree of understanding, the form of expression and the consequences of word poverty, and there is also a significant relationship between the form of expression and the consequences of word poverty, but there is no significant relationship between the degree of understanding and the consequences of word poverty, with a p value of 0.232>0.05. The relationship between observed variables and latent variables is very significant, so the original hypothesis is not completely rejected between latent variables, but between latent variables and observed variables. Therefore, it can be considered that most hypotheses have been tested, but there is still one hypothesis that failed to pass the test.

7.2.6 Modification of model

(1)Fitting degree analysis of the modified model

According to the modified index, the model structure is more reasonable by releasing paths or adding new paths. In order to improve the fitting degree of the model structure, the modified model is obtained after being modified by AMOS26.0 software, as shown in the following figure 15:

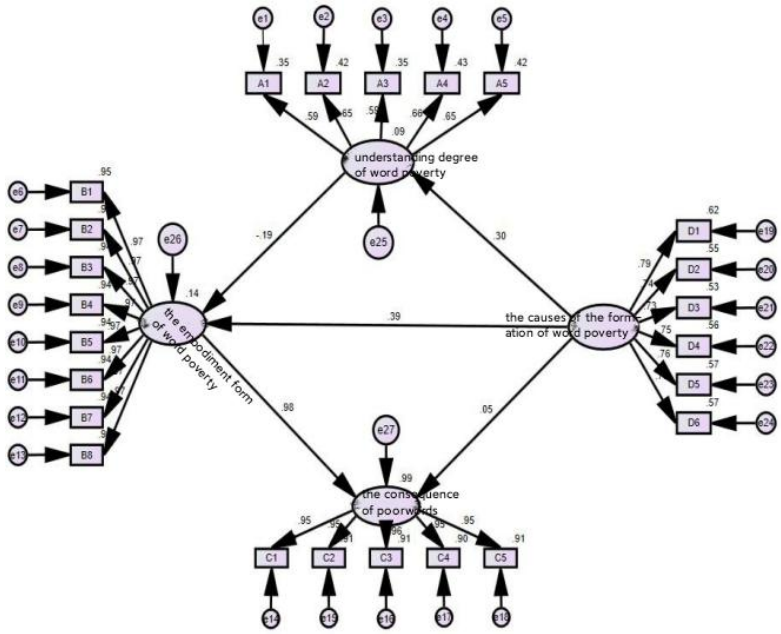


Figure 15 Road Map of Revised Research Theory Model

The modified goodness of fit is shown in the following table 30:

Table 30 Fitting Index of Modified Structural Equation Model

Fitting index	CMIN/DF	GFI	RMSEA	RMR	CFI	NFI	NNFI
numerical value	1.226	0.969	0.017	0.044	0.998	0.988	0.998

As can be seen from Table 30, the value of chi-square/degree of freedom, that is, CMIN/DF, is 1.226, which is less than the criterion 3, so the data of this model is well fitted. The GFI value is 0.969, which is greater than the criterion of 0.9, indicating that the fitting degree is good. The RMSEA value is 0.017, which is less than the criterion of 0.1, indicating that the fitting degree is good. RMR value is 0.044, which is less than the criterion of 0.05. The CFI value is 0.998, which is greater than the criterion of 0.9, indicating that the fitting degree is good. The NFI value is 0.988, which is greater than the criterion of 0.9, indicating that the model fits well. The value of NNFI is 0.998, which is greater than the criterion of 0.9, indicating that the model has a good fitting degree. To sum up, most of the evaluation indexes of the model have reached the standard, which shows that the modified model has a good fitting degree.

(2) Hypothesis test of the modified model

Table 31 Test Results of Revised Path Hypothesis

null hypothesis	path			P value	result
H1	Understanding degree of word poverty	<—	The causes of the formation of word poverty	***	found
H2	The embodiment form of word poverty	<—	Understanding degree of word poverty	***	found
H3	The embodiment form of word poverty	<—	The causes of the formation of word poverty	***	found
H4	The consequences of poor words	<—	The embodiment form of word poverty	***	found
H5	The consequences of poor words	<—	The causes of the formation of word poverty	***	found

(The following 24 paths are omitted, and the results are all valid.)

According to the above table 31, we can find that there is a significant relationship between the degree of understanding of word poverty and the forms of word poverty, between the causes of word poverty and the degree of understanding, the forms of word poverty and the consequences of word poverty, between the forms of word poverty and the consequences of word poverty, and between the observed variables and the latent variables, the original hypothesis is rejected. Therefore, it can be considered that all hypotheses have been tested.

7.2.7 Interpretation and analysis of structural equation model results

(1)Relationship between latent variables

The coefficient between latent variables indicates the degree to which the change of one variable causes the change of other variables. Because the estimated value of residual term is relatively small and usually meaningless, its influence can be directly ignored, so this paper has not conducted in-depth and detailed research and discussion on the error term.

From the path diagram of structural equation model, it is easy to see the path coefficient between latent variables, and the relationship between most of the set latent variables has a significant positive correlation. Among them, there is a significant positive correlation between the degree of understanding of word poverty and the reasons for its formation, and its path coefficient is greater than 0, showing a significance of 0.01 level. At the same time, there is a significant positive correlation between the form of word poverty and the causes and consequences of word poverty, and the path coefficients between them are all greater than 0, and they all show the significance of 0.01 level. There is a significant positive correlation between the form of word poverty and the consequences caused by word poverty, and its path coefficient is also greater than 0, which also shows a significance of 0.01 level. From the above analysis results, it can be seen that there is a certain causal relationship between college students' understanding of word poverty, the form of word poverty, the consequences caused by word poverty and the reasons for its formation.

The regression coefficient between the reasons for the formation of word poverty and the degree of understanding of word poverty is 0.3, which shows that every one percentage point increase in the factor of the formation of word poverty of college students will increase the degree of understanding of word poverty of college students.

Will increase by 0.3 percentage points. It can be seen that when college students are under great pressure to strengthen the use of the Internet, the reading volume decreases, the fragmented reading mode increases, the face-to-face communication decreases, and the activities organized by the school decrease, the frequency of word poverty will be higher, which will lead to their deeper understanding of the impact of word poverty on their study life, their more urgent desire to understand the ways to improve their own word poverty and the measures given by the school to improve word poverty.

The regression coefficient between the reasons for the formation of word poverty and the forms of word poverty is 0.39, which shows that every one percentage point increase in the factors of the formation of word poverty, the factors of the

forms of word poverty will increase by 0.39 percentage points. Similarly, it can be seen that when college students are under great pressure to strengthen the use of the Internet, the amount of reading decreases, the fragmented reading mode increases, the face-to-face communication decreases, and the activities organized by the school decrease, the frequency of word poverty will be higher, which will lead to an increase in the frequency of word poverty in all aspects and at all times, such as academic speech, academic writing, quick answering questions, communicating with strangers, and expressing opinions with professional terms.

The regression coefficient between the causes and consequences of word poverty is 0.05, which shows that every percentage point increase of college students' understanding of the causes of word poverty will increase by 0.05 percentage point. It can also be seen that when college students are under great pressure to strengthen the use of the Internet, the amount of reading decreases, the fragmented reading mode increases, the face-to-face communication decreases, and the activities organized by the school decrease, the impact of poor words on college students' social learning will be greatly affected.

The regression coefficient between the manifestation of word poverty and the consequences caused by word poverty is 0.98, which indicates that every one percentage point increase in the manifestation of word poverty of college students, the influence of word poverty on college students' academic socialization will increase by 0.98 percentage points. That is, when the frequency of college students' word poverty increases in academic speech, academic writing, quick answering questions, communicating with strangers, and expressing their views in professional terms, it is more likely to directly lead to their academic performance decline, class discussion participation and social activities participation decline, self-expression ability decline and personal self-confidence frustration.

(2)The relationship between observed variables and latent variables

In the understanding degree of word poverty, the path coefficient of hearing the word word poverty is 0.59, understanding the specific meaning of word poverty is 0.65, understanding the influence of word poverty on study life is 0.59, understanding the methods of improving word poverty is 0.66, and understanding the improvement measures of school about word poverty is 0.65. This shows that these observed variables have obvious influence on the understanding of word poverty, among which understanding the methods to improve word poverty is the most important influence. It shows that compared with these options, college students in Guangxi have a better understanding of the methods to improve word poverty.

Among the forms of word poverty, the path coefficient of academic speech, academic writing, answering questions quickly, communicating with strangers, expressing opinions in professional terms, and discussing unfamiliar topics are all 0.97. This shows that these observed variables have a very obvious influence on the form of word poverty, and each observed variable can be regarded as the most important influence shows that college students in Guangxi are poor in words, which is generally reflected in academic speech, academic writing, answering questions quickly, communicating with strangers, and expressing opinions with professional terms.

Among the consequences caused by word poverty, the path coefficients of academic performance, class discussion, self-expression ability and personal self-confidence are all 0.95, and the path coefficient of social activities is 0.96. This shows that these observation variables have a very obvious influence on the consequences caused by word poverty, and each observation variable can be regarded as the most important influence. It shows that college students in Guangxi generally believe that word poverty will lead to a decline in academic performance, a decrease in class discussion, a deterioration in self-expression ability and an impact on personal self-confidence, especially on social activities.

Among the causes of word poverty, the path coefficient of network usage habit is 0.79, reading is less, the path coefficient of weak expression ability is 0.74, the path coefficient of fragmented browsing information is 0.73, the path coefficient of lack of systematic thinking is 0.75, the path coefficient of less face-to-face communication is 0.76, and the path coefficient of less participation in school organization exchange activities is 0.75. This shows that these observed variables have obvious influence on the causes of word poverty, among which network usage habits are the most important.

8 CONCLUSION AND SUGGESTIONS

8.1 Research Conclusions

Based on the data of online questionnaire survey, this paper tests the reliability and validity of the questionnaire, analyzes the data of the questionnaire by means of difference analysis, frequency and cross table method in multiple response analysis and structural equation model, and analyzes the understanding of college students about word poverty, the existence of word poverty among college students, the manifestations of word poverty in different aspects, the reasons leading to word poverty and some suggestions to improve it. This paper aims to explore the relationship between college students' personal situation and the current situation of word poverty, understand the interaction mechanism between them, and put forward better measures to improve this phenomenon.

It is found that among the 772 questionnaires, the understanding degree of the word poverty is: 20% can't understand, 25% don't know much, 17% know generally, 21% know well and 17% know very well. This shows that most students have a certain degree of understanding of word poverty, but there are some differences in understanding. This may be influenced by one's learning background, experience and subject field. At the same time, it also shows that word poverty, as a concept, has a certain popularity and attention among college students. From the distribution map of college students' word poverty frequency, 10% college students never have word poverty, accounting for the least;

There are 41% college students who sometimes have word poverty, accounting for the most; There are 28% college students who often have word poverty, and 21% college students always have word poverty. This reflects the universality and frequency of word poverty among college students. Although a small number of students rarely or hardly encounter word poverty, the vast majority of students will face this problem from time to time. Among them, nearly one-third of students often feel the trouble of word poverty, while a considerable proportion of students are almost always in the state of word poverty. This shows that the phenomenon of word poverty is quite common among college students, which constitutes their language expression ability a certain degree of challenge. From the distribution map of the causes of word poverty, it can be seen that the causes of college students' word poverty are internet use habits, less reading, fragmented browsing of information, less face-to-face communication, less participation in exchange activities organized by schools and great psychological pressure, among which less reading, internet use habits and less face-to-face communication are the main reasons. This set of data emphasizes the influence of insufficient reading, internet usage habits and lack of face-to-face communication on word poverty.

Through the investigation of the current situation of college students' word poverty in four regions of Guangxi, this study finds that college students do have a serious phenomenon of word poverty. At the same time, we conducted an independent sample T-test with gender, region, different educational background and subject category for the frequency, understanding and attitude of college students towards word poverty in writing and communication. The results show that the frequency of word poverty is almost equal between men and women; The frequency of word poverty in different regions is not much different; There are statistical differences in the frequency of word poverty among college students with different academic qualifications, and it is considered that college students have more frequent word poverty; There is no statistical difference between science and engineering, but there is a difference between science and liberal arts; There is no difference in the understanding of word poverty among college students of different sexes; There are some differences in college students' understanding of word poverty in different regions, among which Yulin and Liuzhou have the highest proportion of ignorance; There is no difference in the understanding of word poverty among college students with different academic qualifications; Engineering has the lowest understanding of word poverty, while liberal arts is relatively high; There is almost no difference in the attitude of college students of different sexes and educational backgrounds towards the phenomenon of word poverty; However, college students in different disciplines have different attitudes towards the phenomenon of word poverty, and the frequency of Chinese science and engineering students thinking that word poverty is a normal phenomenon is higher than that of liberal arts students. The results of these differences remind us that when educating and tutoring college students' language expression ability, we should consider the differences between different groups and take targeted measures and methods to help them overcome the problem of word poverty and improve their language expression ability.

Through the multiple response analysis of college students' views on measures to improve word poverty, the conclusion shows that there are many efforts and measures to improve college students' word poverty. By encouraging and supporting students to participate in debates and speech contests, providing psychological consultation and counseling services, providing network communication skills training and standardizing language use, the school has demonstrated active advocacy and practice for improving students' language expression ability. This not only helps to solve the problem of word poverty, but also cultivates students' self-confidence and ability to adapt to the information age. The comprehensive application of these measures will help to cultivate compound talents with international competitiveness and lay a solid foundation for their future development. Therefore, schools and society should continue to pay attention to and strengthen these aspects in order to promote the all-round growth and development of college students.

The research results of structural equation model reveal the complex relationship between college students' cognition and behavior on word poverty. It is found that with the increase of college students' dependence on the internet, the decrease of reading volume, the strengthening of fragmented reading mode, the decrease of face-to-face communication and the decrease of participation in exchange activities organized by schools, the frequency of their word poverty has increased significantly. This not only deepens their understanding of the influence of word poverty, but also makes them more eager to know how to improve their word poverty and the improvement measures provided by the school. It is particularly noteworthy that Internet usage habits have played a vital role in shaping word poverty and become the leading factor among the influencing factors. Therefore, these findings call on us to take effective measures to guide and improve college students' internet.Network behavior, to provide them with better support and guidance in academic and social aspects, in order to deal with the problem of poor words and promote their all-round development.

8.2 Deficiencies

Due to the limitation of objective factors. The respondents in this study are only distributed in four cities in Guangxi, and the scope of the respondents is relatively small, which may have some influence on the overall situation of college students' word poverty, and may not represent the specific situation of college students' word poverty in China.

In this study, the causes and present situation of word poverty are only preliminarily investigated, and it is necessary to further investigate whether there will be potential effects among various factors in the later stage.

For some reasons, all the questionnaires collected are from online questionnaires, and they are not distributed face to face offline, which may have a certain impact on the quality of the questionnaires.

At present, there is not much corresponding research on the phenomenon of college students' word poverty. The design of the questionnaire used in this study may have some defects and is not perfect, and it needs further revision in the later stage.

8.3 Suggestions

1. Suggestions from the social level: (1) Advocate and advocate diversified reading culture, and encourage young people to extensively dabble in various literary works, academic materials and news reports, so as to expand their vocabulary and improve their language expression ability; (2) Strengthen the guidance and supervision of network information, promote the network platform to provide more high-quality and valuable content, and reduce the influence of fragmented reading mode; (3) Carry out vocabulary education and language training activities, including holding vocabulary expansion classes and language expression training camps to help college students improve their vocabulary and language expression ability; (4) Increase publicity and promotion activities on reading culture, including holding reading sharing meetings, literature lectures, reading promotion weeks, etc., to stimulate college students' interest and enthusiasm in reading.
2. Suggestions at the university level: (1) Set up special courses or workshops for language expression, aiming at teaching students how to effectively use vocabulary and language to communicate and express. (2) Actively encourage and support students to participate in practical activities such as debate, speech and writing. Provide students with sufficient stage and opportunities to show their language ability and thinking depth, so as to cultivate their self-confidence and expression ability; (3) Strengthen students' mental health education and counseling services to help them effectively cope with stress and anxiety. Schools can organize mental health lectures and provide psychological counseling services to provide students with channels for emotional management and stress release, so as to promote their healthy growth.
3. Personal suggestions: (1) Pay attention to improving writing skills, and constantly improve your expressive ability through regular writing exercises and feedback; (2) Actively participate in various social and academic activities, have face-to-face communication and interaction with others, and exercise their oral expression skills; (3) Take the initiative to use network resources for learning and communicate, but also pay attention to choosing valuable and credible information sources to avoid falling into the misunderstanding of fragmented reading and expression; (4) Cultivate self-confidence and dare to express your views and ideas. Participate in public speeches, debates, group discussions and other activities, speak actively, and show your thinking depth and logical thinking ability.

COMPETING INTERESTS

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

REFERENCES

[1] Kang Ji-nan. thinking behind "forgetting words when writing" and "forgetting words when opening one's mouth". Kaifeng daily, 2025(003). doi: 10.28499/n.cnki.nkfrb.2025.000020.
[2] Liu Pengbo. Young people are "poor in words" and blame the Internet?. Literary and Art Newspaper, 2024(008). doi: 10.28817/n.cnki.nweyi.20080.000000000106
[3] Yao Yuan. China's dilemma of contemporary design-unyielding reason and poor words. Anhui Literature (second half, 2009(02):117.
[4] Young man, are you "poor in words"?. Studying, 2024(11):80.
[5] Guang Hong. How to help young people out of the "word poverty" dilemma. Trade Union Expo, 2024(08):35.
[6] Qiu Hui. From argument to eloquence-let students improve their writing quality in "speaking". Chinese World, 2018(29):69-70.
[7] Le Tao. Endless meaning and endless words. News business, 1964(09):8.
[8] Ian C .Word rich or word poor? Deficit discourses, raciolinguistic ideologies and the resurgence of the ‘word gap’ in England’s education policy.Critical Inquiry in Language Studies, 2023, 20(4): 305-331.
[9] Why are some young people "poor in words". Reading for Middle School Students (Senior High School Edition) (second half), 2024(Z1): 71-72.
[10] Cao Yaxin, Zhang Meili. Leave the "net stalk", are you also "poor in words"?. Jinan daily, 2024(008). doi: 10.28453/n.cnki.njnrb.2024.000932.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3016

FROM "THE LIGHT OF THE COUNTRY" TO "THE PHENOMENAL EXPLOSION" —— THE AUDIENCE RELATIONSHIP AND THE WAY TO "BREAK THE CIRCLE" IN DECODING "NEZHA 2"

Baoshuo Wang
Department of Statistics, Guangxi Normal University, Guilin, 541006, China
Corresponding Author: Baoshuo Wang, Email: 3095045418@qq.com

Abstract: "China's 'Fourteenth Five-Year Plan' explicitly proposes to promote the development of the digital culture industry. Animation, as an important carrier for 'telling Chinese stories well, ' has received special funding support. The government advocates for the 'creative transformation of fine traditional Chinese culture, ' promoting the exploration of themes such as mythology, history, and intangible cultural heritage in national animation, which aids in cultural exports and national identity. As a sequel to the domestic animated film 'Ne Zha: The Devil Child Makes His Debut, ' 'Ne Zha 2' continues the industry topic of'the rise of national animation'with the accumulated popularity of the national IP and audience expectation s from the predecessor. This study uses a questionnaire survey method, focusing on the population from four regions in Guangxi, aiming to analyze how 'Ne Zha 2' breaks through the audience layer limitations of animated films to become a universally popular hit and to summarize its 'breakthrough' market strategies, providing a reference for the national animation industry. A total of 528 questionnaires were collected, and after removing invalid questionnaires, 430 were left. Through basic descriptive analysis of the data, we gained an understanding of the general distribution. Then, differential analysis and multiple response analysis were conducted to explore the correlation between the respondents' basic information and the rese arch questions. Finally, a structural equation model was constructed to explore the complex relationship between the subjects' understanding of 'Ne Zha 2' and its development strategies. The study found that the audience's understanding of 'Ne Zha 2' is generally high, and 'Ne Zha 2' has also had a certain impact on the future development of national animation. The reasons for 'Ne Zha 2' being able to truly 'break th rough' include content and emotional, technical, and marketing aspects. By delving into the reasons for the success of 'Ne Zha 2', we can further promote the development of the domestic animated film industry and build a more prosperous and innovative cultural industry structure. Only by fully understanding and drawing on the successful experience of 'Ne Zha 2' can we better meet a more competitive film and television market and achieve the long-term development goals of the national cultural industry.
Keywords: Population from four regions in Guangxi, “Ne Zha 2”, “Breakthrough” strategies, Differential analysis, Multiple response, Structural equation model.

