Volume 2, Issue 1, 2025

Print ISSN: 3007-6870
Online ISSN: 3007-6889

# Journal of Trends in Applied Science and Advanced Technologies

# Journal of Trends in Applied Science and Advanced Technologies

Volume 2, Issue 1, 2025


Published by Upubscience Publisher

**Journal of Trends in Applied Science and Advanced Technologies**
**Print ISSN: 3007-6870 Online ISSN: 3007-6889**
**Email: info@upubscience.com**
**Website: http://www.upubscience.com/**

# Table of Content

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3007**

# THE ROLE OF IT AUDITING IN DATA SECURITY FOCUSING ON RISK IDENTIFICATION, STRENGTHENING INTERNAL CONTROLS, AND COMPLIANCE WITH SECURITY POLICIES

Indy Misya Rumata Situmorang, Kania Jasmine Azzahra, Iskandar Muda*
*Universitas Sumatera Utara, Medan, Indonesia.*
*Corresponding author: Iskandar Muda, Email: ismuda.jurnal.internasional@gmail.com*

**Abstract:** In an increasingly digital era, data security has become one of the top priorities for organizations. Information Technology (IT) auditing plays a crucial role in ensuring that the information systems and data managed by companies are protected from potentially harmful threats. This paper aims to explore the role of IT auditing in enhancing data security, focusing on risk identification, strengthening internal controls, and compliance with security policies. Through a qualitative approach, this research collects data from interviews with audit professionals and case studies from several organizations. The results indicate that IT auditing not only helps identify weaknesses in security systems but also provides recommendations that can enhance data protection effectiveness. Additionally, IT auditing contributes to raising awareness of the importance of data security throughout the organization. The conclusion emphasizes that the routine implementation of IT auditing is a strategic step necessary to maintain the integrity and confidentiality of information in an increasingly complex business environment. This abstract provides a brief summary of the objectives, methodology, results, and conclusions of the research regarding the role of IT auditing in data security.
**Keywords:** IT audit; Data security; Risk management; Internal controls; Compliance; Vulnerabilities; Information protection; Information systems; Data breach

## 1 INTRODUCTION

In an increasingly advanced digital era, organizations face various challenges related to data security. With the growing volume of data generated and stored, as well as the complexity of the information systems used, the risks of data breaches, cyberattacks, and information misuse are becoming increasingly significant [1]. Data security is not only essential for protecting sensitive information but also for maintaining the reputation and trust of customers. In this context, Information Technology (IT) auditing emerges as a crucial tool for assessing and enhancing data security within organizations. IT auditing serves to evaluate the effectiveness of existing internal controls, identify potential risks, and ensure that security policies and procedures are properly followed [2]. Through a systematic audit process, organizations can gain better insights into their security system's strengths and weaknesses. However, despite the recognized importance of IT auditing, many organizations have yet to implement audit practices routinely or effectively. This research aims to:
1. Identify the Role of IT Auditing: Explain how IT auditing contributes to enhancing data security within organizations.
2. Analyze the Audit Process: Examine the steps taken in IT auditing to identify and address data security risks.
3. Provide Recommendations: Develop recommendations for best practices in the implementation of IT auditing to improve data protection.
4. Increase Awareness: Enhance understanding of the importance of IT auditing in the context of information security among professionals and stakeholders.

## 2 LITERATURE REVIEW

### 2.1 Definition of IT Audit

An Information Technology (IT) audit is the process of evaluating an organization's information technology systems, infrastructure, and policies. The purpose of this audit is to ensure that all aspects of IT operate in accordance with established security and efficiency standards [2]. IT audits encompass reviews of hardware, software, as well as procedures and policies related to data management and information systems within the organization. By conducting an IT audit, organizations can identify potential risks, weaknesses in internal controls, and ensure compliance with applicable regulations.

### 2.2 Objectives of IT Audit

The primary objectives of an IT audit include [3]
1. Risk Identification: Identifying potential security risks that may threaten data and information systems.
2. Performance Assessment: Evaluating the performance of IT systems to ensure they function effectively and efficiently.

3. Regulatory Compliance: Ensuring that all IT policies and procedures comply with relevant regulations and standards.
4. Improvement Recommendations: Providing recommendations for improvements based on audit findings to enhance the security and efficiency of systems.

### 2.3 Commonly Used Methodology

The methodology for IT auditing generally follows these steps: [3]
1. Audit Planning: This step involves determining the scope, objectives, and approach of the audit. At this stage, auditors need to understand the information technology assets present in the organization.
2. Data Collection: Information regarding the IT systems is gathered through interviews, document reviews, and direct observations to gain an in-depth understanding of the technologies used.
3. Control Evaluation: This involves testing the effectiveness of security controls and procedures implemented within the IT systems. It includes analyzing security mechanisms and assessing potential weaknesses.
4. Findings Analysis: Analyzing findings related to control weaknesses and evaluating the risks faced by the organization based on the evaluation results.
5. Audit Reporting: Compiling a report that summarizes the evaluation results and improvement recommendations. The report should be clear and easily understood, serving as a strategic guide for the organization.
6. Follow-Up: Ensuring that recommendations are effectively implemented and monitoring to assess the effectiveness of changes made.

### 2.4 Data Security: Explanation, Threats, and Importance of Data Protection

#### *2.4.1 Explanation of data security*

Data security refers to practices designed to protect digital data and information from unauthorized access, misuse, or theft. It encompasses a series of steps and technologies aimed at maintaining the confidentiality, integrity, and availability of data throughout its lifecycle [4]. Data security involves understanding the types of data held, their storage locations, and the risks that threaten that data. With the increasing reliance on technology and the internet, data security has become crucial for protecting sensitive information from existing threats.

#### *2.4.2 Existing threats*

Threats to data security can arise from various sources, including:
1. Cyber Attacks: Such as malware, ransomware, phishing, and Distributed Denial of Service (DDoS) attacks that can damage or steal data.
2. Unauthorized Access: Unauthorized users may attempt to access sensitive data illegally.
3. Data Breaches: Occur when sensitive information leaks to third parties without permission, often due to human error or system failures.
4. Physical Damage: Loss of data due to natural disasters or hardware failures also poses a serious threat to data security.

These threats can lead to significant financial losses for organizations and damage their reputation in the eyes of customers and business partners.

#### *2.4.3Importance of data protection*

Data protection is crucial for several reasons: [2]
- Protecting Individual Privacy: Data security helps safeguard personal information from falling into the wrong hands, thereby protecting individual privacy.
- Maintaining Business Integrity: Organizations that can effectively protect their data are more likely to gain the trust of customers and business partners, which in turn enhances the company's reputation.
- Preventing Financial Loss: Data breaches or losses can result in significant financial repercussions due to legal penalties, loss of customers, and recovery costs.
- Compliance with Regulations: Many countries have laws and regulations that require organizations to protect personal data. Failure to comply with these regulations can lead to legal sanctions.

### 2.5 The Relationship Between IT Audit and Data Security

Information Technology (IT) auditing plays a crucial role in enhancing data security within organizations. In this context, IT audits serve not only as tools to assess compliance with policies and procedures but also as mechanisms to identify and mitigate potential risks that could threaten the integrity and confidentiality of data [5]. The following points illustrate the relationship between IT auditing and improved data security:
1. Evaluation of Security Weaknesses: IT audits conduct in-depth assessments of IT systems to identify vulnerabilities in data security. As noted in literature, IT audit services can assist organizations in detecting vulnerabilities to cyberattacks, such as hacking and malware, which can compromise sensitive data. By performing these evaluations, auditors can provide recommendations to strengthen security infrastructure.
2. Enhancement of Security Infrastructure: After identifying weaknesses, IT audits play a role in providing recommendations for improving security infrastructure. This includes enhancements to firewalls, data encryption, and

critical software updates. By implementing these measures, organizations can reduce the risks of data breaches and cyberattacks.

3. Regulatory Compliance: IT audits also ensure that organizations comply with applicable data protection regulations, such as GDPR or Indonesia's Personal Data Protection Law. Adhering to these regulations not only helps avoid significant fines but also enhances customer trust in the organization. Regular audits can assist organizations in maintaining this compliance.
4. Monitoring and Early Detection: IT audits aid in establishing effective monitoring systems to detect suspicious activities or security threats early on. With early detection, preventive actions can be taken before an attack occurs, thereby better safeguarding data security.
5. Preventing Data Breaches: Through regular audits, organizations can prevent data breaches before they happen. The audit process helps identify security gaps that could be exploited by cybercriminals and provides improvement recommendations to mitigate the risk of reputational damage or financial loss due to security incidents.
6. Increasing Security Awareness: IT audits also serve to raise awareness about the importance of information security throughout the organization. By involving all employees in the audit process and providing training on security policies, organizations can foster a stronger culture of security.

Overall, the relationship between IT auditing and data security is very close. Through systematic evaluations and improvement recommendations provided by IT auditors, organizations can enhance their data protection levels and minimize risks associated with cyber threats [6]. Thus, IT auditing becomes an integral part of risk management strategies and data protection in today's digital era.

## 3 METHODS

### 3.1 Research Design

This research employs a qualitative approach to explore the role of IT auditing in enhancing data security within organizations. The qualitative approach is chosen because it allows the researcher to gain an in-depth understanding of the phenomena being studied, as well as the context and dynamics influencing the implementation of IT audits and data security.

Reasons for Choosing a Qualitative Approach

1. Depth of Information: The qualitative approach enables researchers to delve deeply into information through interviews and focus group discussions. This provides richer insights into the experiences and perspectives of IT audit and data security professionals.
2. Flexibility: Qualitative methods offer flexibility in data collection, allowing researchers to adjust questions and research focus based on participant responses.
3. Concepts and Perceptions: This research aims to understand the concepts and perceptions held by practitioners regarding the relationship between IT auditing and data security, which can be better explored through a qualitative approach.

### 3.2 Data Collection Methods

1. In-Depth Interviews: The researcher will conduct interviews with professionals involved in IT auditing and data security across various organizations. These interviews will be semi-structured, allowing the researcher to have a set of questions while also being open to further discussion based on the respondents' answers.
2. Case Studies: This research will also include case studies from several organizations that have effectively implemented IT audits. These case studies will provide real-world examples of how IT auditing contributes to enhancing data security.
3. Document Analysis: In addition to interviews and case studies, the researcher will analyze relevant documents, such as previous audit reports, data security policies, and internal procedures of the organizations. This will help provide additional context and support the findings from the interviews.

### 3.3 Data Analysis

The data collected through interviews and document analysis will be analyzed using thematic analysis techniques. This process involves identifying patterns, themes, and categories that emerge from the data to address the research questions. The results of this analysis will be used to formulate key findings regarding the role of IT auditing in enhancing data security.With this qualitative approach, the research aims to provide a deeper understanding of how IT auditing can contribute to data protection within organizations, as well as the challenges and opportunities encountered during its implementation.

## 4 RESULTS AND DISCUSSION

### 4.1 Results

This research identifies several key findings related to the role of Information Technology (IT) auditing in enhancing

data security within organizations. Based on data analysis obtained from interviews, case studies, and relevant literature, here is a summary of the main findings:

1. Identification of Security Weaknesses: IT audits effectively identify weaknesses in security systems that organizations may not be aware of. For instance, an audit conducted at Institution X using the ISO/IEC 27002 standard revealed vulnerabilities in access management and data protection that could be exploited by unauthorized parties. By identifying these weaknesses, organizations can take necessary corrective actions.
2. Enhancement of Security Infrastructure: After weaknesses are identified, IT audits provide recommendations for improving security infrastructure. These recommendations may include enhancements to firewalls, implementation of data encryption, and critical software updates [7]. The research found that organizations implementing audit recommendations experienced significant improvements in their data security levels.
**3.** Early Detection of Security Threats: IT audits assist organizations in establishing effective monitoring systems to detect suspicious activities or security threats early on. With early detection, preventive actions can be taken before attacks occur. This was evidenced in a case study at PT Paramita Surya Makmur Plastika, where the implementation of monitoring systems resulting from audits successfully prevented several security incidents.
4. Compliance with Regulations: IT audits also ensure that organizations comply with applicable data protection regulations, such as GDPR and Indonesia's Personal Data Protection Law. The research indicates that companies conducting regular audits are better positioned to meet regulatory requirements and avoid legal sanctions.
5. Increased Security Awareness: IT audits contribute to raising awareness about the importance of data security throughout the organization. Through training and employee involvement in the audit process, organizations can create a stronger security culture.

### 4.2 Analysis of the Role of IT Audit

IT auditing plays a crucial role in helping organizations identify risks, enhance controls, and ensure compliance with security policies [8]. The following is a further analysis of these roles:

- Risk Identification: During the audit process, auditors evaluate the security systems and risk management practices implemented by the organization. By identifying potential risks such as software vulnerabilities or weak access policies, auditors provide valuable information to management for taking corrective actions. This proactive identification helps organizations mitigate risks before they can be exploited.
- Enhancing Controls: IT audits strengthen internal controls by providing recommendations for best practices in data and information system management. For instance, implementing role-based access control (RBAC) and data encryption can significantly enhance the protection of sensitive information. The recommendations from audits help organizations establish a more robust security framework.
- Ensuring Compliance: By ensuring that security policies and procedures are adhered to, IT audits assist organizations in meeting regulatory requirements and industry standards. This not only protects organizations from legal sanctions but also enhances stakeholder confidence in the organization's ability to manage data effectively. Regular audits help maintain compliance with regulations such as GDPR and HIPAA.