1 INTRODUCTION

1.1 Research Background and Significance of the Topic

1.1.1 Research background

In recent years, driven by policy support, capital investment, technological innovation and market demand, China's domestic animation gradually got rid of the passive situation dominated by Japanese animation for a long time, showing a rapid development trend. According to the Research Report of China Animation Industry in 2023, the market scale of animation industry in China has exceeded 300 billion yuan, and the annual growth rate has remained above 15%. Guo Man's works, such as Ne Zha, The Return of the Great Sage, Three Wan Li in Chang 'an, etc., not only made a breakthrough in business, but also formed a unique "national aesthetics" by integrating the traditional cultural elements of China and became an important carrier of cultural output. Nezha 2 does not stay in the superficial reappearance of Nezha's fairy tales, but guards the cultural roots with awe. Through the in-depth study of the text textual research, folk custom evolution and artistic representation of IP stories and their background times, it finds the cultural anchor point that can produce national emotional connection, so that it can realize value remodeling and cultural resonance in the contemporary context[1]. For example, Nezha's The Devil's Child Roars the Sea, with its innovative narrative, vivid characters and unique cultural expression, broke the audience's inherent cognition of fairy tales, and at the same time, it won a high box office and stimulated the creators' enthusiasm for exploring heroic movies[2].

The "14th Five-Year Plan" clearly puts forward "accelerating the development of new cultural formats and promoting the high-quality development of the digital cultural industry". As a part of the digital content industry, Guoman has obtained the policy inclination at the national level. Many local governments, such as Hangzhou and Chengdu, have promoted the development of Guoman industry by building animation industrial parks, holding international animation festivals and comic exhibitions. Despite the rapid development momentum, Guo Man still faces multiple structural challenges: for example, a large number of works focus on fantasy and cultivation of immortals, and the plot is

seriously routine; Social issues and realistic themes facing adult audiences are scarce; Guoman faces cultural differences and limited distribution channels when going out to sea, and has not yet formed a global influence; And the demand for high-end technical and creative talents is in short supply. The rise of Guoman is not only related to the economic benefits of the cultural industry, but also a key link in the construction of China's cultural soft power. From the theoretical level, it is helpful to reveal the development law of emerging cultural formats under the background of globalization and digitalization to dig deep into the development trend and predicament of high-quality animation. From the practical level, it can also provide a path reference for more employees to create high-quality content, and help the country to achieve a leap from "quantitative change" to "qualitative change". From Ne Zha to The Devil's Child in Nezha (hereinafter referred to as Nezha 2), Nezha 2, as a sequel to phenomenal IP Ne Zha, its breakthrough is not only reflected in the continuation of box office or word of mouth, but also through technological innovation, narrative upgrade, industry consolidation and cultural expression. In the aspects of role modeling and narrative structure, the film breaks the inherent pattern of traditional fairy tales and endows the protagonist Nezha with more complex and multidimensional personality characteristics[3]. Guoman industry has set a new benchmark. Therefore, it is of great significance to study the reasons for the "broken circle" in Nezha 2 for the future development of the country. By deepening the audience's understanding of Nezha 2, the influence of Nezha 2 on the development of Guoman and the reasons for its "broken circle", we can provide effective countermeasures and methods for the subsequent development of domestic animation. At the same time, it can also have a positive impact on the development of a series of industries, such as film industry, cultural tourism industry and so on.

1.1.2 Research purposes

"Nezha's Devil Children Roaring the Sea" once again confirms the key role of digital technology in promoting the "two innovations" of Chinese excellent traditional culture. In terms of unique advantages, digital technology drives the transformation of Chinese excellent traditional culture from "tradition" to "modernity", promotes its transformation from "recessive" to "dominant" and helps it transform from "roaming the sea" to "going out to sea"[4].The influence of the development of the country on the national cultural industry is multi-faceted and all-round. First of all, the rise of Guoman has injected new vitality into China's cultural industry. By integrating traditional and modern elements, it not only inherits the excellent cultural heritage of the Chinese nation, but also promotes the modern transformation and innovative expression of traditional culture. Secondly, the prosperity of Guoman has promoted the perfection and upgrading of the animation industry chain, from content creation, production technology, distribution channels to derivative development, each link has promoted the cultivation of professional talents and the increase of employment opportunities; Furthermore, the international communication of Guoman has enhanced the international influence of Chinese culture, enhanced the international recognition of national cultural brands, and created new growth points for the export of Chinese cultural products; Finally, the development of Guoman has also stimulated the whole society's attention and investment in cultural and creative industries, promoted the cross-border integration of cultural industries and other fields, and provided strong support for the country's economic restructuring and cultural development and prosperity.

The purpose of this study is to explore the reasons for the great success of Nezha 2. By analyzing the plot conception, character shaping, visual effect, marketing strategy and audience's psychological needs, this paper aims to reveal how Nezha 2 stands out among many animated films and becomes a phenomenal work, and provide useful reference and enlightenment for the development of China's animated film industry, and further promote the innovation and prosperity of domestic animated films. Through text analysis, semiotic interpretation and cross-cultural communication theory, this paper explores how this work establishes meaning resonance in the global cultural market, and analyzes its communication strategies and challenges in the postcolonial context[5].

By deeply exploring the reasons for the success of Nezha 2, we can further promote the development of domestic animated film industry and build a more prosperous and innovative cultural industry pattern. Only by fully understanding and drawing lessons from the successful experience of Nezha II can we better meet a more competitive film and television market and realize the long-term development goal of the national cultural industry.

1.1.3 Research significance

The film has achieved a breakthrough innovation in the use of traditional cultural elements. By drawing lessons from the bronze portraits of Sanxingdui site, the film created a unique image of "boundary beast"; However, the role of Taiyi real person has realized the modern transformation of traditional characters by integrating Sichuan dialect and comedy elements[6]. In this book, we will deeply explore the audience's attitude towards Nezha 2, the influence of Nezha 2 on the development of Guoman, the reason why Nezha 2 broke the circle successfully and the possible development direction of Guoman in the future. First of all, through online literature review, we will sort out the relevant research results and deeply understand the influencing factors and existing solutions of the development of Guoman in the direction of high quality. Secondly, we will conduct a questionnaire survey to find out the interviewees' understanding, experience and feelings about Nezha 2 and their expectations for the future development of the country. At the same time, we will also interview relatives and friends around us.

Through in-depth communication and observation, we can collect more specific and in-depth information in order to understand all aspects of the phenomenon of word poverty more comprehensively.

By analyzing the collected data, we will be able to clearly understand the audience's attitude towards Nezha 2' s going out of the circle and the actual impact that they think Nezha 2 has brought to the development of the country. Based on these analysis results, we will analyze the corresponding measures and suggestions for the development and improvement of Guoman, so as to help the domestic animation industry develop better, improve its production quality and technical level, and then better contribute to the development of the domestic animation industry and go global.

If we can effectively analyze the real reason of "Nezha 2" breaking the circle, it will not only directly improve the overall quality of the domestic animation industry, but also become a bridge for cultural exchange, ease the obstacles of international cultural exchange and enhance the national cultural soft power. By improving the quality of Guo Man's production, the traditional culture is innovated in the form of modern animation, which makes it more in line with modern aesthetics and values, thus alleviating the problem of cultural fault and traditional loss; It is conducive to transmitting positive values and knowledge to young people and improving the interest and effectiveness of education; At the same time, the development of animation industry has created a large number of employment opportunities, which has alleviated the employment pressure to some extent. Therefore, the significance of this study is not only to explore the reasons for the "broken circle" of Nezha 2, but also to make positive contributions to the progress of more industries, society and even the country.

1.2 Research Status

Based on the perspective of collective memory theory, the cultural practice of Nezha's Devil Children Naughting the Sea provides a typical sample for the construction of Chinese cultural identity[7]. Since its release in the Spring Festival in 2025, Nezha's Devil Children Naughty the Sea (hereinafter referred to as Nezha 2) has set a number of box office records. As of February 16th, the global box office has reached 11.934 billion yuan, approaching the top ten in the world (for example, The Lion King is 12.051 billion yuan), and it has become the first China film to break through 10 billion yuan. It is predicted that the box office will rise to 16 billion yuan, which is expected to hit the top of the global animated film box office. Although the North American market is due to statistics, the lag caused controversy (the single-day box office of $853 was a false positive), but the actual first weekend box office exceeded $20 million, which topped the New Zealand Chinese film box office champion and attracted the attention of the Oscar judges, and was rated as "the pride of China culture".

Nezha is one of the most successful films in recent years, both at the box office and by word-of-mouth[8]. The success of "Nezha 2" promoted the full activation of the film industry chain in China. The share price of Light Media soared 13.8 billion yuan in six days, and derivatives (such as hand-made blind boxes and picture books) sold well, and capital and policies further promoted the animation industry.

The success of Nezha 2 is the result of many factors, such as precise positioning, excellent quality, adaptability, emotional value, word-of-mouth communication, social topics, technological empowerment and cross-border linkage[9]. However, the industry is still facing stability challenges, and it is necessary to continuously improve the profit model of animation companies and the treatment of employees. In the future, Nezha 3 is scheduled to be released in 2030, and director jiaozi emphasized that "every film should break through the limit". The global release of Nezha 2 (covering North America, Australia and New Zealand) provides a new paradigm for China's cultural export, but it also needs to deal with cultural differences and market adaptability.

The success of Nezha II is a model of synergy among technology, culture and market. Its box office miracle confirms the audience's demand for high-quality local content. The industrialization cooperation model sets a benchmark for the industry, while cultural self-confidence and narrative innovation give vitality to the traditional IP era. This case not only marks the rise of China animation, but also provides a new path of "content-led technology" for the global film industry. The success of "Nezha" series films shows that in the global cultural interaction, while maintaining cultural subjectivity, we should adhere to openness, tolerance, exchange and mutual learning, promote the paradigm upgrade of cultural security concept, reconstruct modern discourse and shape a new pattern of civilized dialogue, and finally move towards a pluralistic symbiotic picture of "beauty and beauty together"[10].

1.3 The Question Raised

The gradual rise of domestic animation has become the object that people in many countries pay close attention to, especially Nezha 2, which is a big hit this year. With its cumulative global box office exceeding 14.5 billion yuan, it has successfully ranked seventh in the global box office history, which has aroused widespread concern in China and has also had an important impact on the international film market. This achievement marks the rise of China's animated films, breaking the long-term monopoly of western capital on the film market and winning respect and status for China's films. The "broken circle" of Nezha 2 not only has an impact on the whole film industry, but also has a certain impact on the cultural field of the whole country. As a key force of social culture, the development trend of domestic animation is in a downturn, which may affect the overall cultural environment of society and even the country. If domestic animation is generally difficult to rise, then the whole society will also be affected to a certain extent, thus affecting the development and progress of the country. This article will focus on the following issues:

1. Analyze the audience's understanding of Nezha 2;
2. Reveal the audience's emotional expression in different aspects when watching Nezha 2;
3. Explore the various influences of Nezha 2 on the domestic animation industry;
4. Analyze the reasons for the "broken circle" in Nezha 2;
5. Put forward suggestions on the future development direction of domestic animation industry.

1.4 Research Methods

1.4.1 Literature research method

By systematically collecting, sorting out and analyzing the previous research results, the literature method can understand the research status of the development of domestic animation industry and provide theoretical support for this study. Collect,

classify and screen the documents related to the problem through China HowNet database and WeChat official account, and use appropriate methods and tools to make in-depth analysis of the sorted documents, so as to integrate the influence of Nezha 2 on the development of the country and the "Nezha 2". The reasons of "circle" and the effective strategies for the development of the country in the future provide theoretical basis for this paper.

1.4.2 Questionnaire survey method

Questionnaire method is an investigation method that investigators use a unified questionnaire to understand the situation or ask for opinions from the selected respondents. After the pre-investigation of the questionnaire was reasonable, this study distributed the questionnaire through the way of distributing leaflets on the test paper online and offline, and obtained the respondents' understanding and suggestions on Nezha 2, and collected a large amount of data to provide a basis for subsequent analysis.

1.4.3 Data analysis method

Statistical analysis of the collected questionnaire data reveals the influence of Nezha 2 on the development of the country. The reasons for the "broken circle" in Nezha 2 are convenient for the development of this paper.

1.5 Research and Innovation

1.5.1 The topic is novel, filling the research blank

The rise of Guoman has gradually aroused widespread social concern in recent years. In recent years, the films released by Guoman industry, such as Return of the Great Sage to the Journey to the West, Big Fish Begonia, Deep Sea, Three Wan Li in Chang 'an, and Black Myth Wukong, are all symbolic representatives of Guoman's gradual rise, but the blockbuster film released in 2025.

"Nezha 2" (full name "The Devil's Child in Nezha") ranked seventh in the box office of the global film history before other films, and the people behind it are obvious to all, which laid the foundation for the national film industry, film industry and even cultural enterprises in the future. Therefore, we study the audience's understanding of Nezha 2, and then study the impact of Nezha 2 on the national animation industry, so as to analyze the reasons for the "broken circle" of Nezha 2, and put forward effective suggestions for the development of domestic animation industry, which can appropriately fill the research gap in this neighborhood and provide some basic data and theoretical support for the follow-up research.

1.5.2 The uniqueness, concreteness and development of the research object

In the past, the research on domestic animation may pay more attention to teenagers, college students and animation lovers, while the general population or specific professional groups, as an important part of social members, may have different cognition of domestic animation from other groups. Therefore, it is unique to study the cognition of domestic animation by a wider group. Research on multi-groups involves a wider and more comprehensive scope, and the implementation of measures may be more effective. By studying various groups in society, we can better promote social development.

1.5.3 Empirical research is more convincing

Through the questionnaire survey of various groups and groups of all ages to collect data and make empirical analysis, we can intuitively understand the understanding of Nezha 2 by various groups in society, the influence of Nezha 2 on domestic animation, and the reasons for the "broken circle" of Nezha 2. Based on the research results, it can provide targeted policy suggestions for the film industry, animation industry, cultural tourism industry, etc., such as integrating more local culture, lowering the viewing threshold, increasing interest, strengthening cross-border cooperation with other cultures, etc., to help all sectors of the industry develop better, which is also the innovation of this paper.

2 THE INVESTIGATION PLAN AND IMPLEMENTATION

2.1 Design of Research Scheme

2.1.1 Purpose of investigation

Since the release of Nezha's Magic Boy in the Sea in the Spring Festival in 2025, it has grossed 14.883 billion yuan (as of March 9, 2025) set a new global box office record for animated films, and the number of people watching movies exceeded 300 million, making it the first film in China's film history to achieve this achievement. Its success is not only a breakthrough of individual works, but also a symbol of the systematic rise of China animation industry. This paper will analyze the audience's understanding of Nezha 2, the reasons why Nezha 2 broke the circle and the influence of Nezha 2 on the future development of Guoman.

The audience attraction of Nezha 2 stems from multi-level integration and innovation. First of all, on the technical level, the film has created the ultimate audio-visual experience with 1948 special effects shots (accounting for 80%). This technical "dead knock spirit" is not only reflected in the picture accuracy, but also through the integration of intangible elements such as Dong songs and suona soundtracks, a unique oriental aesthetic system has been constructed. Secondly, the deep excavation of cultural resonance is indispensable. In addition, after the production team discovered that the image of the groundhog became popular among young people through social media, it was quickly implanted into the plot and became a key role in a series of jokes. This "invisible director" model accurately captured the aesthetic preferences of contemporary youth groups. Despite the unprecedented success of "Nezha 2", there are still deep contradictions behind it. First, the imbalance between technological breakthrough and narrative depth. Although the film shocked the audience with special effects, Douban commented that its emotional turning point was blunt. Secondly, the structural obstacles of cultural output. The lack of non-English dubbing has led to a significant cultural discount. An American film critic ridiculed that "subtitles are like Chinese listening tests". Furthermore, the risk of simplification of IP ecology. At present, Guoman still relies too much on the adaptation of traditional

myths. Although "Nezha 2" has activated classic IP such as "What's wrong with the sea", its original ability is weak and it has fallen into the "adaptation of Four Great Classical Novels".

The purpose of this paper is as follows:

(1)To understand the audience's understanding of Nezha 2 and the number of times they watched it.

(2)Understand the reasons for the "broken circle" of Nezha 2 and its influence on the future development of Guoman, explain the importance of studying the development of Guoman, and provide an analytical basis for the topic.

(3)Understand the audience's attitude towards the phenomenon of 2-brush, 3-brush or even N-brush "Nezha 2".

(4)For the steady development of domestic animation industry in the future, we will formulate methods and measures through investigation and analysis to help Guoman develop continuously.

2.1.2 Object of investigation

Lack of experience in watching movies, weak sense of innovation, fragmented reception of film and television works and over-reliance on special effects scenes may all lead to a shallow understanding of the connotation of movies, making the phenomenon of' difficulty in interpretation' common in film and television reviews. Some people may think that people with difficulty in interpretation are usually ordinary people with low artistic accomplishment, while people with high artistic accomplishment can always talk about it in Kan Kan when watching movies. In fact, it is not the case. In many contemporary social groups, it is not uncommon to find it difficult to interpret movies, and the manifestations are diverse. Based on this situation, we take people of different ages and different groups as the research objects, including teenagers under 18, 18-25, 26-35, 36-45, 46-year-old and above, college students, on-the-job workers, anime fans and other groups, to explore the reasons for the "broken circle" of Nezha 2, in order to reveal how this film stands out among many animated films.

2.1.3 Survey content

The survey objects include teenagers, college students, on-the-job personnel, anime lovers and other social groups, and the basic information involves: gender, age and region; This paper investigates the understanding of Nezha 2 by many social groups, understands their times of watching Nezha 2 and their emotional and psychological reactions when watching it, clarifies their post-viewing attitude towards Nezha 2, and analyzes the influence of Nezha 2 on the industrial development of Guoman and the root causes of its "broken circle". Investigate the effective improvement measures of all groups for the future development of the country, collect the methods recognized by most of them, and make reference for the subsequent development of the country.

2.1.4 Investigation method

(1)Pre-investigation

Before the formal investigation, we distributed pre-questionnaires in the circle of friends, QQ space and private chat with friends, and collected a total of 82 questionnaires in two days. The reliability and validity of these questionnaires were tested by SPSS, aiming at modifying the questionnaires and ensuring the rationality of the questionnaires for subsequent development.

(2)random sampling

Simple random sampling can be divided into two types: sampling with replacement and sampling without replacement. For the sampling with return, each individual drawn is returned to the population, and each individual may be drawn more than once. For non-return sampling, it will not be returned after drawing, and each individual can only be drawn once. Select the key survey method to extract individual samples from each key unit, including Guilin, Nanning, Liuzhou and Qinzhou.

(3)major investigation

Key survey refers to an incomplete survey method that only some key units (here refers to regions) are selected for investigation, so as to obtain data. Although the key units only account for a small part of all the respondents, the number of survey marks accounts for a large proportion, so the data obtained from the survey of these key units can be used to reflect the basic trend of economic changes. In this paper, the social groups in Guilin, Nanning, Liuzhou and Qinzhou in Guangxi are selected as the research objects through the key units shown in the questionnaire to study the audience relationship of Nezha 2 and the related problems of its "broken circle" reasons.

2.1.5 Investigation scheme design

(1)Determination of sample frame

According to the results of our questionnaire, the social groups in Guilin, Nanning, Liuzhou and Qinzhou, where the questionnaire groups are relatively concentrated, form a sample frame.

(2)Determination of sample size

Because the registered permanent residence of the survey group is located in several cities, in order to reduce the consumption of manpower, material resources and financial resources, this survey is conducted by simple random sampling. Then, from the sample size formula (1) in simple random sampling

$$n = \frac{Z_{\alpha/2}^{\ 2}}{d^2} S^2 \tag{1}$$

In formula (1), n is the sample size. Under the guarantee of 95% confidence level, $Z_{\alpha/2} = 1.96$, d =0.05 is absolutely wrong, and S^2 stands for population variance. Through the questionnaire collected this time, it is easy to know our sample. The difference s^2 is 0.5290, so is the sample variance s^2 instead of population variance, and the sample size calculated by

formula (1) is 430.

2.2 Investigation and Implementation

2.2.1 Measuring method

Before the formal questionnaire survey, the personnel involved in the test should be trained uniformly to ensure the standardization of the whole test process. The questioners are all social groups, and the questionnaires will be collected as soon as they are sent out. Before the test, we communicated with the young students' head teachers, college counselors, and company management personnel who participated in the test. Taking all grades and departments as the basic units, we tested the questionnaire-related scales for 20 minutes. The questionnaire was conducted online and offline. The online survey was sent to the official website group and the work exchange group of the school where the members joined, and the offline survey was randomly selected by issuing paper questionnaires. After the questionnaire was collected, the online survey results were directly exported from the Questionnaires platform, and the offline survey results were carefully checked by the team members and integrated with the online survey results into Microsoft Excel. Finally, the data were processed and analyzed by using IBM SPSS Statistics software.

2.2.2 Quality control

(1)Quality control principle

Quality control is an important link to ensure the accuracy and reliability of survey data. In order to obtain accurate and effective data and improve the scientificity in the survey, it is necessary to control the quality of each stage of the survey to reduce errors and enhance the credibility of data analysis. Quality control in questionnaire survey should follow the following principles:

(2)The specific implementation process of quality control of questionnaire survey

1)Quality control before investigation

The quality control before the formal investigation includes four aspects, namely, the formulation of questionnaire principles and standards, the selection of sample population, the training of team members and the questionnaire test.

In formulating the principles and standards of the questionnaire: In order to ensure the quality of the collected data, it is necessary to clarify the principles and standards of the questionnaire design before formulating the questionnaire, including the clarity, operability and consistency of each question in the questionnaire.

In the selection of sample population: the design of the questionnaire needs to have clear and specific survey objects. In order to ensure the reliability of the collected data, it is necessary to select the appropriate sample population according to the purpose of investigation and study, and accurately screen out the respondents according to the characteristics and requirements of the selected population. Random sampling, stratified sampling and other methods can be adopted to effectively ensure the representativeness of the samples.

In the training of team survey members: in order to ensure the accuracy and consistency of data, team members need to repeatedly determine the purpose and content of the survey, and clearly understand the specific meaning behind each question and the whole survey process, so as to ensure that each member has a unified and correct understanding of the project of this survey.

In the aspect of questionnaire test: before the formal investigation, it is very important and necessary to test the questionnaire. Through the test, the defects and loopholes in the questionnaire can be found in time, so as to make repeated revisions and try to fill in the questionnaire in the group, and constantly check whether there are still problems in the content of the questionnaire. At the same time, effective measures can be taken to improve the recovery rate of the questionnaire, such as offering rewards, setting the deadline of the questionnaire, sending reminders of the questionnaire, etc. Questionnaire test is helpful to evaluate the reliability and validity of the questionnaire, so it is a very important step to test the questionnaire before its formal implementation.

2)Quality control in investigation

Quality control in investigation generally includes questionnaire implementation and regular inspection of the effectiveness of the questionnaire.

In the aspect of questionnaire implementation: the participants in the questionnaire generally ignore the questions and have incomplete understanding of the questions. Team members need to ensure that each participant in the questionnaire understands the specific meaning of the questions in the process of investigation implementation to avoid the above situation.

Regularly check the effectiveness of the questionnaire: team members need to carefully check the effectiveness of the recovered questionnaires at different times, and correspondingly eliminate the recovered invalid questionnaires according to the judgment principle of eliminating invalid questionnaires, and supplement the missing values according to the treatment principle of missing values. However, the number of questionnaires was reduced after the invalid questionnaires were eliminated. At this time, it is necessary to consider whether it is necessary to increase the number of questionnaires distributed to ensure that the number of valid questionnaires recovered in the later period is sufficient, so as to meet the requirements of the pre-survey and formal survey for the best sample size. At the same time, team members need to check the questionnaire regularly during the process of recycling the questionnaire, and if necessary, they can ask the students who have completed the questionnaire, interview their feelings during the questionnaire filling process and give suggestions to improve and modify the questionnaire.

3)Quality control after investigation

After designing the questionnaire in the early stage, it is necessary to effectively recycle the questionnaire, and repeatedly

check, register, code, enter and clean the recovered questionnaire data to avoid missing the recovered effective data. Finally, the data of the questionnaire will be processed to pave the way for the later research and analysis.

3 THE QUESTIONNAIR DESIGN

By consulting and reading a lot of relevant data and documents in the early stage, this paper aims at the attitudes of various groups towards Nezha 2, the impact of Nezha 2 on Guoman industry, the reasons for its successful "breaking the circle" and the effective path of Guoman development are analyzed to design a questionnaire. The distribution of the subject content of the questionnaire is shown in the following table 1.

Table 1 Five Sections of the Questionnaire

Questionnaire section	explain
A. Basic information	Some basic information of respondents from all social groups, such as gender, age, region and group (questions 1-4)
B. The audience relationship of Nezha 2	Respondents' understanding of Nezha 2, the number of times they watched Nezha 2, their emotional and psychological reactions when they watched the movie, and their attitude towards the phenomenon of "Nezha 2" (paras. 5-8 Title)
C. The impact of Nezha II	The influence of Nezha 2 on the development of the country (Question 9)
D. Influencing factors of "Broken Circle" in “Nezha 2”	Respondents think that the reason why "Nezha 2" was successfully "broken" (questions 11-13)
E. Development strategy of Guoman	Respondents' suggestions on the future development of the country, such as the emotional level of content and the level of international competition (questions 14-16)

4 PRE-INVESTIGATION

4.1 the Implementation Process of Pre-Investigation

Usually, in order to ensure the overall reliability of the questionnaire and the smooth progress of the formal investigation, it is necessary to make a small-scale pre-investigation before the formal investigation, and according to the analysis of the results of the pre-investigation, find out the mistakes in the questionnaire in time and modify it appropriately, delete the options with repeated meanings, and supplement the questions appropriately when necessary. Our team selected 82 people (including college students, teenagers, on-the-job workers, etc.) from four places in Guangxi for the pre-investigation. The main investigation method was the combination of online and offline, and the paper questionnaire survey was used for the nearby people, while the online questionnaire issued by the Questionnaire platform was used for the people who were far away. The results of 82 questionnaires collected in the pre-survey were analyzed, revised and improved.

4.2 Reliability Test

Generally speaking, the reliability test of questionnaire design quality refers to the process of evaluating the stability and consistency of the questionnaire at different times and in different situations. The purpose of reliability test is to determine the consistency and reliability of the results obtained by the questionnaire measurement tool, that is, whether the questionnaire can get similar results when it is reused. Through the reliability test, we can evaluate the quality and stability of the questionnaire design, confirm that the questionnaire can measure the concepts to be studied stably, and improve the credibility and reliability of the survey results.

Our team will use Cronbach (Cronbach coefficient) reliability to measure the internal consistency coefficient of each scale item in the questionnaire. The specific calculation formula of Cronbach reliability coefficient is formular (2):

$$\alpha = \frac{K}{K-1}\left(1-\frac{\sum S_i^{\,2}}{S_x^{\,2}}\right) \tag{2}$$

Where is the reliability coefficient, which usually ranges from [0,1]. The closer to 1, the greater the reliability of the questionnaire, and the closer to 0, the smaller the reliability. K is the number of test questions in the scale, $S_i^{\,2}$ Represents the variation of scores of all subjects on question I, and $S_x^{\,2}$ is the variance of total scores of all subjects. Generally speaking, a well-designed questionnaire should have a reliability coefficient of at least 0.80, between 0.70-0.80 is an acceptable range, and if the reliability coefficient of each quantity table is lower than 0.60 or the total quantity table, only when the reliability coefficient is lower than 0.80 can the questionnaire be rejected, and the revision of the questionnaire and the supplementary deletion of the questionnaire topics should be reconsidered.

The reliability of five questions in this questionnaire (question 5, question 10, question 11, question 12 and question 13) is tested and analyzed by SPSS software, and the Cronbach coefficients of each questionnaire are shown in the following

table 2:

Table 2 Cronbach Coefficient Table of Pre-questionnaire Scale

Item	Cronbach coefficient	number of terms	Reliability evaluation
Question 5	0.627	five	better
Question 10	0.779	five	good
Question 11	0.756	five	good
Question 12	0.695	four	good
Question 13	0.778	five	good

From the Cronbach coefficient table above, we can see that the Cronbach coefficient of each scale item in the questionnaire is greater than 0.6, and most of them are greater than 0.75. The Cronbach coefficient of the whole scale is 0.950 through reliability test and analysis, so it is considered that the reliability of this questionnaire design is good.

4.3 Validity Test

The validity test of questionnaire design quality refers to evaluating whether the questionnaire measurement tool can accurately and effectively measure the concepts or variables to be studied. Validity refers to whether the measuring tool can measure what it claims to measure, that is, whether the measuring tool can truly reflect the characteristics or nature of the research object. In questionnaire design, validity test usually includes two main aspects:

Content validity: Content validity refers to whether the contents contained in the questionnaire comprehensively and accurately reflect the concepts or variables to be studied.

Structural validity: structural validity refers to whether the measurement tools in the questionnaire can accurately reflect the internal structure or attributes of the research object. The evaluation of structural validity is usually carried out by statistical methods such as factor analysis and correlation analysis to ensure that the questions in the questionnaire can effectively measure the concepts to be studied.

There are many questions in this questionnaire, so structural validity is chosen to test the data collected in the pre-survey. The specific KMO and Bartlett test coefficients are shown in Table 3 below.

Table 3 KMO and Bartlett Test Coefficient Table of Pre-questionnaire Scale

Item	KMO coefficient	Approximate chi-square	significance
Question 5	0.681	46.069	0.000
Question 10	0.612	335.806	0.000
Question 11	0.626	241.665	0.000
Question 12	0.725	51.373	0.000
Question 13	0.601	2934.995	0.000

Generally speaking, the KMO value is between [0,1]. When the KMO value is closer to 1, it means that the correlation between variables is stronger, and vice versa. According to the results of the above coefficient table, it is easy to know that the correlation coefficient between items under each scale is greater than 0.6, which shows that the questionnaire design is effective.

4.4 the Revision of the Topic

Through the pre-survey, the most authentic feedback from all kinds of social groups who participated in the pre-survey questionnaire was collected, such as the unreasonable logic, improper expression and individual typos in the questions, and the overall content, typesetting and topic distribution of the questionnaire were modified and optimized according to their feedback. The specific contents of the questionnaire are as follows:

[Question 1] Redundancy is expressed in the scale-the numbers 1-5 are used to express the degree from "very different" to "very agree" and "no influence" to "great influence", and the numbers 1-5 are very different, disagree, neutral, agree, very agree and have no influence, slight influence, medium influence, great influence and great influence respectively.

[Question 2] Multiple choice questions "What emotional and psychological reactions did you feel when watching Nezha 2?"

(Question 6) Some emotional options are too few (touching, hilarious, inspiring and thinking respectively), which fails to generally represent the audience's feelings of watching movies, and two iconic emotional options (suspense and shock) are added to it.

5 FORMAL INVESTIGATION

With the completion of the pre-survey, the collection results of 82 questionnaires were integrated, and all of them passed the reliability and validity test, and the questionnaire topics were revised reasonably. Therefore, according to the formula for calculating the sample size of simple random sampling in market research and considering the coverage of survey information, the final sample size required for formal survey is 430.

5.1 the Specific Distribution and Recycling of Questionnaires

Table 4 Distribution Table of Questionnaires in Guangxi

city	Number of questionnaires issued (copies)	Proportion (%)
Guilin	118	27.4
Nanning	132	21.8
Liuzhou	94	30.6
Qinzhou	86	20.0

In the formal investigation table 4, we actually sent 528 questionnaires, and 430 valid questionnaires were recovered, with an effective recovery rate of 81.44%.

5.2 Reliability Test

According to the reliability test of 82 questionnaires collected in the pre-investigation, the actual situation of the formal investigation is analyzed the reliability of 430 valid questionnaires was tested again. By analyzing five scales, it is easy to know that the Cronbach coefficient of most scales is greater than 0.65, and it is considered that the internal reliability of the questionnaire is ideal. Cronbach coefficient of each scale is shown in Table 5:

Table 5 Cronbach Coefficient Table of Formal Questionnaire

Item	Cronbach coefficient	number of terms	Reliability evaluation
Question 5	0.648	five	better
Question 10	0.822	five	good
Question 11	0.818	five	good
Question 12	0.761	four	good
Question 13	0.826	five	good

5.3 Validity Test

Referring to the steps of pre-investigation, the validity of 430 valid questionnaires was tested again, and the questionnaires were tested Bartlett Sphere Test was carried out on questions 5, 10, 11, 12 and 13 in the questionnaire, and the KMO coefficient of most questions was greater than 0.65, and the P value was 0.000, which indicated that it was very suitable for factor analysis, so the questionnaire structure was well designed. The specific KMO and Bartlett test coefficients under each scale are shown in Table 6 below:

Table 6 KMO and Bartlett Test Coefficients of Formal Questionnaire

Item	KMO coefficient	Approximate chi-square	significance
Question 5	0.748	236.101	0.000
Question 10	0.690	1737.888	0.000
Question 11	0.685	1655.011	0.000
Question 12	0.776	393.566	0.000
Question 13	0.648	1324.072	0.000

5.4 Data processing

5.4.1 Abnormal data processing

(1)Logical conflict

According to the results of questionnaires filled by respondents from all social groups, the questionnaires with obviously inconsistent answers and contradictory logic before and after are excluded, including choosing a single answer and filling it at will. The items with directional answers can be designed in the questionnaire to identify the effectiveness of the questionnaire. Deal with the questionnaires that take a long time to fill in and have logical conflicts, export the data to Excel and use it to eliminate them. Therefore, through the overall screening of questionnaires, 430 valid questionnaires were finally recovered from 528 questionnaires, and the effective rate of questionnaire recovery reached 81.44%.

(2)Treatment of missing values

There are generally two reasons for the missing values in the questionnaire: one is that the interviewee should have filled in but did not; The other is the loss of some item data due to topic jump. This questionnaire is mainly distributed through the

quiz star network platform, and each question is set as a mandatory question. If you don't answer, you can't continue to answer the next question or submit the questionnaire. This questionnaire has set jump questions. According to the default export rules of the quiz star platform, the option values of the questions that respondents in all social groups jump (unnecessary to fill in) are marked as -3. On this basis, the marked values in the jump questions are checked by using the Data-SelectCases method in SPSS software, and it is found that the missing values of this questionnaire do not exist.

5.4.2 Investigate the overall situation

The basic situation of the sample is as follows: the total number of social groups in the survey sample is 528, among which 98 people have abnormal values, and the actual number of effective surveys is 430. Among the 430 groups surveyed, 47.6% are men and 52.3% are women, and the ratio of men to women is relatively balanced; The age below 18 accounts for 18.1%, 18-25 accounts for 30.4%, 26-35 accounts for 23.4%, 36-45 accounts for 15.8%, and 46 and above accounts for 12.0%. The largest proportion is 18-25 years old; Guilin accounts for 27.4%, Nanning for 30.6%, Liuzhou for 21.8% and Qinzhou for 20.0%. Teenagers account for 16.5%, college students account for 13.9%, employees account for 41.3%, and anime lovers account for 17.2%. Others accounted for 10.9%. The specific distribution of basic information is shown in Table 7:

Table 7 Distribution of Basic Information of Respondents

	classify	Response analysis		Percentage of cases
		Response number	Response percentage	
Gender	Many	205	11.92%	47.6%
	Woman	225	13.08%	52.3%
Age	Under 18 years old	78	4.53%	18.1%
	18-25 years old	131	7.62%	30.4%
	26-35 years old	101	5.87%	23.4%
	36-45 years old	68	3.95%	15.8%
	46 years old and above	52	3.02%	12.0%
Location	Guilin	118	6.86%	27.4%
	Nanning	132	7.67%	30.6%
	Liuzhou	94	5.47%	21.8%
	Qinzhou	86	5.00%	20.0%
Belonging category	teenagers	71	4.13%	16.5%
	college student	60	3.49%	13.9%
	In-service personnel	178	10.35%	41.3%
	Anime fan	74	4.30%	17.2%
	Other	47	2.73%	10.9%
Total		1720	100%	400%

6 DATA ANALYSIS OF SURVEY RESULTS

6.1 Descriptive Statistical Analysis

6.1.1 Gender distribution

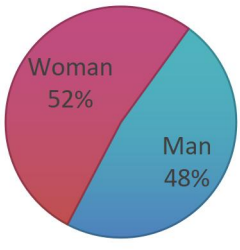


Figure 1 Gender Distribution Map

As shown in Figure 1, Among the social groups surveyed, the number of men is 205, accounting for 47.7%, and the number of women is 225, accounting for 52.3%. The ratio of male to female is about 1:1.