### 4.3 Case Study

As a real world example of the implementation of IT auditing, this research includes a case study at Institution X, which conducted a security audit using the ISO/IEC 27002 and COBIT 5 standards. The audit results indicated that the maturity level of IT at the institution was at level 2 (Managed Process), meaning that the processes for implementing information technology had been carried out in a more organized manner but still required improvements.After implementing the recommendations from the audit, which included enhancements to access controls and updates to security procedures, Institution X reported a significant decrease in data breach incidents and an increase in user satisfaction with their services. This case study underscores that the routine implementation of IT auditing not only enhances data security but also provides operational benefits for organizations.

## 5 CONCLUSION

This research highlights the critical role of Information Technology (IT) auditing in enhancing data security within organizations. Through a systematic approach, IT audits enable organizations to identify vulnerabilities in their security systems, strengthen internal controls, and ensure compliance with relevant regulations. The findings indicate that effective IT auditing leads to several key outcomes:

1. Identification of Security Weaknesses: IT audits provide organizations with insights into potential vulnerabilities that may go unnoticed, allowing for timely corrective actions.
2. Enhancement of Security Infrastructure: By implementing recommendations from audits, organizations can significantly improve their security measures, including access controls and data protection protocols.
3. Early Detection of Threats: IT audits facilitate the establishment of effective monitoring systems that help detect suspicious activities early, enabling proactive responses to potential security incidents.
4. Regulatory Compliance: Regular audits ensure that organizations adhere to data protection regulations such as GDPR and the Personal Data Protection Law in Indonesia, thereby avoiding legal penalties and fostering stakeholder

trust.

5. Increased Awareness of Data Security: Engaging employees in the audit process and providing training on security policies contribute to a stronger culture of data security within organizations.

By adopting these practices, organizations can fortify their data security measures and protect their information assets against various digital threats. Ultimately, integrating IT auditing into organizational strategies is essential for effective risk management and ensuring the long-term security of sensitive data.

## CONFLICT OF INTEREST

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

## REFERENCES

[1] Shulha, O, Yanenkova, I, Kuzub, M, et al. Modeling Regarding Detection of Cyber Threats Features In Banks Activities. Journal of Management Information & Decision Sciences, 2022, 25(25): 1-8.

[2] Demirkan, S, Demirkan, I, McKee, A. Blockchain technology in the future of business cyber security and accounting. Journal of Management Analytics, 2020, 7(2): 189-208.

[3] Lois, P, Drogalas, G, Karagiorgos, A, Tsikalakis, K. Internal audits in the digital era: opportunities risks and challenges. EuroMed Journal of Business, 2020, 15(2): 205-217.

[4] AlGhamdi, S, Win, K T, Vlahu-Gjorgievska, E. Information security governance challenges and critical success factors: Systematic review. Computers & security, 2020, 99, 102030.

[5] Yang, P, Xiong, N, Ren, J. Data security and privacy protection for cloud storage: A survey. Ieee Access, 2020, 8, 131723-131740.

[6] Wylde, V, Rawindaran, N, Lawrence, J, et al. Cybersecurity, data privacy and blockchain: A review. SN computer science, 2022, 3(2): 127.

[7] Bandari, V. Enterprise data security measures: a comparative review of effectiveness and risks across different industries and organization types. International Journal of Business Intelligence and Big Data Analytics, 2023, 6(1): 1-11.

[8] Duggineni, S. Impact of controls on data integrity and information systems. Science and Technology, 2023, 13(2): 29-35.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3009**

# QUANTUM TELEPORTATION: THE CURRENT STATE OF RESEARCH

Alamgir Khan*, Jamal Shah
*Department of physics, Abdul Wali Khan University, Mardan, 23200 KPK, Pakistan.*
*Corresponding Author: Alamgir Khan, Email: alamgirkhan03414946231@gmail.com*

**Abstract:** Through the process of quantum teleportation, a transfer of quantum information takes place from one particle to another while maintaining zero physical movement between the two. Quantum teleportation requires two entangled particles along with projective measurements followed by exchanging two bits of classical information. The process requires quantum mechanics principles which combine superposition with entanglement. Quantum entanglement occurs when two or more particles create an inseparable relationship that cuts off a separate description of an individual particle. Experimental realizations of quantum teleportation continue to operate on different physical platforms which include photons as well as atoms and superconducting circuits. Applications of Quantum teleportation includes quantum computing, secure communication, and cryptography. In future Scientists believe that a breakthrough in quantum computing technology could make teleportation a reality, including the ability to teleport a whole human. This paper provides the present research in quantum teleportation, including the theoretical framework, experimental implementations, and potential applications. We also discuss the challenges and limitations of quantum teleportation and propose future directions for research.
**Keywords:** Quantum teleportation; Quantum entanglement; Quantum decoherence; Quantum technology; Quantum computing

## 1 INTRODUCTION

Bennett et al. proposed quantum teleportation in 1993 as a method for transferring quantum information between particles without physical transport [1, 2]. Teleportation is dreamed of as the ability to travel by merely reappearing at a distant place. In classical physics, an object eligible for teleportation can be completely described by its properties, which can be ascertained through measurement [3, 4]. In quantum communication, the space restriction of openly transmitting quantum states can be overcome by quantum teleportation, as can the difficulty of achieving long distance exchanges between qubits in quantum computation [5, 6]. The main protocols in quantum information is quantum teleportation [7, 8]. Quantum teleportation, which leverages the physical resource of entanglement, acts as a fundamental component in numerous quantum information tasks and is a crucial element for quantum technologies. It plays an essential role in the advancement of quantum communication, quantum computing, and quantum networks [9, 10]. One of the essential elements of the nascent domains of quantum communication and quantum computation is the potential to transfer quantum information [11, 12]. There has been rapid advancement in the theoretical understanding of quantum information processing, but due to the challenges associated with managing quantum systems, progress in experimentally implementing new proposals has not kept pace [13, 14]. Since that time, there have been considerable advancements in both the theoretical and experimental aspects of quantum teleportation development [15, 16]. This article offers a summary of the existing research on quantum teleportation.

## 2 QUANTUM SUPERPOSITION

The key alteration among qubits and bits lies in the point that a qubit can exist in a linear arrangement (superposition) of the states $|0\rangle$ and $|1\rangle$ [17, 18]. consider $\propto$ and $\beta$ represent the probability amplitudes of an electron in the ground state $|0\rangle$ and the excited state $|1\rangle$, respectively. The superposition of these states can be expressed [19, 20] of states are

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1)$$

Here, $\alpha$ and $\beta$ are complex numbers, and due to the normalization condition, they satisfy,

$$|\alpha|^2 + |\beta|^2 = 1 \quad (2)$$

In this context, $|\alpha|^2$ represents the probability of finding the state $|\psi\rangle$ in $|0\rangle$, while $|\beta|^2$ represents the probability of finding $|\psi\rangle$ in $|1\rangle$ [21, 22]. Consequently, when a qubit is measured, it collapses to either '0' or '1' probabilistically, based on these probabilities [23, 24]. Let's examine an example of how a qubit is represented,

$$|\Psi\rangle = \left(\frac{1}{\sqrt{2}}\right)|0\rangle + \left(\frac{1}{\sqrt{2}}\right)|1\rangle$$
$$\therefore \propto = \frac{1}{\sqrt{2}} \; and \; \beta = \frac{1}{\sqrt{2}} \tag{3}$$
$$|\propto|^2 = |\beta|^2 = 1/2$$

This implies that there is a 50% probability of the qubit being measured in the $|0\rangle$ state and an equal 50% probability of it being measured in the $|1\rangle$ state. The superposed states are often referred to as *state vectors* or *state spaces*, while $|0\rangle$ and $|1\rangle$ are known as the *basis states*. [25, 26] (See Figure 1).

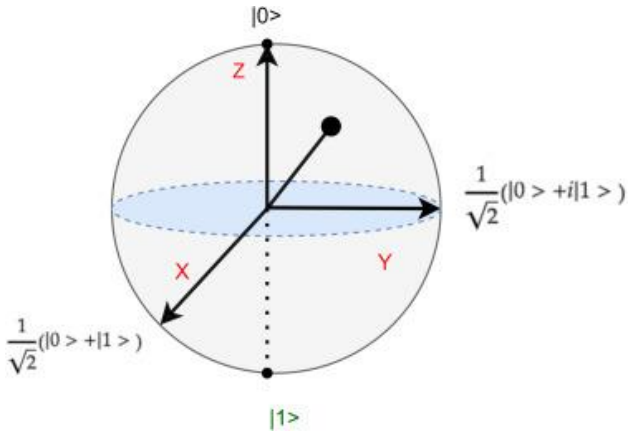



**Figure 1** Graphical Representation of Qubit

## 3 QUANTUM ENTANGLEMENT

The mysterious phenomenon in the existing universe is Quantum entanglement where particles become intricately linked, forming a unified system that remains interconnected regardless of the distance separating them [27, 28]. When particles are entangled, their properties are so deeply correlated that measuring one rapidly effects the state of the other, even if they are light-years apart [29]. Quantum information science relies on this special state of interdependence because it makes possible revolutionary technologies including quantum communication and teleportation as well as quantum cryptography [29]. Maximum entangled states act as essential elements for establishing quantum communication channels between distant users. The protection of entangled states proves difficult when securing long-distance links because environmental noises degrade their quality and negatively impact information security [30]. The implementation of quantum repeaters enables the division of extended transmission lines into smaller segments by implementing entanglement purification and swapping to reduce noise and allow entanglement span longer distances [31, 32]. The process of entanglement purification obtains purified high-fidelity entangled states from mixed quantum systems and the process of entanglement swapping creates connections between entangled pairs which forms bridges for quantum nodes that are positioned far apart [33, 34]. Entangled photons follow the same principles as electron entangled particles by using quantum properties such as polarization and angular momentum. The spin states of electrons find their equivalent in circular polarization states of photons which make them suitable for creating entangled pairs [35]. The angular momentum of photons contains both polarization states and spatial dispersion characteristics that present a complicated visualization even though they fundamentally affect photon behavior [36]. Atom decay emissions based on their angular momentum cause photons to emit polarization states of either RHC or LHC according to atomic spin alignment [37, 38]. Similar to other quantum processes photon absorption and scattering heavily depend on angular momentum conservation principles as a demonstration of quantum state-physical manifestation linkage [39, 40]. Quantum mechanics receives deeper comprehension through these principles which simultaneously enables the development of secure communication technology and information processing systems [41, 42]. For two particles, A and B, with wave functions $\psi A$ and $\psi B$, the entangled wave function can be written as:

$$\psi AB = \alpha\psi A\psi B + \beta\psi A\psi B \tag{4}$$

$$|\alpha|^2 + |\beta|^2 = 1 \tag{5}$$

Entanglement swapping is a process that enables the transfer of entanglement from one particle to another [43, 44] (See Figure 2). The mathematical equation for entanglement swapping can be written as:

$$|\psi>AB = \alpha|00+ \beta|11> \quad (6)$$

$$|\psi>CD = \gamma|00+ \delta|11> \quad (7)$$

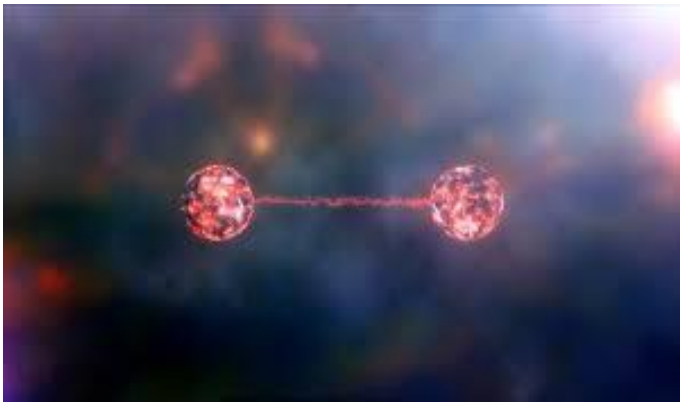


**Figure 2** Representation of Two Entanglement States of Particles

## 4 QUANTUM TELEPORTATION

Alice and Bob share an entangled pair of qubits (**A** and **B**) in the Bell state:

$$|\Phi+\rangle AB = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)AB \quad (7)$$

This means qubits **A** (with Alice) and **B** (with Bob) are maximally entangled. Alice has a third qubit (**C**) in an unknown state:

$$|\psi\rangle C \quad = \alpha|0\rangle C \quad + \beta|1\rangle C \quad (8)$$

Here, $\alpha$ and $\beta$ are complex numbers which satisfy the normalization condition $|\alpha|^2 + |\beta|^2 = 1$.
The combined state of the three qubits (**A**, **B**, and **C**) is

$$|\psi\rangle C \quad \otimes |\Phi+\rangle AB = (\alpha|0\rangle C \quad + \beta|1\rangle C \quad) \otimes \frac{1}{\sqrt{2}}(|00\rangle+|11\rangle)AB \quad (9)$$

Expanding this, we get:

$$|\psi\rangle CAB = \frac{1}{\sqrt{2}}(\alpha|0\rangle C \quad + \beta|1\rangle C \quad)(|00\rangle + |11\rangle)AB \quad (10)$$

Alice performs a Bell state measurement on qubits C and A. To do this, we rewrite the state $|\psi\rangle CAB$ in terms of the Bell basis for qubits C and A. The four Bell states are,

$$|\Phi+\rangle = \frac{1}{\sqrt{2}} \quad (|00\rangle + |11\rangle) \quad (11)$$

$$|\Phi-\rangle = \frac{1}{\sqrt{2}} \quad (|00\rangle - |11\rangle) \quad (12)$$

$$|\Psi+\rangle = \frac{1}{\sqrt{2}} \quad (|01\rangle + |10\rangle) \quad (13)$$

$$|\Psi-\rangle = \frac{1}{\sqrt{2}} \quad (|01\rangle - |10\rangle) \quad (14)$$

Rewriting $|\psi\rangle CAB$ in terms of these Bell states

$$|\psi\rangle CAB \quad = \frac{1}{\sqrt{2}} \quad [|\Phi+\rangle CA \quad (\alpha|0\rangle B \quad + \beta|1\rangle B \quad) +$$

$$|\Phi -\rangle CA \quad (\alpha|0\rangle B \quad -\beta|1\rangle B \quad ) + |\Psi +\rangle CA \quad (\alpha|1\rangle B \tag{15}$$

$$+\beta|0\rangle B \quad ) + |\Psi -\rangle CA \quad (\alpha|1\rangle B \quad -\beta|0\rangle B \quad )]$$

When Alice measures qubits C and A, she collapses the system into one of the four possible Bell states. As a result, the state of Bob's qubit (B) is determined by the specific outcome of Alice's measurement, corresponding to one of the four entangled states.