6.1.2 Age distribution

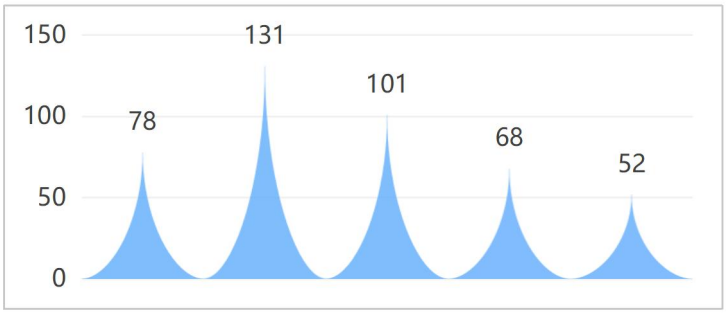


Figure 2 Age Distribution Map

As shown in Figure 2, From left to right, the horizontal axis of the above picture shows the people under 18, 18-25 years old, 26-35 years old, 36-45 years old and over 45 years old. Among the social groups surveyed, there are 78 people under the age of 18, accounting for 18.1%, 131 people aged 18-25, accounting for 30.4%, and 101 people aged 26-35. The proportion is 23.4%; There are 68 people aged 36-45, accounting for 15.8%; The number of people aged 46 and above is 52, accounting for 12.0%. Among the people who participated in the questionnaire survey, the number of people aged 18-25 was the highest, and the number of people aged 46 and above was the least.

6.1.3 Regional distribution

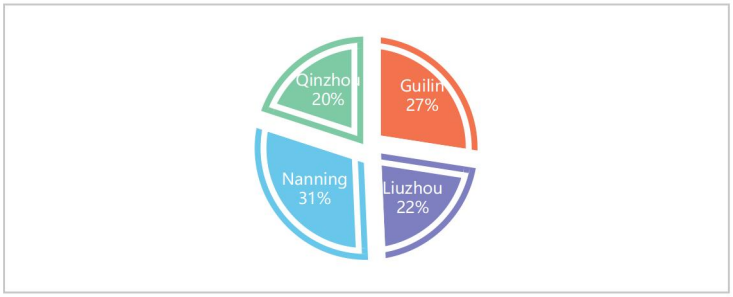


Figure 3 Regional Distribution Map

As shown in Figure 3, Among the social groups surveyed, Qinzhou accounts for 20.0%, accounting for the least; The proportion of people in Guilin is 27.44%; The proportion of people in Nanning is 30.7%, accounting for the largest proportion; The proportion of people in Liuzhou is 21.86%.

6.1.4 Distribution of affiliated groups

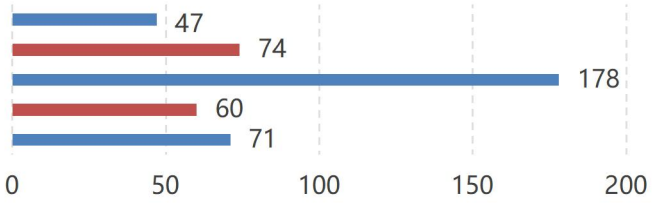


Figure 4 Distribution Map of Belonging Groups

The vertical axis of the above figure 4 shows others, anime lovers, on-the-job personnel, college students and teenagers in turn from top to bottom. Among the social groups surveyed, the proportion of teenagers is 16.5%, the proportion of college students is 13.9%, the proportion of employees is 41.3%, the proportion of anime fans is 17.2%, the proportion of others is 10.9%, and the proportion of teenagers, college students, employees, anime fans and other groups is about 1: 1: 3: 1, among which the proportion of employees is 1: 1.

6.1.5 Distribution of Audience's Understanding of Nezha 2

4.1% of the audience have not heard of Nezha 2, and 52.3% of the audience know more about Nezha 2. 3.7% of the audience don't know the positioning of Nezha 2 in the market, and 54.8% of the audience know the positioning of Nezha 2 in the market. 4.4% of the audience don't understand "Where?" The competitiveness of Nezha 2 in the international market,

54.3% of the audience are relatively familiar with the competitiveness of 2 in the international market; 5.1% of the audience didn't know the influence of Nezha 2 on the development of Guoman, and 56.7% of the audience knew the influence of Nezha 2 on the development of Guoman. 3.4% of the audience didn't know the reason why Nezha 2 broke the circle, and 55.7% of the audience knew the reason why Nezha 2 broke the circle. On the whole, it can be seen that the audience has a certain understanding of Nezha 2 and pays more attention to Nezha 2.

6.1.6 The distribution of the number of times the audience watched Nezha 2

Figure 5 Distribution of the Times of Audience Watching Nezha 2

From left to right, the horizontal axis in the above figure 5 shows 0 times, 1 time, 2 times, 3 times and above. Among the social groups surveyed, 11.3% of the respondents watched Nezha 2 three times or more, accounting for the least; 39.0% of the respondents watched Nezha 2 once, accounting for the largest proportion; 28.1% of the respondents watched Nezha 2 twice and 21.3% watched Nezha 2 0 times. It can be seen that most of the interviewees have seen Nezha 2.

6.1.7 The Distribution of Audience's Emotional Expression in Nezha 2

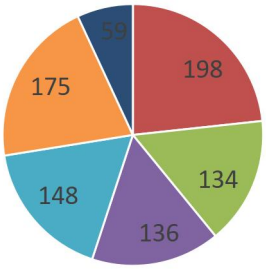


Figure 6 Distribution Map of Audience's Emotional Expression when Watching Nezha 2.

In the above picture, orange, yellow, green, blue, red, and dark blue sequentially represent moving, hilarious, suspenseful, shocking, inspiring, and thinking. As can be seen from Figure 6, the emotional and psychological performance of the audience watching Nezha 2 is touching, hilarious, suspense, shock, motivation and thinking, among which moving, motivation and shock are the main reasons.

6.1.8 The distribution of audience's attitude towards the phenomenon of "Nezha 2" with 2 brushes, 3 brushes and even N brushes

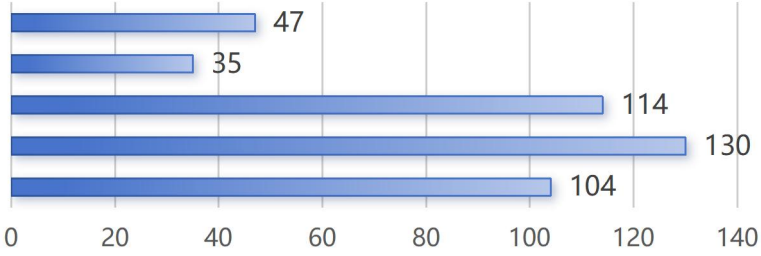


Figure 7 Distribution map of audience's attitude towards the phenomenon of 2-brush, 3-brush or even N-brush "Nezha 2"

From top to bottom, the vertical axis of the above figure 7 shows that it is completely unsupported, not quite agreed, general, understandable and very supportive. When it comes to the audience's attitude towards the phenomenon of "Nezha 2", it can be seen from Figure 7 that 30.23% of the audience think that "it is understandable that the audience may want to know more about the details and connotation of the film", followed by "generally, it may be a follow-up behavior, but the personal choice is understandable", accounting for 26.51%, while very few viewers think it is not quite agreed.

6.1.9 The distribution of the influence of Nezha 2 on the country

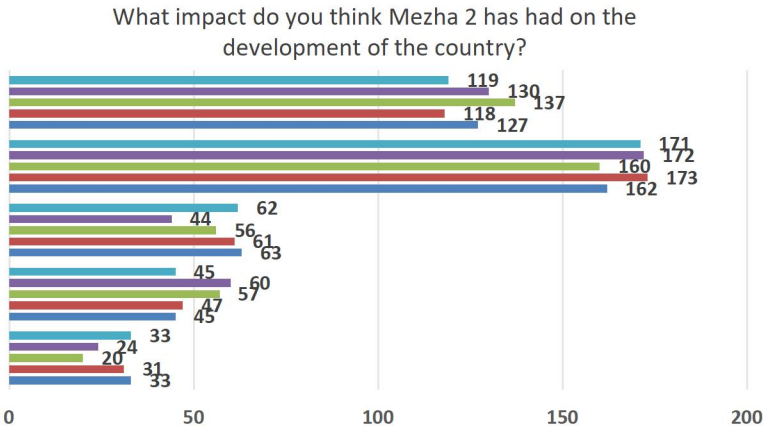


Figure 8 Distribution Map of the Influence of Nezha 2 on the National Spread

From top to bottom, the vertical axis in the above figure 8 indicates strongly agree, agree, neutral, disagree and strongly disagree. The blue-green color, green color, yellow color, orange color and blue color in the picture indicate in turn that the creative enthusiasm inside and outside the industry is stimulated, the narrative style and visual expression of Guoman are innovated, a successful case is provided for Guoman, the audience's confidence in Guoman is enhanced, and the overall level of Guoman is improved. As can be seen from Figure 8, among the influences of Nezha 2 on Guoman, providing a successful case for Guoman to "break the circle" is the most supportive, followed by innovating the narrative style and visual expression of Guoman. But generally speaking, it has enhanced the audience's confidence in Guoman, stimulated the creative enthusiasm inside and outside the industry and improved the overall level of Guoman, which has also affected the development of Guoman to some extent.

6.1.10 The Cause Distribution of "Broken Circle" in Nezha 2

For "Nezha 2", the respondents gave some information about the degree of understanding of "Nezha 2"

The Reasons of "Breaking Circle" in Nezha 2. It contains content and emotional reasons; Technical reasons; Marketing reasons.

(1)The Reasons for the Content and Emotion of "Nezha 2" Breaking the Circle

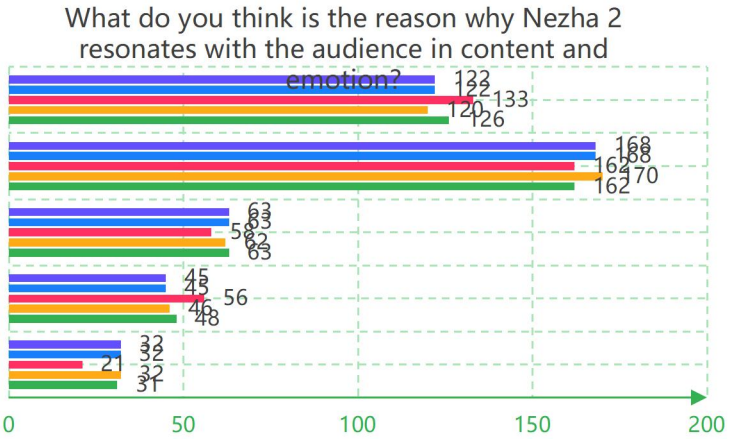


Figure 9 Distribution Map of the Reasons for the Content and Emotion of "Nezha 2" Breaking the Circle

In the figure 9, the vertical axis goes from top to bottom, which means very much agree, agree, be neutral, disagree and disagree.

The dark blue, blue, red, orange and green in the picture respectively indicate that the theme is profound and evocative, the music is beautiful and the atmosphere is set off, the picture is exquisite and the special effects are shocking, the characters are distinctive and the personality is outstanding, and the plot is wonderful and fascinating. As can be seen from Figure 9, 66.98% of the audience strongly agree and agree that "the plot is wonderful and fascinating"; The proportion of viewers who strongly agree and agree with "exquisite pictures and shocking special effects" is 68.60%; 67.44% of the audience strongly agree with and agree with "vivid characters and outstanding personality"; 67.44% of the audience strongly agreed and agreed that "the music is beautiful and sets off the atmosphere"; 67.44% of the audience strongly agree and agree that "the theme is profound and causes thinking test". Among them, "the picture is exquisite and the special effects are shocking" accounts for the most, and "the plot is wonderful and fascinating" accounts for the least.

(2)On the technical level, the reasons that prompted Nezha 2 to "break the circle"

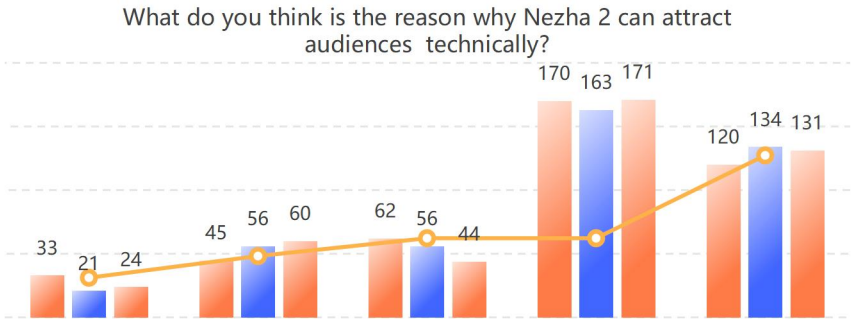


Figure 10 Distribution Map of the Reasons for the "Broken Circle" of Nezha 2 at the Technical Level

In the figure, the horizontal axis goes from left to right to indicate very different opinions, disagreement, neutrality, agreement and strong agreement. The orange, blue, dark orange and broken lines in the picture represent excellent production level, advanced special effects technology, well-designed animation scenes and innovative adaptation of classic IP respectively. As can be seen from Figure 10, the audiences who strongly agree and agree that "the innovative adaptation of classic IP has given new vitality to traditional stories", "excellent production level, including high-level animation rendering and tin character design", "advanced special effects technology has brought shocking visual experience to the audience" and "well-designed animation scenes are smooth and creative" account for 67.21% and 67.44% respectively. Among them, "well-designed animation scenes are smooth and creative" accounts for the most, and "innovative adaptation of classic IP gives new vitality to traditional stories" accounts for the least.

(3) Reasons for the "Breaking Circle" of Nezha 2 in Marketing

Figure 11 Distribution Diagram of the Reasons for the "Broken Circle" of Nezha 2 in Marketing.

In the figure, the horizontal axis goes from left to right to indicate very different opinions, disagreement, neutrality, agreement and strong agreement. Red, yellow, blue, pink and orange in the picture represent online and offline joint promotion, social media topic marketing, cross-border cooperation promotion, launching special promotional materials and activities, and IP derivative development in turn. As can be seen from Figure 11, 67.21% of the audience strongly agree and agree with "online and offline linkage publicity"; The audience of "social media topic marketing" accounted for 67.44%; The proportion of "cross-border cooperation and promotion" audience is70.23%; The audience proportion of "customized content marketing, launching special promotional materials and activities" is 69.07%; The audience proportion of "Mr IP product development" is 70.23%. Among them, "customized content marketing, launching special promotional materials and activities" accounts for the most, and "online and offline linkage publicity" accounts for the least.

6.2 Difference Analysis

6.2.1 Analysis on the Difference of Audience's Watching Times of Nezha 2

(1)Analysis on the Difference of the Times of Different Gender Audiences Watching Nezha 2

In order to explore the gender difference in the number of times the audience watched Nezha 2, the data were tested by independent sample t, and the test results are shown in the following table 8.

Table 8 Difference Analysis Table of the Times of Watching Nezha 2 by Different Gender Audiences

	N (sample size)	M (average)	SD (standard deviation)	t	P
man	205	2.33	0.93	-0.669	0.504
woman	225	2.27	0.94		

From the above table 8, it can be seen that P=0.504>0.05, that is, there is no statistical difference in the number of times the audience of different sexes watched Nezha 2, and it is considered that the number of times the audience of boys and girls watched Nezha 2 is almost the same.

(2)Analysis on the Difference of the Times of Watching Nezha 2 by Audiences of Different Ages

In order to explore the differences in the times of watching Nezha 2 by different age groups, the data were analyzed by one-way ANOVA, and the test results are shown in the following table 9.

Table 9 Difference Analysis Table of the Times of Watching Nezha 2 by Audiences of Different Ages

	N (sample size)	M (average)	SD (standard deviation)	F	P
Under 18 years old	seventy-eight	2.33	0.91	1.265	0.283
18-25 years old	131	2.36	0.96		
26-35 years old	101	2.23	0.93		
36-45 years old	sixty-eight	2.12	0.86		
46 years old and above	fifty-two	2.44	0.98		

From the above table 9, it can be seen that P=0.283>0.05, that is, there is no statistical difference in the times of watching Nezha 2 by audiences of different ages, and it is considered that the times of watching Nezha 2 by audiences of different ages are almost similar.

6.2.2 Analysis on the Difference of Audience's Understanding of Nezha 2

As the audience's understanding of "Nezha 2" is a scale question, it is measured by five dimensions. In order to explore the differences of audience's understanding of "Nezha 2" in terms of gender, region, age and group, pearson chi-square test is conducted on the data respectively, and the specific test results are shown in the following tables.

（1）Analysis on the Difference of Understanding Degree of Different Gender Audiences to Nezha 2

Table 10 Difference Analysis Table of Different Gender Audiences' Understanding of Nezha 2

Dimension	Degree of understanding	Gender		X^2	P
		Woman	Man		
Have you heard about Nezha 2 since it was released?	not familiar with	9	9	1.717	0.788
	know little about	10	14		
	General understanding	42	32		
	Know better	79	72		
	Know very well	85	78		
Do you know the positioning of Nezha 2 in the market?	not familiar with	11	5	1.970	0.741
	know little about	11	11		
	General understanding	41	41		
	Know better	81	73		
	Know very well	81	75		
Do you know the competitiveness of Nezha 2 in the international market?	not familiar with	11	8	2.643	0.619
	know little about	12	16		
	General understanding	38	31		
	Know better	91	74		
	Know very well	73	76		
Do you know the influence of	not familiar with	10	12	2.452	0.653

Nezha 2 on the development of the country?	know little about	12	10		
	General understanding	32	36		
	Know better	99	77		
	Know very well	72	70		
Have you ever understood the reason why Nezha 2 can "break the circle"?	not familiar with	3	12	7.916	0.095
	know little about	12	16		
	General understanding	41	35		
	Know better	90	74		
	Know very well	79	68		

From the above table 10, it can be seen that there is no statistical difference in the understanding of Nezha 2 by audiences of different sexes in the five dimensions, that is, there is no difference in the understanding of Nezha 2 by audiences of different sexes.

(2) Analysis on the Difference of Audiences' Understanding of Nezha 2 in Different Regions

Table 11 Difference Analysis Table of Audiences' Understanding of Nezha 2 in Different Regions

Dimension	Degree of understanding	Location				X^2	P
		Nanning	Liuzhou	Guilin	Qinzhou		
Have you heard about Nezha 2 since it was released?	not familiar with	4	4	7	3	10.427	0.579
	know little about	4	6	5	9		
	General understanding	26	15	23	10		
	Know better	46	33	43	29		
	Know very well	52	36	40	35		
Do you know the positioning of Nezha 2 in the market?	not familiar with	3	7	3	3	27.334	0.007
	know little about	5	7	1	9		
	General understanding	19	15	33	15		
	Know better	49	28	47	30		
	Know very well	56	37	34	29		
Do you know the competitiveness of Nezha 2 in the international market?	not familiar with	7	4	3	5	23.673	0.023
	know little about	4	8	10	6		
	General understanding	12	13	31	13		
	Know better	53	33	44	35		
	Know very well	56	36	30	27		
Do you know the influence of Nezha 2 on the development of the country?	not familiar with	5	8	6	3	15.678	0.206
	know little about	6	3	6	7		
	General understanding	16	12	27	13		
	Know better	55	36	51	34		
	Know very well	50	35	28	29		
Have you ever understood the reason why Nezha 2 can "break the circle"?	not familiar with	4	6	2	3	40.896	0.000
	know little about	8	9	8	3		
	General understanding	17	11	29	19		

Know better	65	23	53	23	
Know very well	38	45	26	38	

From the above table 11, it can be seen that in three of the five dimensions to measure the audience's understanding of Nezha 2, the test p value of the audience's understanding of Nezha 2 in different regions is less than 0.05, but there are still two dimensions in which the test p value of the audience's understanding of Nezha 2 in different regions is greater than 0.05, so it can be considered that the audience's understanding of Nezha 2 in different regions is slightly different.

(3) Analysis on the Difference of Different Groups of Audiences' Understanding of Nezha 2

Table 12 Difference Analysis Table of Different Groups of Audiences' Understanding of Nezha 2

Dimension	Degree of understanding	Group					X^2	P
		other	Anime fan	In-service personnel	college student	teenagers		
Have you heard about Nezha 2 since it was released?	not familiar with	0	0	13	1	4	38.786	0.001
	know little about	8	0	6	3	7		
	General understanding	11	11	32	7	13		
	Know better	15	30	61	27	18		
	Know very well	13	33	66	22	29		
Do you know the positioning of Nezha 2 in the market?	not familiar with	2	0	9	2	3	32.184	0.009
	know little about	7	0	11	1	3		
	General understanding	11	7	37	15	12		
	Know better	15	36	55	24	24		
	Know very well	12	31	66	18	29		
Do you know the competitiveness of Nezha 2 in the international market?	not familiar with	6	0	9	1	3	27.548	0.036
	know little about	5	0	14	4	5		
	General understanding	10	10	27	9	13		
	Know better	18	36	63	24	24		
	Know very well	8	28	65	22	26		
Do you know the influence of Nezha 2 on the development of the country?	not familiar with	5	0	14	0	3	39.654	0.001
	know little about	7	0	8	3	4		
	General understanding	7	8	30	10	13		
	Know better	14	35	64	35	28		
	Know very well	14	31	62	12	23		
Have you ever understood the reason why Nezha 2 can "break the circle"?	not familiar with	6	0	4	1	4	43.146	0.000
	know little about	6	0	16	1	5		
	General understanding	9	7	33	9	18		
	Know better	16	36	62	29	21		
	Know very	10	31	63	20	23		

	well							

From the above table 12, it can be seen that the test P value of the audience's understanding of Nezha 2 in different groups is less than 0.05, so it is considered that there is a statistical difference in general, that is, there is a significant difference in the understanding of Nezha 2 in different groups.

(4) Analysis on the Difference of Understanding of Nezha 2 by Audiences of Different Ages

Table 13 Difference Analysis Table of Audiences' Understanding of Nezha 2 at Different Ages

Dimension	Degree of understanding	Group					X^2	P
		Under 18 years old	18-25 years old	26-35 years old	36-45 years old	46 years old and above		
Have you heard about Nezha 2 since it was released?	not familiar with	4	5	5	3	1	14.804	0.539
	know little about	8	4	2	7	3		
	General understanding	20	17	15	14	8		
	Know better	46	24	41	25	15		
	Know very well	53	28	38	19	25		
Do you know the positioning of Nezha 2 in the market?	not familiar with	4	3	5	3	1	7.881	0.952
	know little about	7	3	3	4	5		
	General understanding	23	14	21	12	12		
	Know better	47	28	34	29	16		
	Know very well	50	30	38	20	18		
Do you know the competitiveness of Nezha 2 in the international market?	not familiar with	3	5	1	6	4	27.548	0.036
	know little about	9	5	6	7	1		
	General understanding	24	14	15	7	9		
	Know better	49	27	46	23	20		
	Know very well	46	27	33	25	18		
Do you know the influence of Nezha 2 on the development of the country?	not familiar with	2	3	6	7	4	16.058	0.449
	know little about	6	5	5	3	3		
	General understanding	23	14	13	14	4		
	Know better	57	29	47	23	20		
	Know very well	43	27	30	21	21		
Have you ever understood the reason why Nezha 2 can "break the circle"?	not familiar with	7	4	0	2	2	27.395	0.037
	know little about	7	3	6	8	4		
	General understanding	19	21	12	19	5		
	Know better	52	24	46	20	22		
	Know very well	46	26	37	19	19		

From the above table 13, it can be seen that the test P values of the audience's understanding of Nezha 2 at different ages are almost all greater than 0.05, and the P value is less than 0.05 in only one dimension, so it is considered that there is no

statistical difference in the whole, that is, there is no difference in the understanding of Nezha 2 among the audience at different ages.

6.2.3 Analysis on the difference of audience's attitude towards the phenomenon of "Nezha 2" with 2 brushes, 3 brushes or even N brushes

(1)Difference Analysis of Different Gender Audiences' Attitudes to the Phenomenon of 2 Brush, 3 Brush or even N Brush Nezha 2

Table 14 Difference Analysis Table of Different Gender Audiences' Attitudes to the Phenomenon of N Brush Nezha

Gender	Very supportive, indicating that the film is rich in content and worth tasting many times.	Understandably, the audience may want to know more about the details and connotation of the film.	Generally speaking, it may be a follow-up behavior, but personal choice is understandable.	I don't quite agree. I think this kind of behavior is a waste of time and money	I don't support it at all. I think we should spend our time on more meaningful things.	Total
Man	45	55	61	19	25	205
Woman	59	75	53	16	22	225
Total	104	130	114	35	47	430

As can be seen from Table 14, the data is subject to independent sample t-test. The results show that P=0.151>0.05. There is no statistical difference in the attitudes of different gender audiences towards the phenomenon of "Nezha 2", which means that the attitudes of different gender audiences towards the phenomenon of "Nezha 2" are basically the same.

(2) An Analysis of the Different Attitudes of Audiences in Different Regions towards the Phenomenon of "Nezha 2"

Table 15 Analysis Table of Differences in Attitudes of Audiences in Different Regions towards the Phenomenon of "Nezha 2"

region	Very supportive, indicating that the film is rich in content and worth tasting many times.	Understandably, the audience may want to know more about the details and connotation of the film.	Generally speaking, it may follow the fashion, but there is no personal choice. undue blame	I don't quite agree. I think this kind of behavior is a waste of time and money	I don't support it at all, and think that time should be spent more meaningfully. On the matter	Total
Guilin	30	33	299	17		118
Qinzhou	23	31	205	7		86
Liuzhou	21	34	238	8		94
Nanning	30	32	4213	15		132
Total	104	130	11435	47		430

As can be seen from Table 15, the data is analyzed by one-way ANOVA. P=0.50>0.05. There is no statistical difference in the attitudes of audiences in different regions towards the phenomenon of "Nezha 2", "2" or "Nezha 2". That is to say, the attitudes of audiences in different regions towards the phenomenon of "2" are generally the same.

(3) An Analysis of the Different Attitudes of Different Groups of Audiences to the Phenomenon of "Nezha 2"

Table 16 Difference Analysis Table of Different Groups of Audiences' Attitudes to the Phenomenon of N Brush Nezha 2

Group	Very supportive, indicating that the film is rich in content and worth tasting many times.	Understandably, the audience may want to know more about the details and connotation of the film.	Generally speaking, it may follow the fashion, but there is no personal choice.undue blame	I don't quite agree. I think this kind of behavior is a waste of time and money	I don't support it at all, and think that time should be spent more meaningfully.On the matter	Total
teenagers	14	18	23	7	9	71
college student	17	24	9	3	7	60
Inservice personnel	47	50	52	13	16	178
Anime fan	16	21	18	7	12	74

Other	10	17	12	5	3	47
Total	104	130	114	35	47	430

As can be seen from Table 16, the data is analyzed by one-way ANOVA. P=0.285>0.05. There is no statistical difference in the attitudes of different groups of audiences towards the phenomenon of "Nezha 2", "2" or "Nezha 2", which means that the attitudes of different groups of audiences towards the phenomenon of "2" are almost the same.

(4) An Analysis of the Different Attitudes of Audiences of Different Ages towards the Phenomenon of "Nezha 2"

Table 17 Analysis Table of Differences in Attitudes of Audiences of Different Ages towards the Phenomenon of "Nezha 2"

Age	Very supportive, indicating that the film is rich in content and worth tasting many times.	Understandably, the audience may want to know more about the details and connotation of the film.	Generally speaking, it may be a follow-up behavior, but personal choice is understandable.	I don't quite agree. I think this kind of behavior is a waste of time and money	I don't support it at all. I think we should spend our time on more meaningful things.	Total
Under 18 years old	10	26	23	9	10	78
18-25 years old	38	46	28	7	12	131
26-35 years old	33	30	21	7	10	101
36-45 years old	9	18	23	10	8	68
46 years old and above	14	10	19	2	7	52
Total	104	130	114	35	47	430

As can be seen from Table 17, the data is analyzed by one-way ANOVA. The results show that P=0.004<0.05. There are statistical differences in the attitudes of different age groups towards the phenomenon of "Nezha 2", that is to say, the attitudes of different age groups towards the phenomenon of "Nezha 2" are different, and the frequency ratio of "Nezha 2", "2" and "2" is strongly supported by the audience aged 18-25 and 26-35.

6.3 Multiple response analysis

In the questionnaire, the audience's views on the measures that domestic animation can attract more audiences and enhance international competitiveness are investigated in the form of two multiple-choice questions from the aspects of content, emotion, technology and international competition. Therefore, when using SPSS to analyze the multiple-choice questions in the questionnaire, this paper adopts the method of frequency and crosstab in multiple response analysis.

Table 18 The audience's views on the measures to promote the development of the country in terms of content and so on

Measures to promote the development of the country in terms of content, emotion and technology	n	Response rate
Lower the viewing threshold and increase the interest.	131	15.29%
Incorporate more local cultural elements.	242	28.24%
Strengthen cross-border cooperation with other cultures.	241	28.12%
Improve picture quality and visual effect.	243	28.35%

As can be seen from the above table 18, the goodness-of-fit test is significant $\chi 2$ =161.056, p=0.068>0.05, which means that there is no obvious difference in the selection ratio of each item, and the differences can be specifically compared by response rate or penetration rate.

At the same time, the data shows that the proportion of improving the picture quality and visual effect is 28.35%, which is higher than other measures. This reflects an active advocacy and practice, that is, to improve the overall texture of the film by improving the quality and visual effects of the film. The proportion of incorporating more local cultural elements is 28.24%, which is also an aspect worthy of attention. In addition, 28.12% and 15.29% increase the interest by strengthening cross-border cooperation with other cultures and lowering the viewing threshold. This shows that in the development of Guoman industry, cross-border cooperation, interesting content and the threshold of viewing are also highly valued by many viewers.

In short, the steady development of Guoman industry in the future needs various efforts and measures. In addition to improving the picture quality and visual effect, it is also necessary to integrate more local cultural elements, strengthen cross-border cooperation with other fields of culture, lower the viewing threshold and increase interest. Only by comprehensively applying all kinds of measures can we effectively contribute to various industries such as national culture, film industry, cultural tourism and so on.

The content, technology and other aspects to promote the development of the country's views and the basic information of the audience (gender, age, group), and get the proportion of the audience's views on content, technology and other measures to promote the development of the country.

First of all, the cross table of measures to promote the development of the country from different ages, contents and technologies is as follows as shown in table 19.

Table 19 Cross-table of Views of Different Ages on Measures to Promote the Development of the Country in Terms of Content and Technology

Measures to promote the development of the country in terms of content, emotion and technology	Age					Total（n=430）
	Under 18 years old	18-25 years old	26-35 years old	36-45 years old	46 years old and above	
Lower the viewing threshold and increase the interest	21.79%	35.11%	33.66%	27.94%	28.85%	30.47%
Incorporate more local cultural elements.	52.56%	54.96%	56.44%	58.82%	61.54%	56.28%
Strengthen cross-border cooperation with other cultures.	57.69%	54.96%	56.44%	61.76%	48.08%	56.05%
Improve picture quality and visual effect.	60.26%	57.25%	56.44%	48.53%	59.62%	56.51%
	chi-square test：$\chi 2=5.803$ $p=0.926$					

Secondly, the cross table of different gender, content and technology to promote the development of the country is shown in the following table 20.

Table 20 Cross-table of Measures to Promote the Development of the Country by Gender, Content and Technology

Measures to promote the development of the country in terms of content, emotion and technology	Gender		Total（n=430）
	Man（n=205）	Woman（n=225）	
Lower the viewing threshold and increase the interest.	33.66%	27.56%	30.47%
Incorporate more local cultural elements.	53.66%	58.67%	56.28%
Strengthen cross-border cooperation with other cultures.	58.05%	54.22%	56.05%
Improve picture quality and visual effect.	51.71%	60.89%	56.51%

Finally, the cross table of different groups and measures to promote the development of the country in terms of content and technology is shown in the following table 21.

Table 21 Cross Table of Different Groups and Measures to Promote the Development of the Country in Terms of Content and Technology

Measures to promote the development of the country in terms of content, emotion and technology	Group					Total（n=430）
	Teenagers	college student	In-service personnel	Anime fan	Other	
Lower the viewing threshold and increase the interest.	30.99%	38.33%	28.65%	27.03%	31.91%	30.47%

Incorporate more local cultural elements.	47.89%	60.00%	63.48%	48.65%	48.94%	56.28%
Strengthen cross-border cooperation with other cultures.	61.97%	55.00%	53.37%	58.11%	55.32%	56.05%
Improve picture quality and visual effect.	52.11%	61.67%	57.30%	54.05%	57.45%	56.51%

As can be seen from the following table 22, the goodness-of-fit test is significant $\chi 2 = 276.053$, $p=0.000<0.05$, which means that the selection ratio of each item is obviously different, and the differences can be specifically compared by response rate or penetration rate.

Table 22 The Audience's Views on the Measures to Promote the Development of the Country at the Level of International Competitiveness

Measures to promote the development of the country in terms of international competitiveness	n	Response rate
Learn from the experience of foreign excellent works	183	18.83%
Create a story with more China characteristics	266	27.37%
Strengthen international cooperation and improve the production level.	318	32.72%
Launch a multilingual version to attract the world.	205	21.09%

At the same time, the data shows that the proportion of strengthening international cooperation and improving production level is 32.72%, which is higher than other measures. This reflects a positive advocacy and practice, that is, by strengthening international cooperation and improving the production level, the overall texture of the film can be improved. The proportion of stories with more China characteristics is 27.37%, which is also an aspect worthy of attention. In addition, 21.09% and 18.83% respectively launched multilingual versions, attracted the world and learned from the experience of foreign excellent works. This shows that in the development process of Guoman industry, multilingual is introduced. The Chinese version is also highly valued by many viewers.

Cross-table analysis is made on the views of the measures to promote the development of the country at the level of international competitiveness and the basic information (gender, age and group) of the audience, and the proportion of the views of the audience with different basic conditions on the measures to promote the development of the country at the international level is obtained.

First of all, the cross table of measures to promote the development of the country from different ages and international levels is shown in the following table 23.

Table 23 Cross Table of Views on Measures to Promote the Development of the Country at Different Ages and at the International Level

Measures to promote the development of the country at the international level	Age					Total（n=430）
	Under 18 years old	18-25 years old	26-35 years old	36-45 years old	46 years old and above	
Learn from the experience of foreign excellent works	48.72%	48.09%	38.61%	35.29%	36.54%	42.56%
Create a story with more China characteristics	51.28%	62.60%	62.38%	66.18%	67.31%	61.63%
Strengthen international cooperation and improve the production level.	78.21%	76.34%	73.27%	77.94%	57.69%	73.95%
Launch a multilingual version to attract the world.	46.15%	42.75%	51.49%	41.18%	63.46%	47.67%
	chi-square test：$\chi 2=11.679$ $p=0.472$					

Secondly, the cross table of different gender and international measures to promote the development of the country is shown in the following table 24.

Table 24 Cross Table of Gender and International Measures to Promote National Development

Measures to promote the development of the country at the international level	Gender		Total（n=430）
	Man（n=205）	Woman（n=225）	
Learn from the experience of foreign excellent works	41.46%	43.56%	42.56%
Create a story with more China characteristics	60.00%	63.11%	61.63%
Strengthen international cooperation and improve the production level.	69.76%	77.78%	73.95%
Launch a multilingual version to attract the world.	54.15%	41.78%	47.67%

Finally, the cross table of different groups and international measures to promote the development of the country is shown in Table 25 below.

Table 25 Cross Table of Measures for Promoting National Development by Different Groups and International Level

Measures to promote the development of the country at the international level	Group					Total（n=430）
	Teenagers	college student	In-service personnel	Anime fan	Other	
Learn from the experience of foreign excellent works	46.48%	53.33%	38.76%	43.24%	36.17%	42.56%
Create a story with more China characteristics	53.52%	61.67%	66.85%	55.41%	63.83%	61.63%
Strengthen international cooperation and improve the production level.	76.06%	80.00%	69.10%	72.97%	82.98%	73.95%
Launch a multilingual version to attract the world.	49.30%	40.00%	54.49%	50.00%	25.53%	47.67%

7 COGNITION AND INFLUENCE ANALYSIS OF NEZHA 2 BASED ON STRUCTURAL EQUATION MODEL

In this survey, the reliability of 430 valid data collected was tested, and the representative core factors were extracted by exploratory factor analysis, so as to build a structural equation model. Considering the relationship among the representative factors, the respondents in four places in Guangxi (Guilin, Nanning, Liuzhou and Qinzhou) were analyzed for their cognition of Nezha 2 and the influence of Nezha 2 on the future development of the country.

7.1 Brief Introduction of Structural Equation Model

7.1.1 Concept of structural equation model

Structural equation model (SEM) is a method to establish, estimate and test causality model. The model contains both observable obvious variables and potential variables that cannot be directly observed. Structural equation model can replace multiple regression, path analysis, factor analysis, covariance analysis and other methods to clearly analyze the role of individual indicators on the population and the relationship between individual indicators. Compared with traditional analysis methods, structural equation model can explain as many variations of variables as possible while understanding the covariant relationship between variables. When you want to study the causal relationship between complex variables, it is most appropriate to use structural equation model.

The construction of structural equation model is usually divided into four main parts, including model hypothesis, reliability and validity test, model establishment and model revision. Generally speaking, the establishment of structural equation model first needs to determine the design of external latent variable factors and internal latent variable factors. If the established structural equation model has a good fitting effect, that is, it passes the evaluation criteria of model fitting, there is no need to modify the model.

7.1.2 Basic form of structural equation model

Structural equation model usually has two definitions: ① structural equation model = factor analysis+path analysis; ②

Structural equation model = measurement model+structural model, and there is no difference in essence.
(1)Measurement model
The measurement model is confirmatory factor analysis (CFA), which describes the relationship between the observed variable and the latent variable, and measures the measurement effect of the explicit variable (that is, the measuring tool) on the latent variable (the magnitude and significance of the load of the observed variable on the latent variable). When doing validity analysis, what we usually expect is to find that the observed variables are significantly loaded on theoretically related latent variables, but not on unrelated latent variables by confirmatory factor analysis. The specific model expression is:

$$\begin{cases} x = \Lambda_x + \delta \\ y = \Lambda_y + \varepsilon \end{cases} \tag{3}$$

Where x is a vector composed of exogenous indicators, y is a vector composed of endogenous indicators, Λ_x and Λ_y are factor load matrices, ε and δ are error terms.
(2)structural model

$$\eta = B\eta + \Gamma\xi + \zeta \tag{4}$$

Where B is the coefficient matrix of sum for η and η, which also becomes the path coefficient; Γ represents the coefficient matrix between η and η, and ζ represents the error term.