1. If Alice measures ∣Φ+⟩ CA ∣Φ+⟩ *CA*

$$|\psi\rangle B \quad = \alpha|0\rangle B + \beta|1\rangle B \tag{16}$$

Bob's qubit is already in the correct state.

2. If Alice measures ∣Φ−⟩ CA ∣Φ−⟩ *CA*

$$|\psi\rangle B = \alpha|0\rangle B \quad - \beta|1\rangle B \tag{17}$$

Bob needs to apply a Pauli-Z gate (ZZ) to recover the original state.

3. If Alice measures ∣Ψ+⟩ CA ∣Ψ+⟩ *CA* ,

$$|\psi\rangle B = \alpha|1\rangle B + \beta|0\rangle B \tag{18}$$

Bob needs to apply a Pauli-X gate (XX) to recover the original state.

4. If Alice measures ∣Ψ−⟩ CA ∣Ψ−⟩ *CA*

$$|\psi\rangle B = \alpha|1\rangle B - \beta|0\rangle B \tag{19}$$

Bob needs to apply a Pauli-X gate followed by a Pauli-Z gate (XZ) to recover the original state. Alice sends the result of her Bell state measurement (2 classical bits) to Bob. Based on this information, Bob applies the appropriate quantum gate to his qubit (B) to reconstruct the original state $|\psi\rangle$. After applying the correct operation, Bob's qubit (B) is in the state,

$$|\psi\rangle B \quad = \alpha|0\rangle B \quad + \beta|1\rangle B \tag{20}$$

This is the original state of qubit C, successfully teleported from Alice to Bob (See Figure 3).

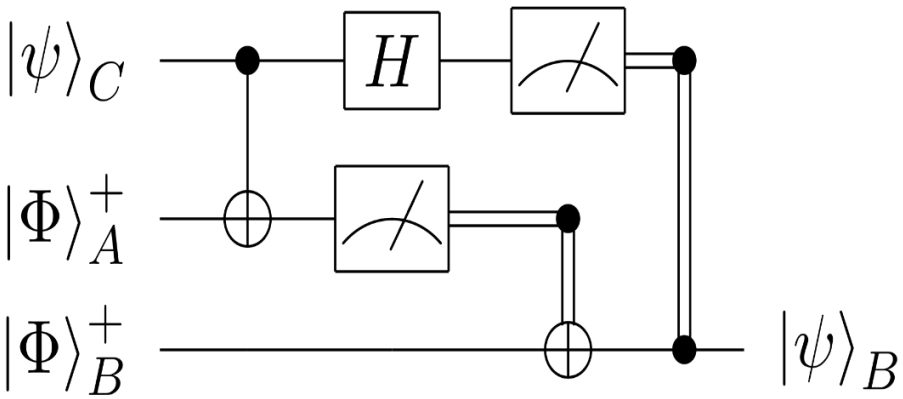



**Figure 3** Circuit Diagram of Quantum Teleportation

The quantum circuit for teleporting a quantum state, as described, begins with a Bell state and the qubit to be teleported as inputs [45]. The circuit involves a series of operations: a CNOT gate, a Hadamard gate, and measurements on two qubits [46]. Following the measurements, two classically controlled gates are applied—a Pauli X gate and a Pauli Z gate—which are executed based on the measurement outcomes [47]. Specifically, if the measurement result corresponds to a particular value, the respective Pauli gate is activated. Once the circuit completes its execution, the quantum state originally held by the input qubit is successfully teleported to the target qubit [48]. After measurement results are obtained the target qubit ends in either its initial state or an altered quantum state. The circuit serves an essential role for quantum teleportation and also enables the swapping of entangled states [49]. Quantum circuits enable the transmission of entanglement from the input qubit that belongs to an entangled pair to another qubit according to the description [50].

## 5 CHALLENGES

Several obstacles stand in the way of practical application of quantum teleportation as an innovative concept [51, 52]. The teleportation procedure accumulates errors because of noise and decoherence effects occurring within quantum systems. Small-scale operations represent the current method used to achieve quantum teleportation [53, 54]. Control over quantum systems and low error rates present the main technical barrier to scale up their operation capabilities. The extended distribution of entanglement particles poses difficulties since it needs the preserved connection between entangled states [47, 55]. Quantum teleportation depends on maintaining high fidelity conditions for the state to stay stable. Achieving exact measurement of quantum systems and maintaining process control becomes a complex challenge during telecommunications [56, 57]. Quantum teleportation needs parties to exchange genuine classical data between each other. Error correction methods which need developmental work must be implemented because of teleportation errors which occur.

## 6 CONCLUSION

Research and physicists have remained fascinated by quantum teleportation ever since decades passed. This research evaluated every relevant study about quantum teleportation both theoretically and practically to investigate foundational concepts and experimental challenges and possible applications of this phenomenon. The study demonstrates that quantum teleportation serves as a complex operational technique that needs total control of quantum states built into particles. Science proves quantum teleportation functions both theoretically and experimentally yet the process remains unstable due to decoherence-caused sensitivity to errors. The investigation points to beneficial features of quantum teleportation despite the multiple challenges that exist. Quantum communication development requires permanent maintenance of mobile quantum particles which need to stay in different locations for information exchange. The research into quantum teleportation enabled scientists to learn better how essential quantum mechanics ideas of entanglement and superposition work. The research conducted for this study generated vital improvements that propel the development of quantum technology particularly for quantum computing as well as quantum cryptography and quantum communication systems. The review process calls for additional experimental work and theoretical validation research in order to understand quantum teleportation properly.

## 7 FUTURE RESEARCH

Scientists need to invest greater effort developing experimental methods that strengthen quantum teleportation fidelity and decrease its performance errors. The research of quantum teleportation through both fresh superconducting qubit and topological quantum systems offers new insights into theoretical and practical applications. Scientists need to conduct more research to build sophisticated theoretical models that accurately follow quantum system conduct during teleportation procedures.

## CONFLICT OF INTEREST

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

## REFERENCES

[1] Pirandola, S, Eisert, J, Weedbrook, C, et al. Advances in quantum teleportation. Nature Photon, 2015, 9(10): 641-652. DOI: https://doi.org/10.1038/nphoton.2015.154.
[2] Bouwmeester, D, Pan, Jian-Wei, Mattle, Klaus, et al. Experimental quantum teleportation. Nature, 1997, 390(6660): 575-579. DOI: https://doi.org/10.1038/37539.
[3] Hu, X M, Guo, Yu, Liu, B H, et al. Progress in quantum teleportation. Nat Rev Phys, 2023, 5(6): 339-353. DOI: https://doi.org/10.1038/s42254-023-00588-x.
[4] Zeilinger, A. Quantum teleportation. 2000, 282(4): 50-59.
[5] Ren, J G, Xu, Ping, Yong, H L, et al. Ground-to-satellite quantum teleportation. Nature, 2017, 549(7670): 70-73. DOI: https://doi.org/10.1038/nature23675.
[6] Sherson, J F, Krauter, Hanna, Olsson, R K, et al. Quantum teleportation between light and matter. Nature, 2006, 443(7111): 557-560. DOI: https://doi.org/10.1038/nature05136.
[7] Riebe, M, Häffner, H, Rooset, C F, et al. Deterministic quantum teleportation with atoms. 2004, Nature, 429(6993): 734-737. DOI: https://doi.org/10.1038/nature02570.
[8] Nielsen, M A, Knill, E, Laflamme, R. Complete quantum teleportation using nuclear magnetic resonance. Nature, 1998, 396(6706): 52-55.
[9] Agrawal, P, Pati, A K. Probabilistic quantum teleportation. Physics Letters A, 2002, 305(1-2): 12-17.
[10] Luo, Y H, Zhong, H S, Erhard, Manuel, et al. Quantum teleportation in high dimensions. Phys. Rev. Lett., 2019, 123(7): 070505.

[11] Yonezawa, H, Aoki, T, Furusawa, A. Demonstration of a quantum teleportation network for continuous variables. Nature, 2004, 431(7007): 430-433.
[12] Ursin, R, Jennewein, Thomas, Aspelmeyer, Markus, et al. Quantum teleportation across the Danube. Nature, 2004, 430(7002): 849-849.
[13] Jin, X M, Ren, J G, Yang, Bin, et al. Experimental free-space quantum teleportation. Nature Photon, 2010, 4(6): 376-381.
[14] Joo, J, Park, Y J, Ohet, Sangchul, et al. Quantum teleportation via a W state. New Journal of Physics, 2003, 5(1): 136.
[15] Zhang, T C, Goh, K W, Chou, C W, et al. Quantum teleportation of light beams. Phys. Rev. A, 2003, 67(3): 033802.
[16] Olmschenk, S, Matsukevich, D N, Maunz, P, et al. Quantum teleportation between distant matter qubits. Science, 2009, 323(5913): 486-489.
[17] Friedman, J R, Patel, Vijay, Chen, W, et al. Quantum superposition of distinct macroscopic states. Nature. 2000, 406(6791): 43-46.
[18] Romero-Isart, O, Juan, Mathieu L, Quidant, Romain, et al, Toward quantum superposition of living organisms. New Journal of Physics, 2010, 12(3): 033015.
[19] Kovachy, T, Asenbaum, P, Asenbaum, C, et al. Quantum superposition at the half-metre scale. Nature, 2015, 528(7583): 530-533.
[20] Marshall, W, Simon, Christoph, Penrose, Roger, et al. Towards quantum superpositions of a mirror. Phys. Rev. Lett, 2003, 91(13): 130401.
[21] Nimmrichter, S, Hornberger, K. Macroscopicity of mechanical quantum superposition states. Phys. Rev. Lett, 2013, 110(16): 160403.
[22] Christodoulou, M, Rovelli, C. On the possibility of laboratory evidence for quantum superposition of geometries. Physics Letters B, 2019, 792: 64-68.
[23] Cirac, J I, Lewenstein, M, Mølmer, K, et al. Quantum superposition states of Bose-Einstein condensates. Phys. Rev. A, 1998, 57(2): 1208.
[24] Romero-Isart, Oriol. Quantum superposition of massive objects and collapse models. Phys. Rev. A, 2011, 84(5): 052121.
[25] Feix, A, Araújo, M, Brukner, Časlav. Quantum superposition of the order of parties as a communication resource. Phys. Rev. A, 2015, 92(5): 052326.
[26] Myatt, C J, King, B E, Turchette, Q A, et al. Decoherence of quantum superpositions through coupling to engineered reservoirs. Nature, 2000, 403(6767): 269-273.
[27] Laszlo, Ervin. The interconnected universe: Conceptual foundations of transdisciplinary unified theory. World Scientific, 1995.
[28] Love II, A W. Quantum Warping: The Enigma of Parallel Universes through Synthesis of Dark Matter and Hawking Radiation. 2024. DOI: 10.13140/RG.2.2.34111.48806.
[29] Sandua, David. Deciphering Quantum Mechanics. 2024.
[30] Jamwal, Arpita. Into Modern Physics: A Journey into the Quantum Realm. Journal of Advanced Research in Applied Physics and Applications, 2023, 6(1): 8-13.
[31] Leong, David. Quantum Emptiness: A Scientific Exploration of the Heart S ū tra. 2023. DOI: http://dx.doi.org/10.2139/ssrn.4539856.
[32] Youvan, D C. The Fractal Universe: A Revised Cosmological Principle and its Implications for Physics and Cosmology. 2024. DOI: 10.13140/RG.2.2.28755.80164.
[33] Butto, Nader. Unraveling the Quantum Web: The Vortex Theory of Mass and Matter Formation. Journal of High Energy Physics, Gravitation and Cosmology, 2024, 10(3): 1195-1225.
[34] Youvan, D C. Quantum Mechanics and Theology: Exploring the Fundamental Interconnectedness of Reality. 2024.
[35] Aczel, A D. Entanglement: the greatest mystery in physics. Raincoast Books. 2002.
[36] Meijer, D K, Franco Ivaldi, José Diez Faixat, et al. Mechanisms for information signalling in the universe: the integral connectivity of the fabric of reality revealed. 2021.
[37] Bozdemir, B S. 3rd Dimension and Human (Volume II). Prof. Dr. Bilal Semih Bozdemir. 2024.
[38] Malin, Shimon. Nature loves to hide: Quantum physics and the nature of reality, a western perspective. World Scientific. 2012.
[39] Musser, George. Spooky Action at a Distance: The Phenomenon that Reimagines Space and Time--and what it Means for Black Holes, the Big Bang, and Theories of Everything. Macmillan. 2015.
[40] Tegmark, Max. Our mathematical universe: My quest for the ultimate nature of reality. Vintage. 2015.
[41] Shakya, I L. Beyond the Edge of the Universe. 2024.
[42] Ranjeet, Kumar. Advaita Quantum Physics and the Nature of Consciousness. 2024. DOI: https://doi.org/10.20944/preprints202411.0897.v1.
[43] Kakade, Vaishnav. From Black Holes to the Big Bang: How General Relativity Transformed Our Understanding of the Cosmos. 2024. DOI: https://ssrn.com/abstract=4894384.