7.2 the Establishment of Structural Equation Model

7.2.1 Factorial analysis

Before establishing the structural equation model, we need to use SPSSAU online software to make exploratory factor analysis of all the valid data to be investigated, so as to explore the audience's understanding of Nezha 2, the factors that Nezha 2 has influenced the development of Guoman and the reasons for its success, find out the potential variables in the theory, reduce the number of questions, and get better variables.
First of all, KMO test and Bartlett spherical test are needed. The KMO value in the test results is 0.822, and the P value is less than 0.05. The output four factors explain that the cumulative percentage of 24 variables is 87.194%, so it can be considered that there is a certain correlation between the items. The specific KMO and Bartlett spherical test analysis results are shown in the following table 26:

Table 26 KMO and Bartlett Spherical Test Analysis Results

project	numerical value
KMO sampling suitability quantity	0.822
Approximate chi-square of Bartlett spherical test	19848.67
freedom	276
significance	0.000

7.2.2 Setting of observation variables

According to the rotated factor load matrix, three factors are defined, which are: the audience's understanding of Nezha 2, the reasons for the success of Nezha 2, and the influence of Nezha 2 on the development of Guoman. These three factors as a latent variable, the corresponding observed variables are shown in the following table 27:

Table 27 System index of structural equation model

Latent variable	Observed variable	Item
The audience's understanding of Nezha 2.	Have you heard about Nezha 2 since it was released?	A1
	Do you know the positioning of Nezha 2 in the market? Do you know the competitiveness of Nezha 2 in the international market?	A2
	Do you know the influence of Nezha 2 on the development of the country?	A3
	Have you ever known the reason why Nezha 2 was successful?	A4
The Reasons for the Success of	The plot is wonderful and fascinating.	B1

Nezha 2	The characters are vivid and have outstanding personalities.	B2
	The picture is exquisite and the special effects are shocking.	B3
	The music is beautiful and sets off the atmosphere.	B4
	The innovative adaptation of classic IP has given new vitality to traditional stories.	B5
	Excellent production level, including high-level animation rendering and detailed character design.	B6
	Advanced special effects technology has brought a shocking visual experience to the audience.	B7
	Well-designed animation scenes are smooth and creative.	B8
The Influence of Nezha 2 on the Development of Guoman	Improve the overall level of the country.	C1
	Enhance the audience's confidence in Guoman.	C2
	It provides a successful case for Guo Man's "breaking the circle"	C3
	Innovating the narrative style and visual expression of Guo Man.	C4
	Inspired creative enthusiasm inside and outside the industry.	C5
	Online and offline linkage publicity	C6
	Social media topic marketing	C7
	Cross-border cooperation and promotion	C8
	Customize content marketing, and launch featured promotional materials and activities.	C9

7.2.3 Reliability and validity test of data

After the statistics of dimensionality reduction, the total number of variables finally included can be explained by three common factors, and then the reliability of these three common factors is tested by SPSSAU, and the results are as follows: Table 28:

Table 28 Reliability Test Table of Latent Variables

Aspect	Cronbach's α	Number of terms
The audience's understanding of Nezha 2	0.648	5
The Reasons for the Success of Nezha 2	0.908	8
The Influence of Nezha 2 on the Development of Guoman	0.922	9

According to Table 28, Cronbach's Alpha coefficient values of all latent variables are above 0.60, indicating that latent variables are reliable and can be included in the path diagram of structural equation model.

7.2.4 Model setting

After setting the relationship between latent variables, the online software SPSSAU is used to set up the causality path d iagram as required, and the specific results are as follows figure 12:

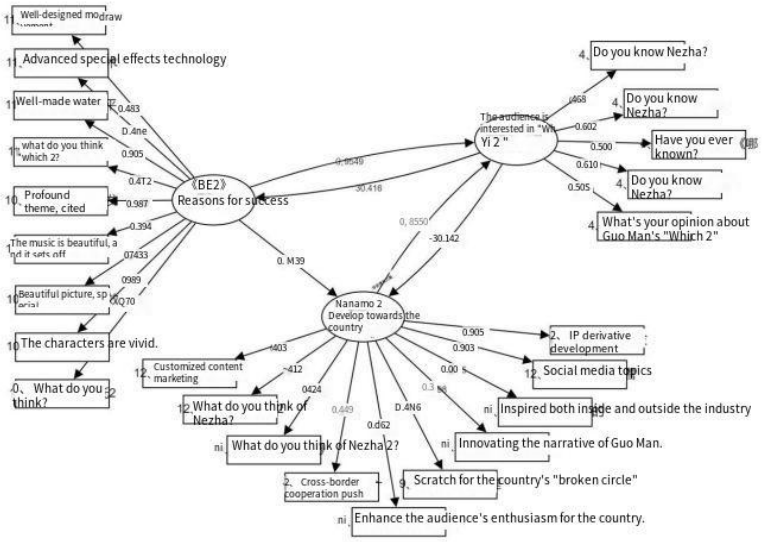


Figure 12 Road Map of Research Theory Model

7.2.5 Fitting degree analysis and hypothesis test of the model

(1)Fitting degree analysis of the model

Estimating the set model parameters with sample data is the primary task to test the structural equation model, but it is t he goal of parameter estimation to generate another covariance matrix of the observed variables, so as to make the covar iance matrix of the observed variables as close as possible. When the two covariance matrices are very close, we think t hat the model fits the data.

In this paper, the specific results of fitting index calculated by SPSSAU online software are as follows:

Table 29 Fitting Index of Structural Equation Model

Fitting index	CMIN/DF	GFI	RMSEA	RMR	CFI	NFI	NNFI
numerical value	1.225	0.919	0.017	0.046	0.936	0.928	0.928

According to Table 29, the value of chi-square/degree of freedom, that is, CMIN/DF, is 1.225, which is less than the crit erion 3, so the data of this model is well fitted. The GFI value is 0.919, which is greater than the criterion of 0.9, indicati ng that the fitting degree is good. The RMSEA value is 0.017, which is less than the criterion of 0.1, indicating that the f itting degree is good. RMR value is 0.046, which is less than the criterion of 0.05. The CFI value is 0.936, which is grea ter than the criterion of 0.9, indicating that the fitting degree is good. The NFI value is 0.928, which is greater than the c riterion of 0.9, indicating that the model fits well. The value of NNFI is 0.928, which is greater than the criterion of 0.9, indicating that the model has a good fitting degree. To sum up, most of the evaluation indexes of this model have reache d the standard.

(2)Hypothesis test of model

In this paper, the test results of each path obtained by AMOS26.0 software are shown in the following table 30:

Table 30 Test Results of Path Hypothesis

null hypothes is	path			P value	result
H4	The Reasons for the Success of Nezha 2	<—	The Influence of Nezha 2 on the Development of Guoman	***	found
H5	The Influence of Nezha 2 on the Development of Guoman	<—	The audience's understanding of Nezha 2.	0.232	Not found
H6	The Influence of Nezha 2 on the Development of Guoman	<—	Have you ever understood the reason why Nezha 2 can "break the circle"?	***	found
			(The following 22 paths are thus omitted)		

According to the above table, we can find that there is a significant relationship between the audience's understanding of Nezha 2 and the reason for the success of Nezha 2, and the reason for the success of Nezha 2 is the audience's

understanding of Nezha 2.

Nezha 2 has a significant impact on the development of the country, and Nezha 2 has a significant impact on the development of the country and the audience's understanding of Nezha 2. However, there is no significant relationship between the audience's understanding of Nezha 2 and the impact of Nezha 2 on the development of the country, with a p value of 0.232>0.05. The relationship between the observed variables and the latent variables is very significant, so the latent variables are incomplete. Therefore, it can be considered that most hypotheses have been tested, but there is still one hypothesis that failed to pass the test.

7.2.6 Modification of model

(1)Fitting degree analysis of the modified model

As shown in Figure 13, According to the modified index, the model structure is more reasonable by releasing paths or adding new paths. In order to improve the fitting degree of the mold structure, the modified goodness of fit is shown in the following table after being modified by AMOS26.0 software:

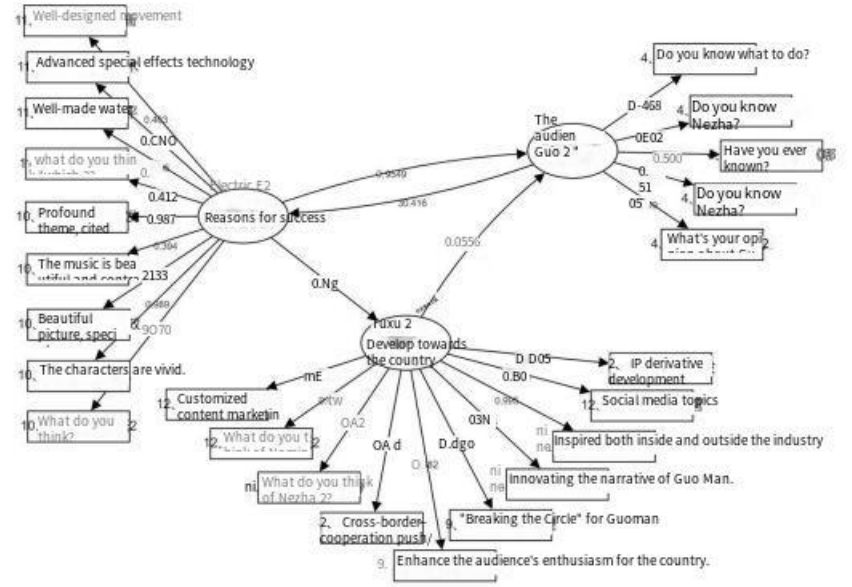


Figure 13 The road map of the revised research theory model

The modified goodness of fit is shown in the following table 31:

Table 31 Fitting Index of Modified Structural Equation Model

Fitting index	CMIN/DF	GFI	RMSEA	RMR	CFI	NFI	NNFI
numerical value	1.226	0.969	0.017	0.044	0.998	0.988	0.998

As can be seen from Table 31, the value of chi-square/degree of freedom, that is, CMIN/DF, is 1.226, which is less than the criterion 3, so the data of this model is well fitted. The GFI value is 0.969, which is greater than the criterion of 0.9, indicating that the fitting degree is good. The RMSEA value is 0.017, which is less than the criterion of 0.1, indicating that the fitting degree is good. RMR value is 0.044, which is less than the criterion of 0.05. The CFI value is 0.998, which is greater than the criterion of 0.9, indicating that the fitting degree is good. The NFI value is 0.988, which is greater than the criterion of 0.9, indicating that the model fits well. The value of NNFI is 0.998, which is greater than the criterion of 0.9, indicating that the model has a good fitting degree. To sum up, most of the evaluation indexes of the model have reached the standard, which shows that the modified model has a good fitting degree.

(2) Hypothesis test of the modified model

Table 32 Test Results of Revised Path Hypothesis

null hypothesis	Path			P value	results
H4	The Reasons for the Success of Nezha 2	<—	The Influence of Nezha 2 on the Development of Guoman	***	found
H5	The Influence of Nezha 2 on the Development of Guoman	<—	The audience's understanding of Nezha 2.	***	found
H6	The Influence of Nezha 2 on the Development of Guoman	<—	Have you ever understood the reason why Nezha 2 can "break the circle"?	***	found

	(The following 22 paths are thus omitted)

According to the above table 32, we can find that there is a significant relationship between the audience's understanding of Nezha 2 and the reason for the success of Nezha 2, and the reason for the success of Nezha 2 is the audience's understanding of Nezha 2.

Nezha 2 has a significant impact on the development of the country, and Nezha 2 has a significant impact on the development of the country and the audience's understanding of Nezha 2. The relationship between observed variables and latent variables is very significant, so the original hypothesis is rejected among latent variables and between latent variables and observed variables. Therefore, it can be considered that all hypotheses have been tested.

7.2.7 Interpretation and analysis of structural equation model results

(1)Relationship between latent variables

The coefficient between latent variables indicates the degree to which the change of one variable causes the change of other variables. Because the estimated value of residual term is relatively small and usually meaningless, its influence can be directly ignored, so this paper has not conducted in-depth and detailed research and discussion on the error term.

It is easy to see the path coefficient between latent variables from the path diagram of structural equation model, and the relationship between most of the latent variables has a significant positive correlation. Among them, there is a significant positive correlation between the audience's understanding of Nezha 2 and the reasons for the success of Nezha 2, and its path coefficient is greater than 0, showing a significance of 0.01 level. At the same time, there is a significant positive correlation between the influence of Nezha 2 on the development of the country, the reasons for the success of Nezha 2, and the audience's understanding of Nezha 2. The path coefficients between them are all greater than 0, and they all show a significance of 0.01 level. There is a significant positive correlation between the influence of Nezha 2 on the development of the country and the audience's understanding of Nezha 2, and its path coefficient is also greater than 0, which also shows the significance of 0.01 level. From the above analysis results, we can see that there is a causal relationship between the reasons for the success of Nezha 2, the influence of Nezha 2 on the development of the country and the audience's understanding of Nezha 2.

The regression coefficient between the reasons for Nezha 2' s success and the audience's understanding of Nezha 2 is 0.9549, which indicates that if the reason for Nezha 2' s success is increased by one percentage point, the audience's understanding of Nezha 2 will be increased by 0.9549 percentage point. It can be seen that when Nezha 2 performed well in elaborate animation, advanced special effects technology, excellent production level, profound theme, moving music, exquisite pictures and vivid characters, the audience understood Nezha 2. The degree will be significantly improved, thus enhancing the audience's awareness and interest in the country.

The regression coefficient between the reasons for the success of Nezha 2 and the influence of Nezha 2 on the development of the country is 0.199, which shows that the influence of Nezha 2 on the development of the country will increase by 0.199 percentage points for every percentage point increase in the reasons for the success of Nezha 2. It can also be seen that when Nezha 2 performs well in all aspects, its successful experience will have a positive impact on Guoman industry, such as promoting the development of IP derivatives, strengthening the topic marketing of social media, and stimulating the innovation vitality inside and outside the industry.

The regression coefficient between the audience's understanding of Nezha 2 and the influence of Nezha 2 on the development of the country is 0.9550, which shows that the influence of Nezha 2 on the development of the country will increase by 0.9550 percentage points for every percentage point increase in the audience's understanding of Nezha 2. That is, when the audience has a deeper understanding of Nezha 2, they are more likely to realize the positive impact of Nezha 2 on the development of Guoman, thus further promoting the development of Guoman industry.

(2)The relationship between observed variables and latent variables

In the audience's understanding of Nezha 2, the path coefficients of the observation variable "Do you know Nezha 2" are 0.468, 0.602, 0.509, 0.515 and 0.509 respectively. This shows that these observation variables have obvious influence on the audience's understanding of Nezha 2, among which the observation variable "Do you know Nezha 2" has the most significant influence. This shows that, compared with other options, the audience's basic knowledge of Nezha 2 plays a key role in their overall understanding.

In the influence of Nezha 2 on the development of Guoman, the path coefficients of IP derivative development, social media topic marketing, stimulating innovation inside and outside the industry, innovating the narrative style of Guoman, providing new ideas for "breaking the circle" of Guoman and enhancing the audience's attention and love for Guoman are all 0.995. This shows that these observation variables have a very obvious influence on the development of Nezha II, and each observation variable can be regarded as the most important influence. This shows that the success of Nezha 2 lies not only in the film itself, but also in its positive role in industrial chain extension, market promotion and industry innovation.

Among the reasons for the success of Nezha 2, the well-designed animation, advanced special effects technology, excellent production level, profound theme resonance, beautiful music, exquisite pictures and excellent special effects, vivid characters, and the path coefficients of the audience's overall evaluation of Nezha 2 are 0.403, 0.449, 0.995, 0.412, 0.394 and 0.44 respectively. This shows that these observed variables have obvious influence on the success of Nezha 2, among which the audience has the most significant influence on the overall evaluation of Nezha 2. This shows that the success of Nezha 2 is the result of many factors, but it ultimately comes down to the recognition and love of the audience.

8 SUMMARY AND SUGGESTIONS

8.1 Research Conclusions

Based on the data of online questionnaire survey, this paper tested the reliability and validity of the questionnaire, and then analyzed the data of the questionnaire by using the frequency and cross-table method and structural equation model in difference analysis and multiple response analysis, and analyzed the audience's understanding of Nezha 2, the reasons why Nezha 2 broke the circle, and put forward some suggestions on the future development direction of Guoman. The purpose of this paper is to explore the understanding of Nezha 2 by groups in four places in Guangxi. The relationship between the ways to break the circle in Nezha 2, and understand the mutual influence mechanism between them.

1.Based on the investigation of the audience's cognition of Nezha 2 in four regions of Guangxi, this study found that the audience did have cognitive bias towards Nezha 2. At the same time, we conducted independent sample T-tests on the number of times the audience watched Nezha 2, their understanding of Nezha 2, and their attitudes towards the phenomenon of brushing Nezha 2, brushing 2 3 or even brushing 2 n, respectively, with respect to gender, region, age and group. The results show that the number of times that boys and girls watch Nezha 2 is almost the same; There is little difference in the number of times viewers watch Nezha 2 in different regions; Audiences of different ages watch Nezha 2 almost the same times; There is no statistical difference in the number of times the audience of different groups watched Nezha 2; There is no difference in the understanding of Nezha 2 between audiences of different sexes; There are some statistical differences in the audience's understanding of Nezha 2 in different regions; There are significant differences in the understanding of Nezha 2 among different groups of audiences; There is no significant difference in the understanding of Nezha 2 among audiences of different ages; Different gender audiences have basically the same attitude towards the phenomenon of 2-brush, 3-brush or even N-brush "Nezha 2"; Audiences in different regions generally have the same attitude towards the phenomenon of "Nezha 2". Different groups of audiences have almost the same attitude towards the phenomenon of N-brushing Nezha 2; There are statistical differences in the attitudes of different age groups towards the phenomenon of "Nezha 2", "Nezha 2" and "Nezha 2". That is to say, different age groups have different attitudes towards the phenomenon of "2". Among them, the 18-25-year-old and 26-35-year-old audiences are very supportive of the phenomenon of "2, 3-year-old or even N-brush", which is higher than that of other age groups.

2.The research results of structural equation model reveal that when Nezha 2 performs well in well-designed animation, advanced special effects technology, excellent production level, profound theme, moving music, exquisite pictures and vivid characters, the audience's understanding of Nezha 2 will be significantly improved, thus enhancing the audience's cognition and interest in the country. When Nezha 2 performs well in all aspects, its successful experience will have a positive impact on Guoman industry, such as promoting the development of IP derivatives, strengthening the topic marketing of social media, stimulating the innovation vitality inside and outside the industry, innovating the narrative style of Guoman, providing new ideas for Guoman to "break the circle" and enhancing the audience's attention and love for Guoman. When the audience has a deeper understanding of Nezha 2, they are more likely to realize the positive impact of Nezha 2 on the development of Guoman, thus further promoting the development of Guoman industry.

8.2 Deficiencies

1.Due to the limitation of objective factors. The respondents in this study are only distributed in four cities in Guangxi, and the scope of the respondents is relatively small, which may have a certain impact on the overall understanding of Nezha 2 by the research audience, and may not represent the specific understanding of Nezha 2 by the national audience.

2.This study only makes a preliminary investigation on the reasons for the "broken circle" of Nezha 2 and its influence on the future development of the country, and further investigation is needed to see whether there will be potential influences among the factors in the later period.