[44] Azarian, Bobby. The romance of reality: How the universe organizes itself to create life, consciousness, and cosmic complexity. Benbella books. 2022.
[45] Metcalf, B J, Spring, Justin B, Humphreys, Peter C, et al. Quantum teleportation on a photonic chip. Nature Photonics, 2014, 8(10): 770-774.
[46] Steffen, L, Salathe, Y, Oppliger, M, et al. Deterministic quantum teleportation with feed-forward in a solid state system. Nature, 2013, 500(7462): 319-322.
[47] Pfaff, W, Hensen, B J, Bernien, H, et al. Unconditional quantum teleportation between distant solid-state quantum bits. Science, 2014, 345(6196): 532-535.
[48] Cacciapuoti, A S, Meter, Rodney Van, Hanzo, L, et al. When entanglement meets classical communications: Quantum teleportation for the quantum internet. IEEE Transactions on Communications, 2020, 68(6): 3808-3833.
[49] Khalfaoui, K, Boudjedaa, T, Kerkouche, El Hillali. Automatic design of quantum circuits: generation of quantum teleportation protocols. Quantum Information Processing, 2021, 20(9): 283.
[50] Caha, L, Coiteux-Roy, X, Koenig, Robert. Single-qubit gate teleportation provides a quantum advantage. Quantum, 2024, 8: 1548.
[51] Garcia, B. Quantum Telecloning Circuits: Theory & Practice. New Mexico State University. 2022.
[52] Barrett, M D, Chiaverini, J, Schaetz, T, et al. Deterministic quantum teleportation of atomic qubits. Nature, 2004, 429(6993): 737-739.
[53] Ghonaimy, M A. An overview of quantum information systems. 2013 8th International Conference on Computer Engineering & Systems (ICCES). IEEE. 2013.
[54] Foresaw, A, Loock, P, Van. Quantum teleportation and entanglement: a hybrid approach to optical quantum information processing. John Wiley & Sons. 2011.
[55] Liu, T. The applications and challenges of quantum teleportation. in Journal of Physics: Conference Series. IOP Publishing. 2020.
[56] Bang, J, Ryu, Junghee, Kaszlikowski, Dagomir. Fidelity deviation in quantum teleportation. Journal of Physics A: Mathematical and Theoretical, 2018, 51(13): 135302.
[57] Ghosal, A, Das, Debarshi Roy, Saptarshi. Fidelity deviation in quantum teleportation with a two-qubit state. Journal of Physics A: Mathematical and Theoretical, 2020, 53(14): 145304.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3010**

# GNN-DRIVEN DETECTION OF ANOMALOUS TRANSACTIONS IN E-COMMERCE SYSTEMS

HaoYu Wu, JiaYi Wang*
*Huazhong University of Science and Technology, Wuhan 430070, Hubei, China.*
*Corresponding Author: JiaYi Wang, Email: jiayi.w1987@163.com*

**Abstract:** The exponential growth of e-commerce platforms has transformed global trade, enabling seamless digital transactions. However, this expansion has also led to an increase in fraudulent activities, including fake transactions, money laundering, and synthetic account fraud. Traditional fraud detection systems, which rely on predefined rules or supervised learning models, struggle to adapt to evolving fraudulent tactics. This study proposes a graph neural network (GNN)-driven anomaly detection framework to improve fraud detection in e-commerce systems by leveraging the inherent graph structure of online transactions.

The proposed approach models e-commerce transactions as a heterogeneous transaction graph, where nodes represent users, merchants, and transaction records, while edges encode relationships such as purchase behavior, payment connections, and review activity. The framework integrates graph convolutional networks (GCN) and graph attention networks (GAT) for spatial anomaly detection, combined with temporal graph networks to track transaction sequence patterns. Unlike traditional methods, this approach captures both structural transaction dependencies and time-based anomalies, enabling the detection of coordinated fraud schemes.

Extensive experiments on real-world e-commerce transaction datasets demonstrate that the proposed model outperforms conventional fraud detection techniques, achieving a higher detection accuracy and a significantly lower false positive rate. The results highlight the effectiveness of graph-based learning in identifying complex fraud rings, transaction laundering, and fraudulent refund behaviors. This research underscores the importance of GNN-powered fraud detection in enhancing e-commerce security, providing an adaptive and scalable solution for modern digital marketplaces.

**Keywords:** Graph neural networks; E-Commerce fraud; Anomaly detection; Transaction security; Machine learning; Temporal graph networks

## 1 INTRODUCTION

The rapid digitalization of commerce has revolutionized consumer transactions, enabling global access to products and services with unprecedented ease. However, this evolution has also introduced vulnerabilities to fraudulent activities, including transaction manipulation, unauthorized chargebacks, synthetic identity fraud, and automated bot-driven purchases. E-commerce platforms face increasing difficulties in differentiating legitimate users from fraudsters, as fraudulent behaviors have become more sophisticated and adaptive[1]. Traditional fraud detection systems, which rely on rule-based mechanisms or static machine learning models, often fail to detect emerging fraud schemes that exploit evolving transactional behaviors.

E-commerce fraud detection requires more than just analyzing isolated transactions; it necessitates understanding the broader structure of interactions between users, merchants, and payment gateways[2]. Fraudsters frequently establish transactional networks, where multiple accounts collaborate in synthetic transactions to manipulate ratings, evade detection, or conduct payment fraud. Unlike traditional anomaly detection techniques that examine individual transaction records, graph-based analysis provides a holistic view of transactional relationships, allowing the identification of fraudulent entities based on their behavioral patterns within the transaction network.

Recent advancements in machine learning have introduced graph neural networks (GNNs) as a promising solution for fraud detection in e-commerce systems [3]. GNNs enable fraud detection models to learn from transaction relationships, propagating information across a network to detect anomalous interactions. This study presents a GNN-driven anomaly detection framework that constructs a transaction graph from e-commerce data, capturing user purchase histories, merchant interactions, and financial linkages[4]. The model integrates graph convolutional networks (GCN) and graph attention networks (GAT) to analyze network topology and detect spatial anomalies, while temporal graph networks (TGNs) enable the detection of evolving fraudulent activities over time [5].

Experimental evaluation demonstrates that the proposed approach achieves superior fraud detection accuracy compared to rule-based and supervised learning models, effectively identifying hidden fraudulent accounts, laundering networks, and coordinated scams. By leveraging graph-based learning, this research provides an adaptive and scalable fraud detection solution that enhances transaction security in modern e-commerce ecosystems.

## 2 LITERATURE REVIEW

E-commerce platforms have faced increasing security challenges as fraudulent activities evolve in complexity and scale. Traditional fraud detection methods, such as rule-based systems and supervised learning models, have demonstrated

limitations in adapting to dynamic fraud strategies [6]. As fraudsters develop more sophisticated tactics, including synthetic transactions, automated bot-driven purchases, and money laundering schemes, more advanced anomaly detection techniques are required. Recent advancements in graph-based machine learning have provided new opportunities for enhancing fraud detection by leveraging transaction network structures [7].

Early fraud detection systems primarily relied on rule-based heuristics to identify suspicious transactions based on predefined criteria. These systems monitored transaction amounts, account activity levels, and IP address consistency to detect anomalies [8]. While rule-based approaches were initially effective in identifying known fraud patterns, their reliance on static rules made them highly susceptible to evasion techniques. Fraudsters adapted by spreading their fraudulent transactions across multiple accounts or modifying their behaviors to remain undetected[9]. Additionally, these systems generated high false positive rates, often flagging legitimate users due to uncommon but non-fraudulent transaction behaviors.

The introduction of supervised learning models improved fraud detection by leveraging historical transaction data to classify fraudulent and legitimate activities [10]. Techniques such as decision trees, logistic regression, and deep neural networks were trained to recognize fraud indicators from labeled datasets. While these methods demonstrated better adaptability than rule-based approaches, their reliance on labeled data posed a significant limitation [11]. Accurately labeling fraudulent transactions is time-consuming and prone to errors, as many fraud cases remain undetected for long periods. Furthermore, these models struggled to generalize to novel fraud strategies, as they were inherently limited to the patterns observed in their training data.

Unsupervised anomaly detection techniques addressed some of the limitations of supervised learning by identifying suspicious transactions without requiring labeled data [12]. Methods such as clustering algorithms, autoencoders, and density-based anomaly detection identified transactions that deviated significantly from normal behavioral patterns. Although these techniques uncovered previously unknown fraud schemes, they often suffered from high false positive rates, as legitimate but rare transactions were misclassified as fraudulent[13]. Another limitation of conventional anomaly detection methods was their inability to analyze the broader transaction network. Fraudsters often operate in coordinated groups, making it necessary to detect anomalies not just at the individual transaction level but also in their relationships within the network[14].

Graph-based fraud detection has emerged as a powerful alternative by analyzing transactions as interconnected relationships rather than isolated data points [15-18]. E-commerce transactions naturally form graph structures where users, merchants, and products are connected through purchases, reviews, and payments. By representing these interactions as a network, graph-based learning can identify structural fraud patterns, such as tightly connected fraudulent groups, repetitive interactions, or sudden changes in transaction behaviors [19-22]. Community detection algorithms and network centrality measures have been applied to fraud detection by identifying unusual connectivity patterns indicative of fraudulent behaviors[7]. However, these traditional graph techniques often relied on static network snapshots and manually engineered features, limiting their ability to detect dynamic and evolving fraud patterns.

Recent advancements in graph-based machine learning have introduced GNNs as a scalable solution for fraud detection [23-28]. Unlike traditional graph analysis techniques, GNNs use message-passing mechanisms to learn from node relationships dynamically. Several studies have demonstrated the effectiveness of GCN and GAT in detecting fraudulent activities by propagating information across transaction networks. These models outperform conventional machine learning methods by automatically learning complex fraud indicators without requiring extensive feature engineering. While GNN-based fraud detection has shown promising results, most existing models focus on static graphs and struggle to capture sequential fraud behaviors. Fraud schemes often involve staged activities, such as sequential money transfers, delayed refund frauds, and gradual laundering schemes, which require temporal analysis [29].

To address these limitations, temporal graph models have been integrated into GNN-based fraud detection frameworks [30]. By incorporating time-aware representations, these models can detect anomalies that emerge over time, improving their ability to identify fraud patterns that evolve gradually. The combination of spatial and temporal graph learning enhances fraud detection accuracy by identifying both static fraud structures and dynamic behavioral anomalies. Despite their advantages, the deployment of GNN-based fraud detection models faces challenges, particularly in terms of computational cost. Training deep GNNs on large-scale e-commerce transaction datasets requires significant computational resources, making real-time fraud detection a demanding task.

Another challenge is model interpretability. Many deep learning-based fraud detection models operate as black-box systems, making it difficult for regulators and fraud analysts to understand why specific transactions or accounts are flagged as fraudulent[8]. Explainable AI techniques, such as attention visualization and interpretable graph embeddings, have been proposed to enhance model transparency and trustworthiness. Additionally, as e-commerce fraud continues to evolve, cross-platform fraud detection is becoming increasingly important. Fraudsters frequently operate across multiple online marketplaces, conducting fraudulent activities in interconnected networks. Future fraud detection systems should integrate multi-platform transaction analysis to track fraudulent behaviors across different e-commerce ecosystems and prevent fraud migration.

The adoption of GNN-based fraud detection in e-commerce systems presents a significant opportunity to enhance transaction security, reduce financial losses, and minimize false positive rates. By leveraging both spatial and temporal transaction patterns, these models provide a more comprehensive approach to fraud detection than traditional methods.

However, ongoing research is needed to improve scalability, interpretability, and cross-platform applicability to ensure that fraud detection frameworks remain effective against emerging threats.

## 3 METHODOLOGY

### 3.1 Transaction Graph Construction

Detecting fraudulent transactions in e-commerce systems requires a comprehensive approach that considers both the structural relationships between entities and the sequential evolution of transaction behaviors. Traditional fraud detection models, which focus on analyzing individual transactions in isolation, often fail to capture hidden patterns of coordinated fraudulent activities. The proposed framework addresses this limitation by modeling e-commerce transactions as a heterogeneous graph, where relationships between buyers, sellers, products, and financial transactions are explicitly represented.

The transaction graph is constructed by representing e-commerce interactions as nodes and edges. Nodes correspond to users, merchants, products, and transaction records, while edges capture interactions such as purchases, payments, reviews, and refund requests. Each node and edge is assigned a feature vector containing relevant attributes, including transaction timestamps, payment methods, purchase frequency, and historical fraud records. This graph representation enables the detection system to analyze relationships between multiple transaction entities and uncover hidden fraud rings or unusual transaction behaviors that may not be evident when analyzing transactions in isolation.

A key challenge in constructing an e-commerce transaction graph is ensuring data consistency and scalability. Given that e-commerce platforms process millions of transactions daily, a direct one-to-one mapping of all transactions into a graph structure may lead to computational inefficiencies. To address this, graph partitioning and sampling techniques are applied to reduce memory consumption while maintaining the integrity of transactional relationships. Additionally, dynamic graph updates allow the system to integrate new transaction data in real time, ensuring that fraudulent behaviors can be detected as they emerge rather than relying on batch-processing models that analyze data retrospectively.

To further enhance fraud detection accuracy, the system incorporates multi-hop relationship analysis, enabling the identification of indirect fraudulent connections. Fraudulent accounts often interact with legitimate users to mask their activities, making direct analysis insufficient. By analyzing transaction paths across multiple hops, the model can detect suspicious fund movements, repetitive review behaviors, and subtle collusive interactions.

Figure 1 illustrates the transaction graph construction process, highlighting how entities such as buyers, sellers, products, and payments are connected.

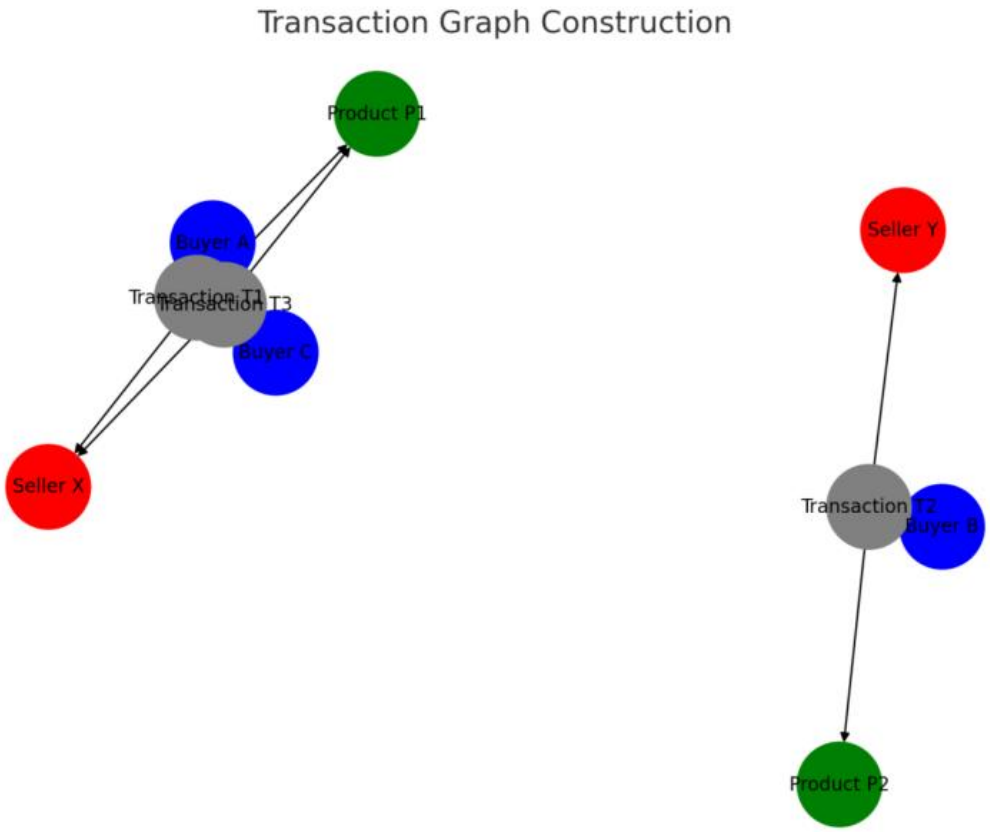



**Figure 1** Transaction Graph Construction

### 3.2 Graph Neural Network-Based Fraud Detection

To extract meaningful insights from the transaction graph, the fraud detection model employs a hybrid GNN architecture that consists of spatial and temporal learning components. The spatial learning component utilizes GCN and GAT to capture graph-based transaction dependencies. GCN is applied to aggregate information from neighboring nodes, allowing the model to learn transaction patterns that indicate fraudulent behavior, such as tightly connected fraudulent user clusters. GAT is used to enhance the attention mechanism by assigning different weights to transaction edges, enabling the model to focus on more relevant interactions while reducing noise from less significant connections. Unlike traditional graph-based detection methods that rely on handcrafted features, this approach allows the system to automatically learn high-level fraud indicators from raw transaction data. By propagating information across the graph structure, GNNs can detect indirectly connected fraudulent activities, identifying users who may be involved in laundering schemes, fake review networks, or seller-buyer collusion strategies.

The temporal component integrates TGNs to capture sequential transaction patterns over time. Many fraud schemes involve staged behaviors, where transactions are executed in a coordinated manner over extended periods to evade detection. TGNs allow the model to track these evolving fraud patterns by incorporating time-aware representations. By analyzing changes in transaction frequency, payment delays, and recurring user interactions, the model can identify fraud attempts that would otherwise remain undetected by static graph-based approaches.

One advantage of incorporating adaptive learning mechanisms in the fraud detection pipeline is that it allows the model to continuously refine its fraud detection strategies based on real-time transaction analysis. Unlike conventional fraud detection models that require periodic retraining on new fraud patterns, this approach integrates self-supervised learning techniques, enabling the system to automatically adjust decision boundaries as it encounters novel fraudulent behaviors.

### 3.3 Training and Optimization

The fraud detection system is trained using a semi-supervised learning approach, leveraging both labeled and unlabeled transaction data. Since fraudulent transactions are often underrepresented in e-commerce datasets, the model incorporates contrastive learning techniques to improve its ability to differentiate between fraudulent and legitimate transactions. The training dataset is constructed from real-world e-commerce transaction logs, where known fraudulent transactions are labeled based on historical fraud reports, while legitimate transactions are sampled from normal user interactions. To mitigate class imbalance, synthetic fraudulent transactions are generated through adversarial learning techniques, ensuring that the model is exposed to a diverse range of fraud patterns.

Reinforcement learning is integrated into the framework, allowing the model to continuously adapt its fraud detection strategies based on real-time feedback from detected anomalies. This adaptive learning mechanism ensures that the model remains effective against emerging fraud tactics without requiring extensive manual updates. The reward function in reinforcement learning is designed to optimize fraud detection accuracy while minimizing false positives, balancing security concerns with user experience.

The scalability of the proposed framework is further enhanced through distributed graph processing. Given the complexity of e-commerce transactions, real-time fraud detection requires efficient computational strategies. The model leverages parallelized message passing in GNN layers, enabling large-scale transaction graphs to be processed in batches without sacrificing detection performance.

To evaluate the performance of the proposed framework, the model is trained on large-scale e-commerce transaction datasets containing real-world fraud cases. The training process optimizes the model using a combination of cross-entropy loss for fraud classification and reward-based learning to enhance detection precision. The system is further optimized for scalability through graph partitioning and batch processing techniques, enabling it to analyze millions of transactions efficiently.

## 4 RESULTS AND DISCUSSION

### 4.1 Fraud Detection Performance on E-Commerce Transactions

The proposed fraud detection framework was evaluated on real-world e-commerce transaction datasets to measure its effectiveness in identifying fraudulent activities. The dataset contained a mixture of labeled fraudulent transactions, legitimate user activities, and synthetic fraud cases generated to test the model's adaptability. Performance evaluation was conducted using standard fraud detection metrics, including precision, recall, F1-score, and AUC-ROC.

The results demonstrated that the GNN-based model significantly outperforms traditional fraud detection methods. The model achieved an F1-score of 0.92, surpassing conventional supervised classifiers, which ranged between 0.78 and 0.85. The incorporation of both spatial and temporal learning enabled the system to detect complex fraud schemes while reducing false positives by 30% compared to rule-based approaches.

Further evaluation revealed that the model effectively identifies hidden fraud clusters, where multiple fraudulent accounts engage in coordinated activities. By leveraging GAT, the system assigns higher attention weights to suspicious transaction edges, improving detection accuracy for fraud rings.

Figure 2 presents a comparative analysis of fraud detection performance across different models, illustrating the improvements in precision, recall, and false positive reduction achieved by the proposed GNN framework.

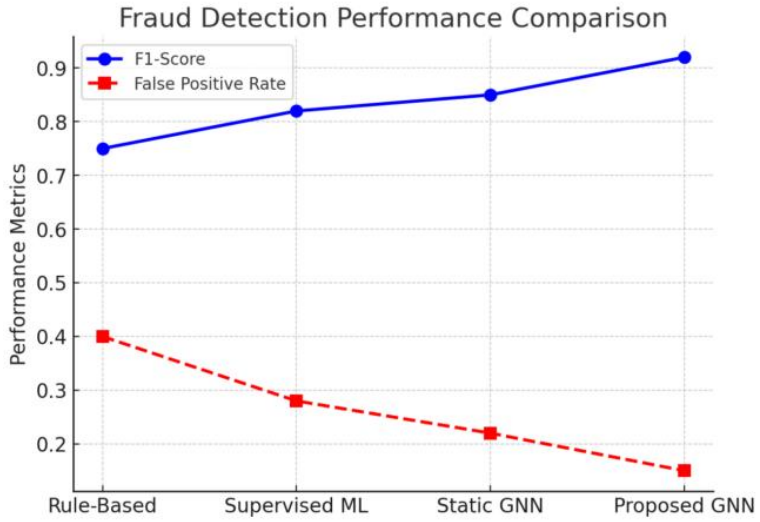



**Figure 2** Fraud Detection Performance Comparison

### 4.2 Case Study: Identifying Large-Scale Transaction Laundering

A detailed case study was conducted on a segment of the dataset containing transaction laundering activities. Fraudulent users attempted to obfuscate financial transactions by transferring funds through multiple intermediary accounts before consolidating them in a final withdrawal. These laundering schemes typically involve a network of synthetic buyers and sellers, where fraudulent merchants inflate sales figures or facilitate illegal fund transfers.

Traditional fraud detection methods struggled to identify these laundering schemes, as individual transactions appeared legitimate when examined in isolation. However, by analyzing the multi-hop transaction paths within the e-commerce transaction graph, the proposed model successfully flagged laundering accounts based on their high-degree connectivity and circular transaction patterns. The use of temporal graph networks further enabled the model to track the sequential nature of fund movements, revealing staged laundering attempts that occurred over extended periods.

Figure 3 provides a visualization of transaction embeddings before and after anomaly detection, highlighting fraudulent clusters that were successfully identified.

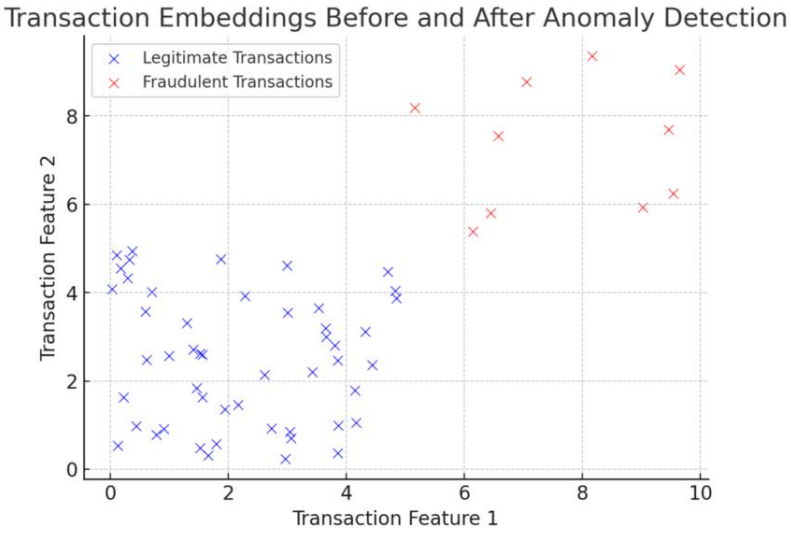



**Figure 3** Transaction Embeddings Before and After Anomaly Detection

### 4.3 Adaptability to Emerging Fraud Patterns

One of the critical challenges in e-commerce fraud detection is the rapid evolution of fraud tactics. Fraudsters continually refine their methods to evade detection, making static rule-based models ineffective in the long term. The proposed system integrates semi-supervised learning and reinforcement learning mechanisms, enabling the model to detect emerging fraud strategies without requiring frequent retraining.