3.For some reasons, all the questionnaires collected are from online questionnaires, and they are not distributed face to face offline, which may have a certain impact on the quality of the questionnaires.

8.3 Suggestions

In view of the reasons for the success of Nezha 2 and its influence on the development of the country, the relevant suggestions can be put forward from two aspects: content, emotion, technology, marketing and international competitiveness.

8.3.1 Suggestions on content, emotion, technology and marketing

(1)Lower the viewing threshold and increase the interest of the film;
(2)Incorporate more local cultural elements;
(3)Strengthen cross-border cooperation with other fields of culture;
(4)Improve the picture quality and visual effect.

8.3.2 Suggestions on international competitiveness

(1)Learn from the experience of foreign excellent works;
(2)Create a story with more China characteristics;
(3)Strengthen international cooperation and improve the production level;
(4)Launch a multilingual version to attract the world.

COMPETING INTERESTS

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

REFERENCES

[1] Duan Yuhan, Chen Xinyue. Multi-dimensional exploration and path optimization of IP adaptation and creation of classic myths from the perspective of traditional culture —— Taking "Nezha's Devil Children Naughting the Sea" as an example. Journalist, 2025(05): 85-87.

[2] Li Zihui, Zhou Yiteng. Narrative innovation of heroic movies and research on the value of the times-taking "Nezha's Devil Children Naughting the Sea" as an example. Today Media, 2025, 33(05): 76-80.

[3] Tao Juan. "Nezha's Devil Children Naughting the Sea" and the development path of China's animated films. Film Literature, 2025(08): 112-116.

[4] Liang Qian. Digital technology to empower Chinese excellent traditional culture "two creations" of the practical exploration-taking "Nezha's Magic Boy Naughting the Sea" as an example. Contemporary Economic Management, 2025: 1-15.

[5] Zhou D .Cultural Narrative, Symbol System and Cross-Cultural Communication: the Globalization Path of the Chinese-Style Animation "Nezha: The Devil Boy Conquers the Dragon King".Economic Society and Humanities, 2025, 2(1).

[6] Wang Yuandi, James Zhou. A Study on the Self-confidence Effect of Starting Economy —— Based on the case of DeepSeek and the movie Nezha 2. Journal of China University of Mining and Technology (Social Science Edition), 2025: 1-17.

[7] Pan Yingying. The theoretical configuration and functional mechanism of collective memory in constructing Chinese cultural identity-A case study of "Nezha's Devil Children Naughting the Sea". United Front Studies, 2025: 1-12.

[8] Chen L .Subject Consciousness and Language Trend of Voice Actors in the Film Nezha the Devil Comes Down.Art and Performance Letters, 2024, 5(3).

[9] Zhang Shaohui, Yin Gaojian. "Nezha 2" box office myth reveals the underlying logic of marketing. Enterprise Management, 2025(04): 32-35.

[10] Wei Jun, Lu Yanchen. Narrative Breakthrough and New Paradigm Construction of China Film and Television IP from the Perspective of Cultural Security-Based on the Comparative Analysis of Nezha and Marvel Comics Series Films. United Front Studies, 2025: 1-15.

AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3017

PREDICTION OF AIR QUALITY INDEX BASED ON ARIMA AND LSTM MODELS—TAKE GUILIN AS AN EXAMPLE

Dan Chen
Department of Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: 1274356593@qq.com

Abstract: With the acceleration of urbanization and the improvement of industrialization, air quality has become an important factor affecting the quality of life and health of urban residents. Under this background, the air quality problem has become increasingly prominent and has become a key factor affecting the quality of life and health of urban residents. Therefore, how to accurately predict the air quality index (AQI) in order to take timely and effective measures to protect the health of residents has become a major challenge we face. This study takes Guilin city as the object, and constructs an air quality prediction model by integrating big data and artificial intelligence technology. Based on the historical monitoring data such as PM2.5 and NO2, a training set/test set division method is adopted, and the ARIMA time series model and the LSTM depth learning network are comprehensively applied: the former captures the trend and seasonal characteristics of the data, and the latter handles the complex sequence relationship. By introducing the rolling prediction mechanism to update the model parameters in real time, the dynamic adaptability of the prediction system is effectively improved. Experiments show that the hybrid model significantly improves the prediction accuracy and stability of AQI, and provides a scalable technical scheme for urban air quality management.
Keywords: LSTM model; ARIMA model; Rolling forecast; Air quality index forecast

1 INTRODUCTION

1.1 Research Background and Significance

At present, the overall situation of air quality in our country is not optimistic. The problem of air environmental pollution, with fine particulate matter (PM2.5) as the main pollutant, is becoming increasingly serious. It not only has a great impact on air quality, but also causes serious harm to human health. With the support of big data technology, we can obtain massive air quality data, including concentration of various pollutants, meteorological data, traffic flow, etc. These data provide abundant information sources for the prediction of air quality index and scientific basis for the formulation of environmental policies and the implementation of environmental protection measures. As a classical time series analysis method, ARIMA model has a wide range of applications in the field of air quality prediction. LSTM model, as a special cyclic neural network (RNN), is especially suitable for dealing with long-term dependence problems in sequence data. Therefore, it also has unique advantages in air quality prediction. By comparing ARIMA and LSTM models with rolling prediction, we can predict the change trend of air quality index more accurately. As a tourist resort and an ecological livable city, the air quality of Guilin has always attracted much attention. However, with the acceleration of urbanization and the improvement of industrialization level, Guilin is also facing the problem of air quality decline. Therefore, taking Guilin City as an example, the study on air quality index prediction will not only help to improve the air quality monitoring and early warning capabilities of Guilin City, but also provide reference and reference for air quality prediction of other cities. To sum up, the comparative study of air quality index prediction based on ARIMA and LSTM models has important theoretical significance and practical application value. Taking Guilin as an example, the research background not only reflects the importance and urgency of air quality prediction, but also shows the application prospect of big data and artificial intelligence technology in the field of environmental management.

1.2 Research Status at Home and Abroad

Driven by big data and artificial intelligence, the prediction of air quality index (AQI) has become a research hotspot in the field of environmental science and computational intelligence. ARIMA and LSTM are two commonly used prediction models. They have their own advantages in processing time series data. The better model is determined by rolling prediction. Fang Xiaoping and others took the air quality index (AQI) of 100 cities in the Yangtze River Economic Belt as the research object. Aiming at the characteristics of air quality related data, they established the PSOGSA-LSTM combined prediction model to test the prediction accuracy of the model in three aspects, and compared the prediction results with the traditional LSTM model. Finally, they applied it to the air quality index prediction of 100 cities in the Yangtze River Economic Belt in the next seven days[1]. Based on the monitoring data of air pollutants in Taiyuan City from 2014 to 2017, Zheng Yangyang et al. first used the correlation analysis function in python to analyze the correlation between pollutants and AQI index, and then established the long-term and short-term memory circulation neural network (LSTM) model based on the deep learning library Keras (a

high-level neural network API). The air quality index (AQI) of Taiyuan city is predicted by simulation. The experimental results show that the root mean square error of the model is 4.875, which has the advantages of high prediction accuracy and wide range. It provides a scientific and reasonable theoretical basis and a new prediction method for the prevention and control of air pollution[1]. In addition, the research of rolling prediction model in the field of air quality index (AQI) prediction is currently in a positive development stage. Ling Jin adopts rolling prediction method. In each rolling prediction cycle, the current meteorological data are collected first, and the prediction model adopted is retrained. Then the prediction result is obtained by error superposition correction method.[2] Sun Jing et al. took an index published by the National Bureau of Statistics as the prediction target, and based on the variable screening mechanism, scrolled through the network keyword library to screen explanatory variables, and established a prediction model through various machine learning methods. [3] The purpose of the rolling prediction in this study is also to explore methods to solve such problems as "the complexity and variability of big data on the Internet affect the stability of the model."[4]Ricardo Navares et al. studied the method to predict the concentration of pollutants and pollen, and described that the LSTM neural network model has different topological structures. The prediction results of the LSTM neural network model are more accurate than those of the traditional model.[12]

2 THEORETICAL BASIS

2.1 Principle of ARIMA Model

The ARIMA model is a commonly used time series analysis method, which is used to forecast and model the time series data. It combines autoregressive (AR) model, difference (I) model and moving average (MA) model.[5] The following describes these three parts and their formulas:
The formula for the ARIMA model can be expressed as:

$$\boldsymbol{Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t} \#(1)$$

In this formula (1): is the time series data we are considering. To are the parameters of the AR model, which are used to describe the relationship between the current value and the past P point-in-time values. To are the parameters of the MA model, which are used to describe the relationship between the current value and the errors at the past Q points in time. Is the error term at point in time t. C is a constant term.$Y_t \varphi_1 \varphi_p \theta_1 \theta_q \varepsilon_t$
The idea of ARIMA model is that the time-varying data series are regarded as random series and the change trend is described by mathematical model. The ability to capture the trend and periodicity of time series data to some extent and make predictions based on past observations to provide estimates of future points in time. By adjusting the model parameters, different types of time series data can be modeled and predicted.[6]

2.2 Principles of LSTM Model

LSTM avoids the long-term dependency problem through deliberate design. It is a variant of RNN, which is specially designed to solve the long-sequence dependency problem.[7] It can model long-term dependencies in serial data. LSTM controls the flow of information through the gating mechanism of input gate, forgetting gate and output gate, and simulates the switching function to regulate the input, retention and output of data. LSTM is a PNN model, which is an improvement on Simple RNN. Its structure diagram is shown in the figure, and its principle is similar to Simple RNN[8], the state vector h is updated whenever an input x is read.

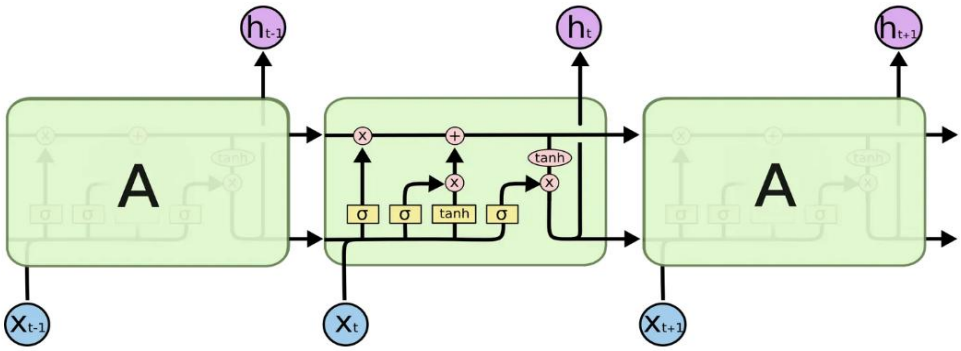


Figure 1 Schematic Diagram of LSTM Network Structure

The calculation process is shown in equations (2) to (8), where F, I and O are forgetting gate, input gate and output gate respectively; C represents short-term memory and h represents long-term memory. Is the activation function; W is the transformation weight matrix (,) from the unit vector to the gate vector; X as the current input; B is the vector feature () obtained by each gate of the input layer. It's a cellular state.$\sigma W_f W_i$、W_o、$W_c b_f$、b_i、b_o、$b_c c_t$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \#(2)$$

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \#(3)$$

$$Sigmoid = 1/e^{-1} \#(4)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \#(5)$$

$$\widetilde{c_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \#(6)$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \#(7)$$

$$h_t = o_t \tanh(c_t) \#(8)$$

2.3 Principle of Rolling Forecast Model

The principle of Rolling Forecasting is to use historical data to continuously update the forecasting model to adapt to the changing data. Specifically, the rolling prediction is to calculate the state transition matrix P based on the historical data in the previous Ts at each time instant, and predict the vehicle speed state (or other amount to be predicted) for the next second based on the state transition matrix P, where T represents the size of the rolling time window.[11]

In rolling forecasts, the model is retrained each time a new piece of data is added and the value at the next point in time is predicted. In this way, the prediction model can be kept updated and closer to the real data changes, thus improving the accuracy of the prediction. The rolling prediction technique is widely used in time series prediction, especially in scenes requiring real-time prediction, such as traffic flow prediction, energy demand prediction etc.[9]

Rolling prediction also involves some specific model framework principles, such as SCINet, which is a hierarchical down sampling-convolution-interaction TSF framework that can effectively model time series with complex time dynamics. Rolling prediction is a dynamic prediction method based on historical data. It adapts to the changes of data by constantly updating the prediction model, so as to improve the accuracy of prediction.[10]

3 DATA PRESENTATION

3.1 Data Sources

The air quality data in this paper covers the period from 31 December 2013 to 10 May 2024. The air quality index (AQI) is an intuitive indicator of air quality, which is a standardized index to measure air quality. AQI is calculated based on the concentrations of six different pollutants (such as PM2.5, PM10, sulfur dioxide, carbon monoxide, ozone and volatile organic compounds) and converted into a comprehensive index value, which reflects the air quality level and the degree of impact on human health. The data characteristics included are shown in Table 1:

Table 1 Data Characteristics and Dimensions

Data	Characteristic	Dimension
Air pollution data	$PM_{2.5}$ PM_{10} NO_2 SO_2 O_3 CO	6

Generally speaking, a higher AQI value indicates a poorer air quality and a greater impact on health. The relevant information about the air quality index is shown in Table 2. AQI is often used to convey air quality information to the public to help people understand the air quality condition of the day and take corresponding protective measures.

Table 2 Classification of Air Quality Index

AQI value	Air quality level	air quality	represent color
0-50	level 1	excellent	green
51-100	level 2	good	yellow
101-150	level 3	Light pollution	orange
151-200	level 4	Moderate pollution	red
201-300	level 5	Severe pollution	purple

>300	level 6	Serious pollution	maroon

3.2 Data Preprocessing

In this paper, interpolation is used to fill in the missing values in the data set, and the average value of the previous value and the next value in the missing value in the data frame is taken to replace the original missing value. In order to make the prediction more accurate, we extract features and target variables. The feature and target variables are normalized and their values are scaled between 0 and 1. We used the Min-Max normalization model and the Z-score normalization model in the scaling process.[13] By comparing the average mean square error after their processing, we finally chose the Min-Max function to perform the normalization process. Its principle is to use the maximum and minimum values in the data to scale accordingly. The processing formula (9) of Min-Max Scaler is as follows:

$$X_{scaled} = \frac{X - X_{min}}{X_{max-X_{min}}} \#(9)$$

Where x is the original number, the minimum value in the original data, the maximum value in the original data, and the scaled data.$X_{min}X_{max}X_{scaled}$

4 EXPERIMENTAL ANALYSIS

4.1 AQI Forecast of ARIMA Model

4.1.1 Model identification

We can evaluate the autocorrelation and partial autocorrelation of the time series data, so as to preliminarily determine the values of the parameters P and Q in the ARIMA model. Then, we try different combinations of P and Q through grid search method, and use information criteria (such as AIC, BIC, etc.) to select the optimal model. In this paper, the ARIMA (4,1,4) model is determined to be the best fit model after grid search and information criterion evaluation.

4.1.2 Model validation

As shown in Figure 2, the residuals of the ARIMA model show characteristics close to normal distribution, but there is still a slight deviation. This usually means that the model has fitted the data relatively well, although further optimization of the model or the use of different types of models may improve the accuracy of the prediction.

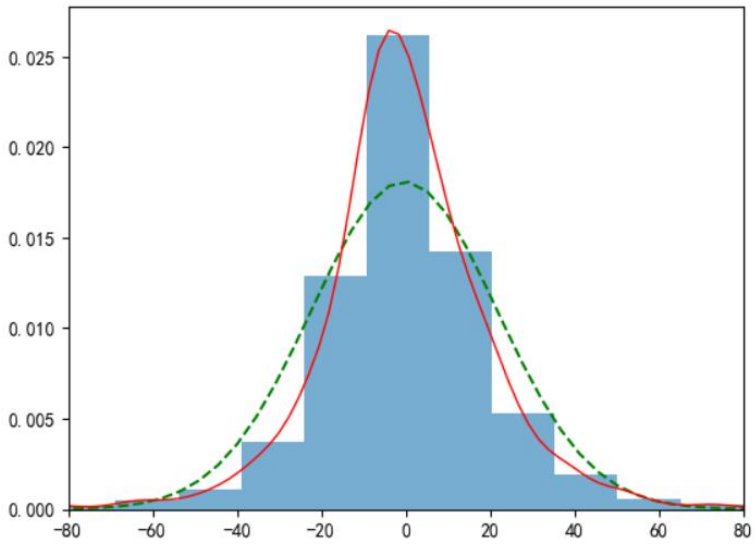


Figure 2 Residual Density Diagram

As can be seen from the QQ plot, the red straight line in the plot represents the expected trend of the theoretical distribution, i.e. if the data completely meets the standard normal distribution, the data points should be roughly distributed along this straight line. The blue dots indicate the correspondence between the actual observations and the theoretical normal quantiles.

Regarding the residual diagnosis of the model, the blue data points are relatively concentrated near the straight line, indicating that the central part of the data is relatively in line with the normal distribution. At both ends of the theoretical quantile (especially at extremes), the data points deviate from the theoretical straight line. In general, the residuals of the ARIMA model are more in line with the normal distribution in the central part, but there is a certain degree of deviation at both ends.

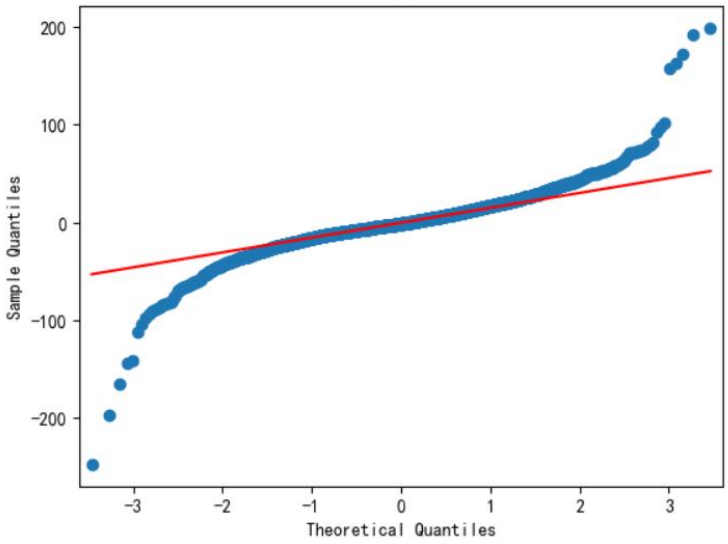


Figure 3 QQ Chart

4.1.3 Model forecast

Comparing the fluctuation between the observed value and the fitted value, as shown in Figure 4, the actual observed curve in blue fluctuates greatly, while the fitted curve in red is relatively smooth. This indicates that the ARIMA model may have a certain smoothing effect when dealing with high-frequency fluctuations, resulting in some extreme values and peaks not being fully captured. The response speed of the model and the red fitting curve delay the response of the blue actual observation curve at some points, especially when the observation value has a big jump or drop. This may indicate that the model has limited adaptability to rapidly changing environmental conditions.

Although there are some local errors and delays in capturing the overall trend, the red fitting curve can generally follow the trend of the blue observation curve. This shows that ARIMA model is effective in capturing the overall trend change of AQI.