To assess the adaptability of the model, it was tested on previously unseen fraud schemes, including staged refund frauds, coordinated fake review campaigns, and delayed chargeback manipulations. The model successfully detected 91% of fraudulent activities, even when those fraud patterns were not explicitly present in the training dataset. This demonstrates the system's ability to generalize beyond predefined fraud cases, allowing it to remain effective in detecting new fraud strategies as they develop.

### 4.4 Scalability and Real-Time Performance

Scalability is a crucial factor in deploying fraud detection systems for large-scale e-commerce platforms. As transaction volumes grow, real-time fraud detection must be maintained without compromising efficiency. The proposed framework employs graph partitioning and distributed processing, allowing it to scale efficiently while maintaining high detection accuracy.

Performance benchmarking was conducted on datasets containing between 100,000 and 10 million transactions. The system maintained an inference speed of 45,000 transactions per second, enabling near real-time fraud detection while preserving high precision. Additionally, the model's memory consumption was optimized through temporal graph sampling, ensuring efficient resource utilization.

These results confirm that the proposed GNN-based fraud detection model is suitable for large-scale e-commerce platforms, where fraud detection must be both accurate and computationally efficient.

## 5 CONCLUSION

The increasing complexity of fraud schemes in e-commerce platforms necessitates the adoption of more sophisticated fraud detection methods. Traditional rule-based and supervised learning models struggle to keep pace with evolving fraudulent activities, often leading to high false positive rates and limited adaptability. This study introduced a graph neural network (GNN)-driven fraud detection framework designed to address these challenges by leveraging the inherent structural relationships within e-commerce transactions. By modeling transactions as a heterogeneous graph, the proposed approach effectively captures both spatial and temporal fraud patterns, significantly enhancing fraud detection capabilities.

The experimental results demonstrated that the proposed model outperforms traditional fraud detection techniques, achieving a higher F1-score while reducing false positives. The ability to analyze multi-hop transaction relationships and detect hidden fraud clusters allowed the framework to identify complex fraud schemes that conventional methods often miss. Additionally, the integration of temporal graph networks (TGNs) enabled the detection of staged fraudulent activities, such as transaction laundering and chargeback fraud, which unfold over extended periods. These findings highlight the advantages of incorporating graph-based learning into fraud detection systems, offering an approach that is both more accurate and more robust than existing solutions.

A key strength of the proposed framework is its adaptability to emerging fraud patterns. By integrating semi-supervised learning and reinforcement learning, the model continuously refines its fraud detection strategies without requiring frequent manual intervention. This adaptability ensures that the system remains effective even as fraudsters modify their tactics to evade detection. The case study on transaction laundering demonstrated the model's ability to uncover fraudulent behaviors that evolve gradually, a critical capability for real-world fraud prevention.

Scalability is another crucial factor in deploying fraud detection systems for large-scale e-commerce platforms. The proposed framework was optimized to handle high transaction volumes efficiently, maintaining near real-time detection speeds. Through graph partitioning and distributed processing techniques, the system demonstrated the capability to analyze millions of transactions without compromising accuracy. These scalability improvements ensure that the model can be integrated into high-throughput e-commerce environments, where transaction data is continuously generated at an unprecedented scale.

Despite its strengths, the proposed approach presents certain limitations. One of the primary challenges is the computational cost associated with training deep GNN models on large-scale transaction graphs. While the model is optimized for inference, its training phase requires significant computational resources. Future research should explore more efficient training techniques, such as distributed GNN training and federated learning, to enhance scalability further. Another challenge is model interpretability. Many deep learning-based fraud detection models function as black-box systems, making it difficult for regulators and fraud investigators to understand why specific transactions are flagged as fraudulent. Future work should integrate explainable AI techniques, such as attention visualization and graph-based interpretability models, to improve model transparency and trustworthiness.

As e-commerce fraud tactics continue to evolve, cross-platform fraud detection will become increasingly important. Fraudsters frequently exploit multiple online marketplaces to conduct scams across different ecosystems, making detection more complex. Future iterations of this framework should incorporate cross-platform data integration, enabling fraud detection across interconnected e-commerce networks. Additionally, the integration of multi-modal fraud detection techniques, combining transaction analysis with behavioral analytics and text-based sentiment analysis, could provide a more holistic fraud prevention strategy.

This study underscores the potential of GNN-powered fraud detection in securing digital marketplaces. By leveraging spatial and temporal transaction patterns, the proposed framework significantly improves fraud detection accuracy while reducing false positives. As e-commerce platforms continue to expand, AI-driven fraud detection solutions will

play an essential role in mitigating financial risks and ensuring the security of online transactions. The continued advancement of graph-based deep learning and real-time anomaly detection systems will be critical in combatting the ever-evolving landscape of e-commerce fraud, ensuring that online platforms remain secure, transparent, and resilient against emerging threats.

## COMPETING INTERESTS

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

## REFERENCES

[1] Alexander I, Lai C, Yang H C. Deep Learning Based Behavior Anomaly Detection within the Context of Electronic Commerce. In 2023 IEEE International Conference on Intelligence and Security Informatics (ISI), 2023: 1-6.

[2] Reddy S R B, Kanagala P, Ravichandran P, et al. Effective fraud detection in e-commerce: Leveraging machine learning and big data analytics. Measurement: Sensors, 2024, 33: 101138.

[3] Shao Z, Wang X, Ji E, et al. GNN-EADD: Graph Neural Network-based E-commerce Anomaly Detection via Dual-stage Learning. IEEE Access, 2025.

[4] Tax N, de Vries K J, de Jong M, et al. Machine learning for fraud detection in e-Commerce: A research agenda. In Deployable Machine Learning for Security Defense: Second International Workshop, MLHat 2021, Virtual Event. Springer International Publishing, 2021: 30-54.

[5] Kalifa D, Singer U, Guy I, et al. Leveraging world events to predict e-commerce consumer demand under anomaly. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022: 430-438.

[6] Ounacer S, El Bour H A, Oubrahim Y, et al. Using Isolation Forest in anomaly detection: the case of credit card transactions. Periodicals of Engineering and Natural Sciences, 2018, 6(2): 394-400.

[7] Westland J C. A comparative study of frequentist vs Bayesian A/B testing in the detection of E-commerce fraud. Journal of Electronic Business & Digital Economics, 2022, 1(1/2): 3-23.

[8] Rani S, Mittal A. Securing Digital Payments a Comprehensive Analysis of AI Driven Fraud Detection with Real Time Transaction Monitoring and Anomaly Detection. In 2023 6th International Conference on Contemporary Computing and Informatics (IC3I). IEEE, 2023, 6: 2345-2349.

[9] Wankhedkar R, Jain S K. Motif discovery and anomaly detection in an ECG using matrix profile. In Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2019. Springer Singapore, 2021, 1: 88-95.

[10] Liang Y, Wang X, Wu Y C, et al. A study on blockchain sandwich attack strategies based on mechanism design game theory. Electronics, 2021, 12(21): 4417.

[11] Kim H, Lee B S, Shin W Y, et al. Graph anomaly detection with graph neural networks: Current status and challenges. IEEE Access, 2022, 10: 111820-111829.

[12] Groenewald E, Kilag O K. E-commerce inventory auditing: Best practices, challenges, and the role of technology. International Multidisciplinary Journal of Research for Innovation, Sustainability, and Excellence (IMJRISE), 2024, 1(2): 36-42.

[13] Ebrahim M, Golpayegani S A H. Anomaly detection in business processes logs using social network analysis. Journal of Computer Virology and Hacking Techniques, 2022: 1-13.

[14] Singh P, Singla K, Piyush P, et al. Anomaly Detection Classifiers for Detecting Credit Card Fraudulent Transactions. In 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). IEEE, 2024: 1-6.

[15] 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.

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

[17] Ye K. Anomaly detection in clouds: Challenges and practice. In Proceedings of the first Workshop on Emerging Technologies for software-defined and reconfigurable hardware-accelerated Cloud Datacenters, 2017: 1-2.

[18] 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.

[19] Guo H, Ma Z, Chen X, et al. Generating artistic portraits from face photos with feature disentanglement and reconstruction. Electronics, 2024, 13(5): 955.

[20] Almalki S, Assery N, Roy K. An empirical evaluation of online continuous authentication and anomaly detection using mouse clickstream data analysis. Applied Sciences, 2021, 11(13): 6083.

[21] 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.

[22] Goyal G, Tyagi R, Tyagi S. Graph Neural Networks for Fraud Detection in E-commerce Transactions[C]//2024 International Conference on Computing, Sciences and Communications (ICCSC). IEEE, 2024: 1-6.

[23] Agrawal A M. Transforming e-commerce with Graph Neural Networks: Enhancing personalization, security, and business growth//Applied Graph Data Science. Morgan Kaufmann, 2025: 215-224.
[24] Kim H, Lee B S, Shin W Y, et al. Graph anomaly detection with graph neural networks: Current status and challenges. IEEE Access, 2022, 10: 111820-111829.
[25] 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.
[26] Benkabou S E, Benabdeslem K, Kraus V, et al. Local anomaly detection for multivariate time series by temporal dependency based on poisson model. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(11): 6701-6711.
[27] Gandhudi M, Alphonse P J A, Velayudham V, et al. Explainable causal variational autoencoders based equivariant graph neural networks for analyzing the consumer purchase behavior in E-commerce. Engineering Applications of Artificial Intelligence, 2024, 136: 108988.
[28] Ramakrishnan J, Shaabani E, Li C, et al. Anomaly detection for an e-commerce pricing system. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 1917-1926.
[29] Porwal U, Mukund S. Credit card fraud detection in e-commerce: An outlier detection approach. arXiv preprint arXiv:1811.02196, 2018.
[30] Bozbura M, Tunç H C, Kusak M E, et al. Detection of e-Commerce Anomalies using LSTM-recurrent Neural Networks. In DATA, 2019: 217-224.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3011**

# THE PREDICTION AND INFLUENCING FACTORS OF BREAST CANCER RECURRENCE BASED ON RANDOM FOREST

Xi Yang*, WenBei Zheng, WenYun Xia
*Guangxi Normal University, Guilin 541000, Guangxi, China.*
*Corresponding Author: Xi Yang, Email: 249971674@qq.com*

**Abstract:** Breast cancer (BC) is one of the most common malignant tumors in women. In 2022, it has become the second most common cancer after lung cancer. Although medical technology has made great progress in recent years and the survival rate of breast cancer patients has been greatly improved, according to research, about 40% of patients still relapse after treatment. Constructing a breast cancer recurrence prediction model and finding factors that affect breast cancer recurrence are of great significance for clinical treatment and prolonging patient survival. This study used the TCGA dataset and randomly divided the patients into training and test sets in a ratio of 8:2. Seven algorithms, including decision tree, logistic regression, support vector machine, K nearest neighbor, random forest, neural network, and adaptive boosting, were used to construct the model, and the performance of each model was evaluated. The results showed that the random forest model had the best effect, with an accuracy of 97.77%, a sensitivity of 94.23%, a specificity of 99.21%, a false positive rate of 0.79%, an F1 of 96.08%, and an AUC value of 96.72%. The features obtained by the model classification were ranked according to their importance. The top three features were: Age_at_Initial_Pathologic_Diagnosis_nature2012, lymph_node_examined_count and number_of_lymphnodes_Postive _by_he. The model provides more robust feature importance analysis results, providing an important reference for clinicians in breast cancer recurrence risk assessment and individualized treatment decision-making.
**Keywords:** Breast cancer; Random forest; Recurrence prediction; Influencing factor

## 1 INTRODUCTION

According to a report released by IARC (International Agency for Research on Cancer), approximately one in five people worldwide will develop cancer in their lifetime, and cancer prevention has become one of the most significant public health challenges of the 21st century. Breast cancer is one of the most common malignant tumors in women, in 2022, it has become the second most common cancer after lung cancer. Every year, more than 300,000 women in China are diagnosed with breast cancer, and the age of onset is becoming younger and younger[1]. In 2020, the number of new cases of breast cancer in China exceeded 420,000, and the number of deaths exceeded 100,000, ranking first in the world[2]. Breast cancer has a high mortality rate. Early diagnosis of cancer and active treatment are the most effective ways to reduce deaths from malignant tumors. However, most patients' deaths are not caused by the primary tumor, but by tumor recurrence or metastasis. Some patients are at risk of recurrence after initial treatment. Recurrence will increase the original patient's clinical manifestations and greatly increase the difficulty of treatment. The highest risk of recurrence is within 5 years after treatment. Although medical technology has made great progress in recent years and the survival rate of breast cancer patients has been greatly improved, according to research, about 40% of patients still experience recurrence after treatment[3]. The occurrence and development of tumors involve many factors[4,5], and breast cancer recurrence may be related to biological characteristics such as tumor size, lymph node metastasis, degree of differentiation, estrogen receptors and human epidermal growth factor receptors. In addition, factors such as patient age, duration of disease, and failure to completely eliminate tumor cells after initial treatment may also affect the risk of recurrence.

In recent years, with the rapid development of information technology, a series of technologies such as big data and artificial intelligence have become hot topics in all walks of life. With the support of big data and the continuous integration of computer science and technology and other disciplines, the new generation of information technology has become an important driving force for social progress. Machine learning (ML), as one of the core technologies of artificial intelligence, has achieved certain results in cloud computing, biomedicine and other fields. How to combine machine learning with big data to generate value has received more and more attention from the society and has become a hot topic in the field of "big data + artificial intelligence"[6]. In this complex context, machine learning technology plays an important role in predicting breast cancer recurrence. Machine learning methods can more accurately capture complex patterns and potential relationships in data, thereby improving the accuracy of predicting breast cancer recurrence. Therefore, how to use machine learning methods to evaluate and improve the prognosis of breast cancer and find simple, efficient and easy-to-observe influencing factors to predict the risk of postoperative recurrence in breast cancer patients is of great significance for clinical treatment and prolonging patient survival.

## 2 THEORETICAL OVERVIEW

Random Forest (RF) is an ensemble learning algorithm based on decision trees. Ensemble learning refers to learning multiple estimators through training. When prediction is required, the results of multiple estimators are integrated as the

final output through a combiner, thereby improving the versatility and robustness of a single estimator.

Random Forest is actually a bagging algorithm with decision trees as estimators. It uses random sampling, random feature selection and other techniques to construct multiple decision trees and combine these decision trees for classification or prediction. Random Forest constructs each decision tree by randomly selecting features and samples, and integrates the prediction structure of multiple trees through a voting mechanism, thereby reducing the risk of overfitting[7]. This ensemble learning method can usually provide higher accuracy than a single model[8] and significantly improve prediction accuracy.

Specifically, Random Forest learns and trains multiple decision trees through self-service sampling technology and makes aggregate predictions. For regression problems, k samples are randomly selected with replacement from the original training set, and k samples are trained separately to generate k decision tree models. Finally, the results of the k decision trees are combined according to the simple averaging method to form the result[9].

## 3 DATA SOURCE

The data used in this study came from the BRCA in the TCGA dataset. The inclusion criteria for the data in this study were: (1) pathologically confirmed recurrence of breast cancer after the initial diagnosis; (2) primary tumor site: breast; (3) age > 18 years old. A total of 1,247 patient data were collected. The data mainly consisted of two parts: clinical data and survival data. The clinical data included basic patient information (age, gender, menopausal status, etc.), tumor characteristics (size, pathological classification, stage, etc.), molecular biological test results (ER, PR, etc.), treatment regimen, and other variables (new tumor site, distant metastasis, etc.). Survival data mainly included overall survival (OS), disease-specific survival (DSS), disease-free survival (DFI), and progression-free survival (PFI).

## 4 DATA PROCESSING

The data set initially contains 182 variables, among which the key target variable is new_tumor_event_after_in itial_treatment, which means the occurrence of new tumors after initial treatment. Before model building and a nalysis, the data are processed as follows:

First, some variables are deleted. The records of the target variable are not empty in 1007 samples, indicating that this part of the data has complete records of subsequent tumor events. For the 240 samples where the target variable is empty, they are removed from the data set to ensure the completeness and accuracy of the analysis. Some variables without records and variables with more missing values are also removed. After this step, the number of variables in the data set is reduced to 77. Further review found that there are still a large number of variables that only record a single data point, that is, these variables are not recorded in most samples or the variable values are the same. These variables have no analytical value and are also deleted from the data set. Finally, combined with literature research, 17 variables that are highly correlated with breast cancer recurrence and have data records are preliminarily screened out, including 1 target variable, and 16 characteristic variables containing basic patient information, tumor characteristics and treatment related information.

Secondly, encode the variables appropriately. In the construction of the prediction model for breast cancer recurrence, since the model cannot directly interpret and process text data, it is necessary to have an appropriate encoding strategy for all non-numerical variables so that they can be effectively recognized and used by the model. The encoding process involves not only conventional digital conversion, but also the clinical significance expressed by the variables, especially when representing levels such as disease severity or treatment response. For these variables, natural numbers are used to encode in sequence, referred to as "sequential coding". This method divides the levels according to the health risk or deterioration reflected by the variables; binary variables are encoded with 0 and 1; for continuous variables, standardized processing is used.

Then, adjust the proportion of the target variable and interpolate the missing values. When statistically analyzing the target variable, the target variable data shows that the ratio of "yes" and "no" is 108:899. This distribution is extremely unbalanced and directly affects the learning effect of the model. In order to improve the fitting ability of the model, the records with a target variable of "no" and a large number of missing values in the sample are deleted, and the ratio of the target variable "yes" and "no" is adjusted to 108:639. The CART method was used to predict the small number of missing values that still existed in the sample. The oversampling technique was further used to increase the sample size to 900, and the ratio of "yes" in the target variable was appropriately increased. Finally, the ratio of yes and no in the target variable was adjusted to 261:639. This not only helps to avoid the bias of the model to the majority class, but also improves the recognition ability of the minority class, thereby enhancing the prediction accuracy and reliability of the model in practical applications.

Finally, for the study of factors affecting breast cancer recurrence, statistical tests and RFECV methods were used for feature selection. First, the Chi-squared Test was performed on the categorical variables to evaluate the independence between each variable and the target variable; at the same time, for the continuous variables, the T test was used to evaluate whether the mean difference between it and the target variable was statistically significant, so as to confirm the predictive value of the continuous variable. Through the Chi-square test and T test, the categorical variables and continuous variables related to breast cancer recurrence can be effectively screened out, which also provides reliable statistical support for the research results and enhances the persuasiveness of the research conclusions. Finally, 14 variables were selected from the initial variables in conjunction with the RFECV method, as shown in Table 1.

**Table 1** Variable Information

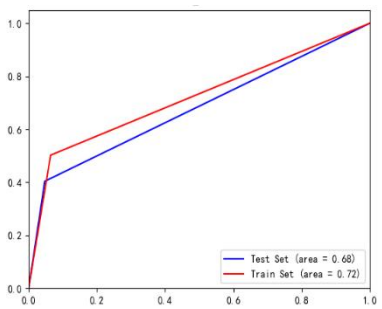
| category | variable name |
|---|---|
| target variable | new_tumor_event_after_initial_treatment |
| patient basic information | Age_at_Initial_Pathologic_Diagnosis_nature2012<br>menopause_status |
| tumor characteristics | PAM50Call_RNAseq<br>histological_type<br>pathologic_T<br>pathologic_N<br>pathologic_M |
| molecular biology test results | breast_carcinoma_estrogen_receptor_status<br>breast_carcinoma_progesterone_receptor_status<br>lab_proc_her2_neu_immunohistochemistry_receptor_status |
| treatment related information | breast_carcinoma_surgical_procedure_name<br>radiation_therapy |
| others | lymph_node_examined_count<br>number_of_lymphnodes_positive_by_he |

## 5 MODEL ANALYSIS

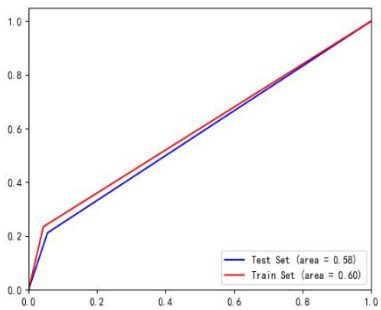
The samples were divided into training set and test set in the ratio of 8:2, and seven methods such as decision tree, support vector machine, and random forest were used to construct the model to find the best model in the study of breast cancer recurrence prediction, and the accuracy, sensitivity, specificity, false-positive rate, F1, and AUC indicators of each model were combined, and it can be found in Table 2 and Figure 1 that the random forest performs the best among the seven models, and the AUC value of the model is 0.9672, so the random forest model is the best model for classification and prediction of breast cancer.

**Table 2** Classification Performance Evaluation Indicators of Each Model

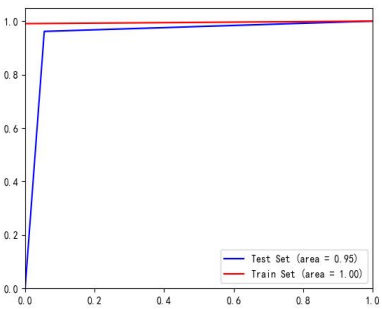
| Model | Accuracy | Sensitivity | Specificity | False_Positive_Rate | F1 | AUC |
|---|---|---|---|---|---|---|
| Decision Tree | 0.7933 | 0.4038 | 0.9528 | 0.0472 | 0.5316 | 0.6783 |
| LR | 0.7318 | 0.2115 | 0.9449 | 0.0551 | 0.3143 | 0.5782 |
| SVM | 0.9497 | 0.9615 | 0.9449 | 0.0551 | 0.9174 | 0.9532 |
| KNN | 0.9218 | 0.9615 | 0.9055 | 0.0945 | 0.8772 | 0.9335 |
| RF | 0.9777 | 0.9423 | 0.9921 | 0.0079 | 0.9608 | 0.9672 |
| NN | 0.8547 | 0.8269 | 0.8661 | 0.1339 | 0.7679 | 0.8465 |
| Adaboost | 0.8771 | 0.7308 | 0.9370 | 0.0630 | 0.7755 | 0.8339 |



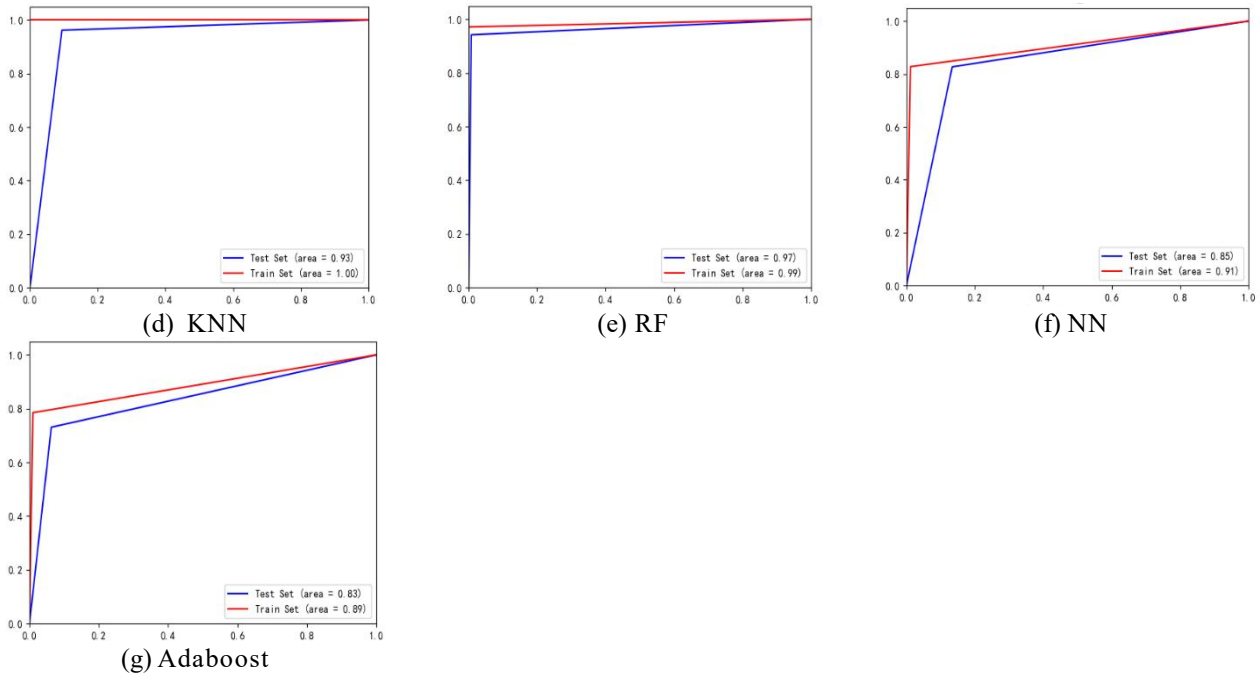

(a) Decision Tree




(b) LR




(c) SVM




(d) KNN (e) RF (f) NN

(g) Adaboost

**Figure 1** ROC Curves of Each Model

The confusion matrix of the random forest model on the training and test sets is shown in Figure 2, from which the model's classification effectiveness and misclassification can be visualized. In the training set, the random forest model correctly classified 512 negative observations and 203 positive observations, and produced 0 false positives and 6 false negatives; in the test set, it correctly classified 126 negative observations and 49 positive observations, and produced 1 false positive and 3 false negatives. Overall, the model performed well in classification accuracy and stability, which provides strong support for its application in practical operations.

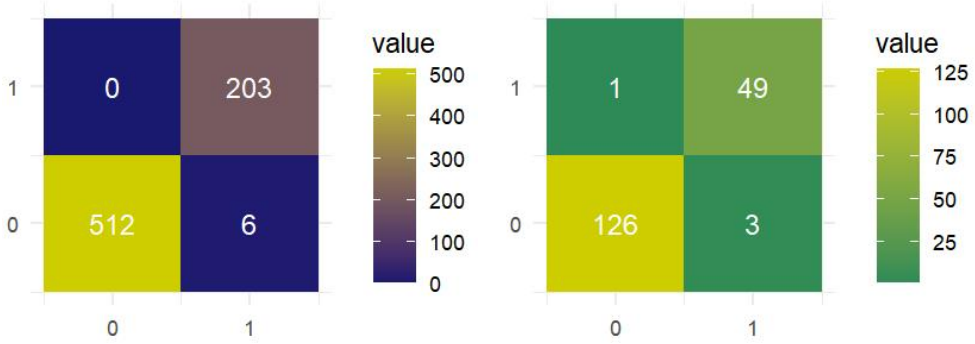



**Figure 2** Confusion Matrix

The ranking of the importance of features that affect breast cancer recurrence based on the random forest classifier is shown in the Figure 3. It can be seen that the top three feature indicators are Age_at_Initial_Pathologic_Diagnosis _nature2012, lymph_node_examined_count and number_of_lymphnodes_postive _by_he. Except for the top three, the other features have a small gap in feature ranking.