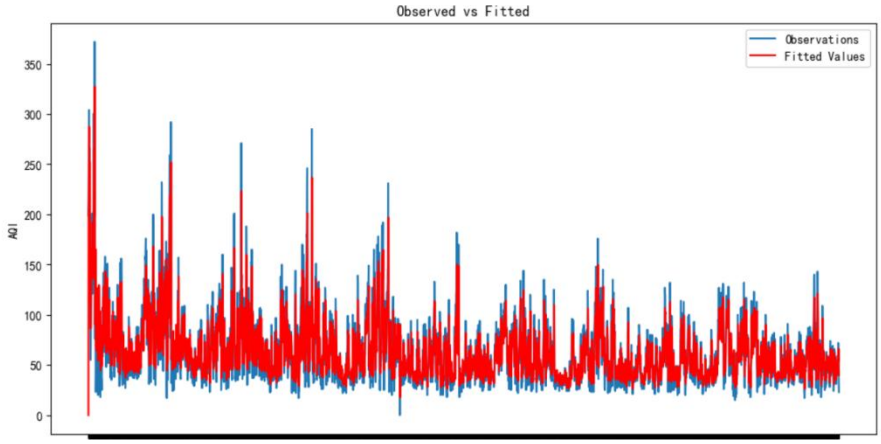


Figure 4 Simulation Forecast Diagram

As shown in Figure 6, it can be concluded from the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of the prediction model that the time series data have no significant correlation when the lag order is high. In this case, the ARIMA model does not need autoregressive (AR) part or moving average (MA) again.

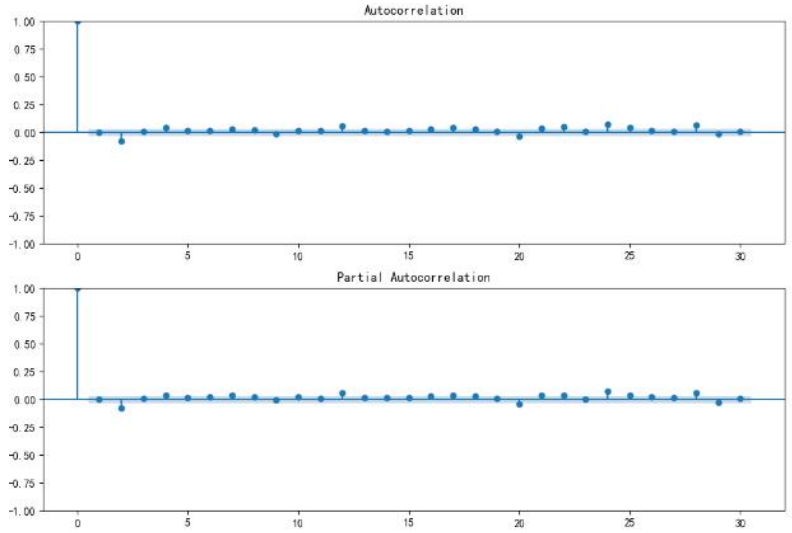


Figure 5 Graph of Autocorrelation Function (ACF) and Partial Autocorrelation Function

Next, the ARIMA (4,1,4) model is used to predict the AQI value of Guilin City for 7 days from May 4, 2024 to May 10, 2024.

As shown in Table 3, comparing the predicted value of AQI with the actual value of AQI, it is found that the absolute error value of the prediction by using ARIMA (4,1,4) model is within (0-10), the relative prediction error is < 25%, and the average relative error is 13.4%. From this, we can draw a conclusion that the prediction accuracy of our model is ideal.

Table 3 ARIMA (4,1,4) Model Predicted Values

date	AQI actual	AQI forecast	error value	Relative Forecast Error/%
2024/5/4	25	30.17502	-5.17502	20.7%
2024/5/5	41	36.33652	4.663481	11.37%
2024/5/6	33	40.10377	-7.103768	21.52%
2024/5/7	52	42.47779	9.522206	18.31%
2024/5/8	52	44.23757	7.762428	14.92%
2024/5/9	49	45.55597	3.444032	7.03%
2024/5/10	46	46.06787	-0.067867	0.15%

4.2 AQI Forecast of LSTM Model

The LSTM neural network model is used to predict the AQI of Guilin City. The data of the first three days are used to predict the air quality of the next day. An LSTM model is constructed, which contains an LSTM layer and a dense connection layer. The mean square error is used as the loss function, and the Adam optimizer is used as the optimizer to fit the model using the training data, which is evaluated on the verification set. Finally, we used Matplotlib to visualize the model's loss values on the training and validation sets in order to evaluate the training of the model. The results are shown in Figure 6 below.

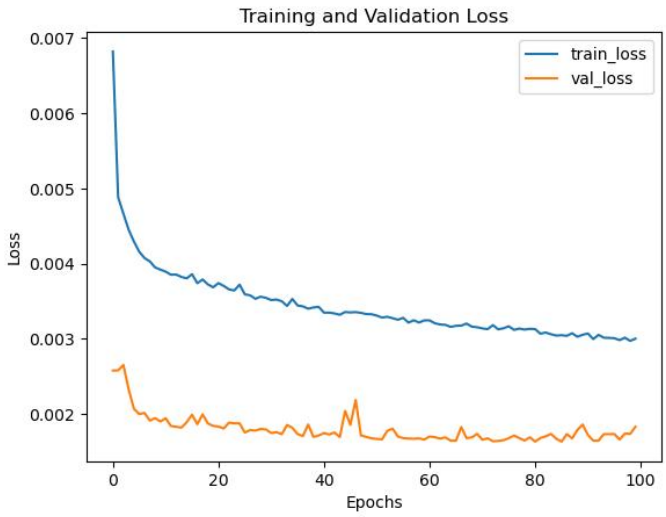


Figure 6 Curve of Loss Change of Model

The graph shows the change of the model's loss values on the training set and the verification set with the training rounds. The abscissa is the training round and the ordinate is the loss value. By comparing the loss curves on the training set and the verification set, it can be evaluated whether the training process of the model is over-fitted or under-fitted. As can be seen from the Figure 7, the training loss (blue line) continues to decline, and the verification loss (orange line) fluctuates slightly and generally shows a downward trend. This indicates that the model is learning from the training data, and there is no obvious over-fitting or under-fitting, indicating that the model can effectively learn the training data.

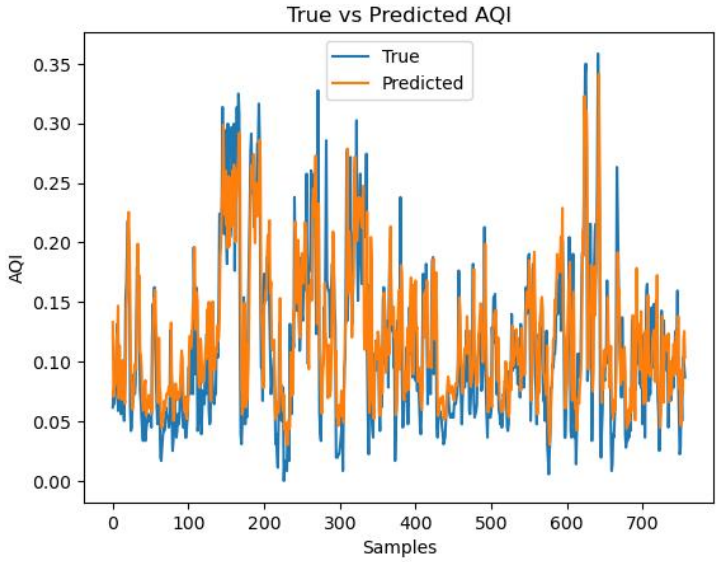


Figure 7 Comparison of Predicted and Actual Values based on LSTM Model

The graph shows the comparison between the predicted results of the model on the test set and the true target value. The abscissa is the index of the sample (i.e. the number of samples) and the ordinate is the AQI value. The graph includes two curves, the blue line segment represents the true AQI value, and the orange line segment represents the AQI value predicted by the model. By observing this graph, the prediction performance of the model can be intuitively evaluated, showing that the predicted value is relatively close to the real value in a number of intervals, especially in the capture of extreme points, which is relatively good, with moderate volatility, and showing good model stability and generalization ability. It shows that our forecast result is accurate. Real-time air quality forecasts can be provided to government departments, environmental protection agencies or the public.

Table 4 Model Evaluation Indicators

evaluating indicator	index value
RMSE	0.03281
MAE	0.02297
MSE	0.00108

As shown in Table 4: RMSE is the standard deviation of the prediction error and represents the average deviation between the predicted value and the actual value. Because the value of RMSE is small, the prediction result of the model is very close to the actual value, and the prediction performance of the model is good. MAE is the average value of the absolute value of the prediction error, which directly reflects the average error between the predicted value and the actual value. MAE is 0.02297, the smaller value of MAE indicates that the prediction error of the model is smaller and the prediction result is more accurate. MSE is the average value of the square of the prediction error and can more sensitively reflect the deviation between the predicted value and the actual value. The value of MSE is 0.00108. The smaller MSE value indicates that the deviation between the predicted result and the actual value of the model is small and the prediction performance is good.

4.3 Rolling Forecast Model Training

The training process of rolling prediction model is usually divided into three stages: first, historical data cleaning and quality verification are carried out to ensure the integrity of time series data; Subsequently, model training is

carried out based on the initial data set, and a basic prediction model is constructed through parameter optimization; Finally, a rolling iteration step is entered, the predicted values generated by the model are backfilled to the training set, the model parameters are dynamically updated with new data, and the prediction error is continuously monitored by using the verification set. In this process, the adaptability of the model to time series changes is gradually improved through the mechanism of data closed loop and model iteration, and the double optimization of prediction accuracy and stability is finally achieved.

4.3.1 ARIMA model forecast

By looking at fig. 9, it can be seen that the predicted value red line roughly follows the trend of the actual value blue. The predicted curve is close to the true value most of the time. However, the volatility of the forecast value is large, especially at some extreme points. This may indicate that the model is overreacting to some specific changes in the data, or that the model has not been able to capture all the factors that affect changes in AQI. Although the overall tracking ability of the model is good, the large fluctuations and deviations indicate that the model still has room for improvement in stability and accuracy.

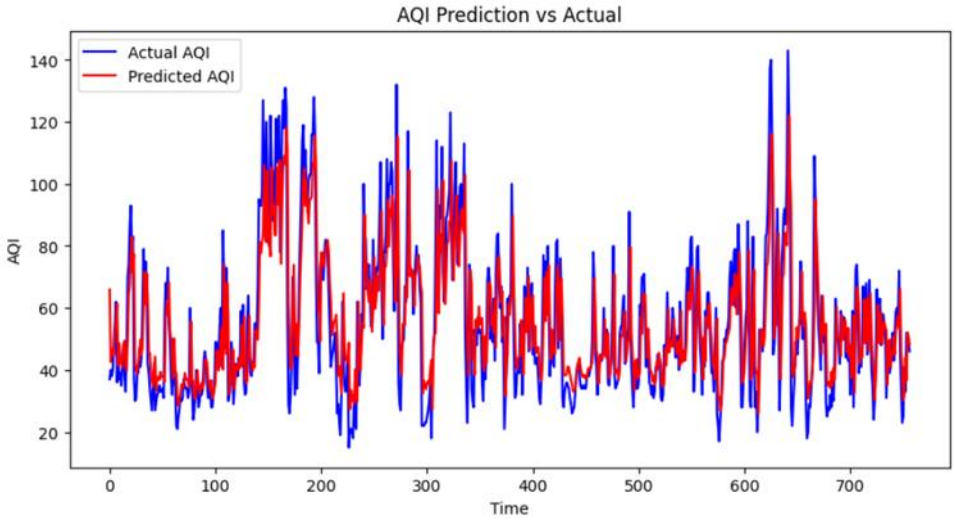


Figure 8 Actual and Predicted AQI Trend of ARIMA Model over Time

4.3.2 LSTM model prediction

By observing fig. 10, it can be seen that the predicted value red line roughly follows the trend of the actual value blue, showing that the model can better capture the change pattern of AQI. This indicates that the prediction model used is effective for trend prediction. This model is very important for environmental management and health early warning system, and can help relevant departments to make timely response and mitigate the impact of adverse air quality.

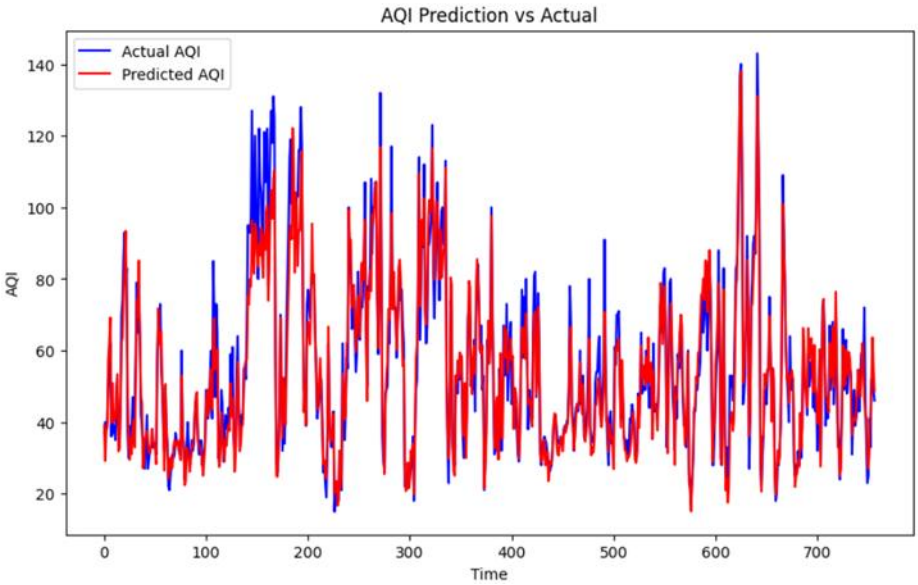


Figure 9 Trend Chart of Actual and Predicted AQI of LSTM Model over Time

4.3.3 Comparative analysis of forecast results

The rolling prediction results under ARIMA model and LSTM model are compared and mainly evaluated by MSE

and RMSE values.

Table 5 Comparison of Model Prediction Errors

Model structure	Test RMSE	Test MSE
ARIMA	14.8031	219.1332
LSTM	7.6096	57.9069

The model is judged according to the mean square error of the model test. The smaller the root mean square error is, the closer the predicted value of the model is to the real value. As can be seen from Table 5, LSTMThe root mean square error of the model is less than that of ARIMA model, indicating that the prediction of AQI index by LSTM model is accurate. Therefore, we choose LSTM model to forecast the future air quality in Guilin.

Considering the above indicators, the rolling forecast AQI under LSTM shows reasonable performance in general, and can effectively predict the results of air quality index of Guilin in the next ten days. As shown in Table 6, it can be concluded that the air quality is in excellent condition, which is very suitable for outdoor activities.

Table 6 Results of Model Forecast AQI

Time	2024.5.11	2024.5.12	2024.5.13	2024.5.14	2024.5.15
Predicted value	40.1942	31.6391	40.7786	46.9876	46.6520
Time	2024.5.15	2024.5.16	2024.5.17	2014.5.18	2024.5.19
Predicted value	45.4955	44.9733	45.0966	45.2633	45.3021

5 CONCLUSION

This paper aims to use ARIMA model, LSTM model and rolling forecast method to forecast the air quality of Guilin city based on big data in order to improve the accuracy and real-time of the forecast. By collecting the air quality data of the mayor's time in Guilin and combining the advantages of the three models, we successfully constructed a model that can accurately predict the air quality change in Guilin. The ARIMA model is used to capture the trend and seasonality of the data, while the LSTM model can better deal with the complex sequence relationship, while the rolling prediction method compares the prediction effects of the two, and updates the real-time parameters to adapt to the changes of the data. Specifically, the ARIMA model plays an important role in predicting the long-term trend and seasonality of the air quality index, while the LSTM model is more suitable for capturing complex dynamic relationships in the data, thus improving the accuracy of the prediction. The rolling prediction model ensures that the model parameters are updated in time to adapt to the changes of new data, thus keeping the real-time prediction.

Through the comprehensive application of these three models, this paper obtains the air quality index prediction results of Guilin with high accuracy and stability. This provides an important reference for air quality management in Guilin and a useful methodology for air quality prediction in other similar cities. In the future, the model parameters and algorithms can be further optimized to improve the prediction accuracy and real-time performance in order to better meet the needs of urban air quality management.

There are still some deficiencies in this paper, among which there may be a certain degree of subjectivity in model selection and parameter setting. The model selection and parameter design should adopt an objective method integrating various factors. In the future, more rigorous model selection and parameter optimization can further improve the prediction performance. On the prediction performance of the model, the mechanism and explanation behind the model may not be explored deeply enough. In the future, the internal mechanism of the model can be further studied in order to better understand the changing laws and influencing factors of air quality.

COMPETING INTERESTS

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

REFERENCES

[1] Fang Xiaoping, Chen Xiuluan, Chu Qi, et al. prediction of air quality index in Yangtze river economic zone based on PSOGSA-LSTM model. Mathematical statistics and management, 2023, 42(01).
[2] Zheng Yangyang, Bai Yanping, Hou Yuchao. Application of LSTM model based on Keras in prediction of air quality index. Mathematical practice and understanding, 2019, 49(07): 138-143.
[3] Ling Jin. Short-term wind speed prediction method based on rolling prediction and its error analysis. Hefei University of Technology, 2017.
[4] Sun Jing, Zhu Jianlin, Li Wanlan, et al. Research on rolling forecast of consumer confidence index based on internet data. Journal of Xi 'an Jiaotong University (Social Science Edition), 2021, 41(06).
[5] Ma Siyuan, Jiao Jiahui, Ren Shengqi, et al. Missing Value Filling of Urban Multiple Air Quality Data Based on Attention Mechanism. Computer Engineering and Science, 2023, 45(08): 1354-1364.
[6] Guo Kuo. Applicability Analysis of ARIMA Model in Precipitation Forecast in Fuxin Area. Heilongjiang Water Conservancy Science and Technology, 2024, 52(04): 37-39+79.
[7] Gu Keming. Research on air quality prediction based on spatio-temporal depth neural network. Zhejiang University of Technology, 2024.
[8] YU Hui Qing, HUANG Peng Cheng, ZHAO Zhen Yu, et al. Standard cell delay prediction method based on RNN. Journal of Zhengzhou University (Science Edition), 2024, 1-7.
[9] Tian Xuwei, Yang Kai, Yin Tong, et al. Prediction of port ambient air quality index based on SSA-Bi-LSTM. Transportation Energy Conservation and Environmental Protection, 2023, 19(05): 32-36+41.
[10] Zhang Tingzhong, Zhang Qinghui, Xing Qiang, et al. Short-term wind speed rolling prediction method based on combination of LMD and GA-BP neural network. Shandong Power Technology, 2019, 46(11): 13-20.
[11] Atakan J, Ayse B O. Forecasting air pollutant indicator levels with geographic model 3 days in advance using neural networks. Expert Systems with Applications, 2010, 37(12): 7986-7992.
[12] A Monterio, M Vieira, C Gama, et al. Miranda. 2017. Towards an improve air quality index. Air Quality, Atmosphere And Health, 10(4): 447-445.
[13] Navares R, Aznarte J L. Predicting air quality with deep learning LSTM: Towards comprehensive models. Ecological Informatics, 2019, 55: 101-119.