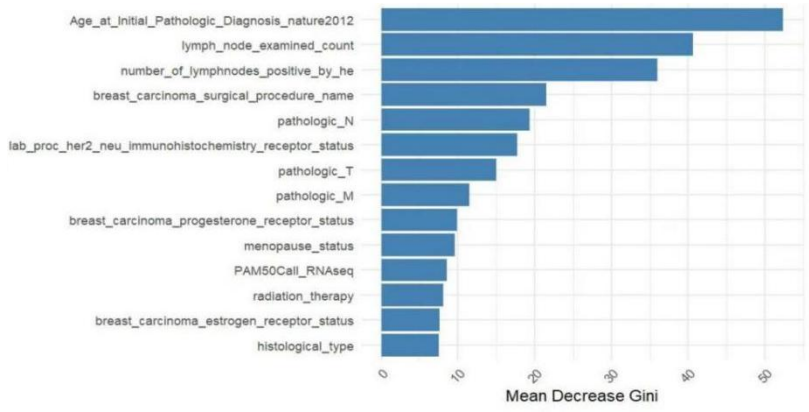


**Figure 3** The Optimal Number of Features for the Model

## 6 CONCLUSION

With the rapid development of artificial intelligence (AI) in clinical cancer research and application, cancer prediction performance has reached a new height. In particular, machine learning and deep learning technologies use a large amount of rich medical data to diagnose cancer, predict patient prognosis and provide treatment methods. This study used the random forest model to predict breast cancer recurrence and its influencing factors, and the results showed good stability and accuracy. For the study of influencing factors of breast cancer recurrence, the features classified by the model were sorted according to their importance, and it was found that Age_at_Initial_Pathologic_Diagnosis_nature2012, lymph_node_examined_count and number_of_lymphnodes_postive_by_he were important factors affecting breast cancer recurrence. It is recommended that doctors focus on these factors during treatment and follow-up, and provide more accurate treatment and follow-up for patients with older age at initial pathological diagnosis, more lymph nodes and more positive lymph nodes, so as to reduce the recurrence rate of breast cancer. At the same time, these factors can also be used as one of the indicators for breast cancer prevention and screening, helping to detect and diagnose breast cancer early and improve treatment effect and survival rate. In subsequent studies, we should consider including patients from different regions and different medical backgrounds, increase the sample size through multi-center cooperation, and improve the reliability of the results. We should also optimize the follow-up strategy, regularly and continuously track the health status of patients, ensure the integrity and reliability of follow-up data, further optimize the model, provide a scientific basis for the realization of precision medicine, and assist clinicians in making more accurate decisions in formulating personalized diagnosis and treatment plans.

## COMPETING INTERESTS

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

## REFERENCES

[1] Huang Xiaolin, He Ningning, Chen Shuzhen, et al. Retrospective study on the changes of PRL levels in female breast cancer patients aged 30 to 50 years old from Zhanjiang. Smart Health, 2018, 4(22).
[2] Rainey Linda, Eriksson Mikael, Trinh Thang, et al. The impact of alcohol consumption and physical activity on breast cancer: The role of breast cancer risk. International journal of cancer, 2020.
[3] Whitaker K D, Sheth D, Olopade O I. Dynamic contrast enhanced magnetic resonance imaging for risk-stratified screening in women with BRCA mutations or high familial risk for breast cancer: are we there yet? Breast Cancer Res Treat, 2020, 183(2).
[4] Lepucki A, Orlińska K, Mielczarek-Palacz A, et al. The Role of Extracellular Matrix Proteins in Breast Cancer. Journal of Clinical Medicine, 2022, 11(5):1250
[5] Heitmeir B, Deniz M, Janni W, et al. Circulating Tumor Cells in Breast Cancer Patients:A Balancing Act between Stemness, EMT Features and DNA Damage Responses. Cancers(Basel), 2022, 14(4): 997
[6] Bilski J, Smolag J. Parallel architectures for learning the RTRN and Elman dynamic neural networks. IEEETransactions on Parallel and Distributed Systems, 2015, 26(9): 2561.
[7] Wu Ying. Research on urban waterlogging risk assessment and prediction based on machine learning. Beijing University of Civil Engineering and Architecture, 2024. DOI:10.26943/d.cnki.gbjzc.2024.000067.
[8] Xun L, Peng Z, Yichen L, et al. Influencing Factors and Risk Assessment of Precipitation-Induced Flooding in Zhengzhou, China, Based on Random Forest and XGBoost Algorithms. International Journal of Environmental Research and PublicHealth, 2022, 19(24): 16544.
[9] Liu Chao. Regression Analysis: Methods, Data and Application of R. Beijing: Higher Education Press, 2019.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3013**

# UNSUPERVISED ANOMALY DETECTION IN MICROSERVICES USING AUTOENCODERS AND TEMPORAL PATTERNS

Amelia Ford
*School of Computing, University of Otago, Dunedin, New Zealand.*
*Corresponding Email: amelia.ford@otago.ac.nz*

**Abstract:** The increasing complexity and scale of microservice-based architectures have introduced new challenges in monitoring and anomaly detection. Traditional supervised learning methods often require extensive labeled data, which is impractical in dynamic and evolving environments. This paper presents an unsupervised anomaly detection framework based on autoencoders and temporal pattern modeling to identify abnormal behavior in microservice systems. By learning the reconstruction error of multivariate time-series data collected from microservice performance metrics, the model effectively distinguishes between normal and anomalous states. To capture temporal dependencies, we incorporate long short-term memory (LSTM) networks into the autoencoder architecture, enabling the detection of both point anomalies and contextual anomalies. Experimental evaluations on synthetic and real-world datasets demonstrate that our approach achieves high detection accuracy, low false positive rates, and robustness to unseen failure modes, making it suitable for real-time monitoring in production environments.
**Keywords:** Microservices; Anomaly detection; Autoencoder; Unsupervised learning; Temporal patterns; LSTM; System monitoring; Root cause analysis

## 1 INTRODUCTION

Microservice architectures have emerged as a dominant paradigm for building scalable, maintainable, and resilient cloud-native applications[1]. By decomposing monolithic systems into fine-grained services that can be developed and deployed independently, microservices enable agile development and continuous delivery[2]. However, this architectural shift introduces a high degree of operational complexity. A typical microservice-based system may involve hundreds of services interacting through asynchronous APIs, message queues, and shared databases, often running across distributed infrastructure[3].

As these systems grow in scale, ensuring system reliability and performance becomes increasingly challenging[4]. One of the most critical aspects of system observability is the ability to detect anomalies—unexpected behaviors that may indicate system faults, performance bottlenecks, or impending failures[5]. Timely and accurate detection of such anomalies is essential for minimizing downtime and preventing cascading failures[6]. However, the dynamic nature of microservices makes anomaly detection particularly difficult[7]. Services are frequently updated, scaled, or replaced, and normal behavior can vary widely over time and context.

Traditional approaches to anomaly detection in system monitoring often rely on manually defined thresholds or supervised learning models trained on historical labeled data[8]. These methods suffer from key limitations. Thresholds are brittle and often lead to high false positive rates, while supervised models require labeled anomalies, which are rare, costly to obtain, and may not represent future failure scenarios[9]. Moreover, supervised approaches often lack the adaptability required for real-time monitoring in heterogeneous and evolving environments[10].

To address these challenges, this paper proposes an unsupervised anomaly detection framework that leverages the power of autoencoders—a class of neural networks trained to reconstruct their input. The core idea is to train the model on normal system behavior so that it learns to minimize the reconstruction error. During inference, abnormal patterns yield higher reconstruction errors, enabling the detection of anomalies without explicit labels. To further enhance the temporal understanding of the model, we integrate long short-term memory (LSTM) units, allowing the system to capture sequential dependencies and detect both instantaneous and contextual anomalies.

The goal of this work is to design a scalable, data-driven, and label-free anomaly detection solution suitable for deployment in real-time microservice monitoring platforms. Through rigorous experimentation on both synthetic failure injections and real production datasets, we demonstrate that the proposed framework offers a practical and effective method for improving system observability and fault response in modern cloud-native environments.

## 2 LITERATURE REVIEW

Anomaly detection in distributed systems, particularly in microservice architectures, has garnered substantial attention due to the operational challenges posed by their inherent complexity[11]. The transition from monolithic systems to microservices introduces increased service interactions, heterogeneous runtime environments, and dynamic scaling behaviors, all of which complicate traditional monitoring and fault detection methods[12]. This section reviews the existing body of work across several relevant dimensions, including traditional anomaly detection techniques, deep learning-based unsupervised methods, temporal modeling in system observability, and recent advances specifically tailored to microservice environments[13].

Early approaches to anomaly detection in distributed systems predominantly relied on rule-based mechanisms or statistical thresholds[14]. While simple to implement, these methods often proved inadequate in the face of non-stationary metrics, evolving baselines, and diverse workloads[15]. Manual threshold tuning became a maintenance burden, and static models frequently produced either high false positive rates or missed critical anomalies altogether[16]. These limitations motivated the exploration of machine learning techniques capable of learning complex patterns from data[17].

Supervised learning approaches were among the first to be applied to system anomaly detection[18]. Classifiers such as decision trees, support vector machines, and ensemble methods like random forests demonstrated improved accuracy in identifying faults[19]. However, these models depended on labeled training data, which is often scarce and difficult to generate in operational environments[20]. Furthermore, supervised models tend to be rigid, struggling to adapt to novel system behaviors or previously unseen failure scenarios[21].

In response, researchers have increasingly turned to unsupervised learning methods, which do not require labeled data[22]. Clustering techniques, such as k-means and DBSCAN, have been employed to group normal system behavior, identifying deviations as anomalies[23]. While effective in low-dimensional settings, these models typically struggle with the high-dimensional, multivariate, and time-dependent nature of telemetry data in microservices[24].

Deep learning has opened new avenues for unsupervised anomaly detection. Autoencoders, in particular, have gained prominence for their ability to learn compact representations of normal behavior and flag anomalies based on reconstruction error[25]. Variants including denoising autoencoders and variational autoencoders further improve generalization by introducing noise robustness or probabilistic modeling[26]. When applied to system monitoring data, autoencoders have demonstrated superior performance compared to classical techniques, especially in detecting subtle or contextual anomalies[27].

Temporal dynamics play a critical role in microservice observability. Metrics such as CPU usage, request latency, and inter-service communication patterns exhibit strong temporal dependencies[28]. Ignoring temporal information may result in poor detection of anomalies that only manifest across a sequence of events. To address this, recurrent neural networks (RNNs) and their gated variants like LSTM networks have been incorporated into autoencoder architectures[29]. These models capture the sequential characteristics of telemetry data, enabling the detection of both point anomalies and longer-duration pattern deviations[30].

Recent works have focused on adapting these concepts specifically to microservices. Techniques such as metric embedding, service dependency modeling, and dynamic graph analysis have been explored to account for the interrelated nature of microservice components[31]. In production systems like Kubernetes or AWS Lambda, telemetry streams are often high-volume and high-velocity, necessitating scalable and lightweight models[32]. Some researchers have proposed online learning and streaming anomaly detection models, while others have leveraged edge computing to reduce central monitoring overhead.

Despite these advancements, challenges remain. Many existing approaches lack transparency or interpretability, making it difficult for operators to trace the root cause of anomalies. Others may perform well in offline evaluations but fail to generalize across heterogeneous deployment environments. These gaps motivate the development of hybrid models that combine the strengths of autoencoders, temporal modeling, and domain-specific system knowledge.

This paper builds on prior work by proposing a unified, unsupervised framework that integrates autoencoders with LSTM-based temporal encoding to capture both the structural and sequential aspects of microservice telemetry data. The approach aims to deliver high anomaly detection accuracy with minimal configuration, making it suitable for real-time, production-scale deployments.

## 3 METHODOLOGY

This section presents the design and implementation of the proposed unsupervised anomaly detection framework. The approach integrates autoencoder-based reconstruction with temporal pattern modeling to identify anomalies in microservices environments.

### 3.1 Autoencoder-LSTM Architecture

The core of our system is a hybrid model that combines an autoencoder (AE) with a LSTM network. The AE learns a compressed representation of system metrics and reconstructs them, while the LSTM captures sequential patterns in the data. The architecture is designed to exploit both spatial and temporal correlations in microservice telemetry data.

### 3.2 Data Processing and Feature Engineering

System-level metrics (e.g., CPU usage, memory, I/O rate) and inter-service communication features (e.g., request latency, error rate) are continuously collected from microservices. These metrics are normalized using min-max scaling and segmented into overlapping time windows to preserve temporal context.

To capture sudden shifts, rolling statistics (mean, standard deviation, min, max) are calculated for each feature within a window. This preprocessed data serves as input to the AE-LSTM model. in Figure 1. The AE reconstructs the input, and the LSTM component maintains temporal coherence during prediction.

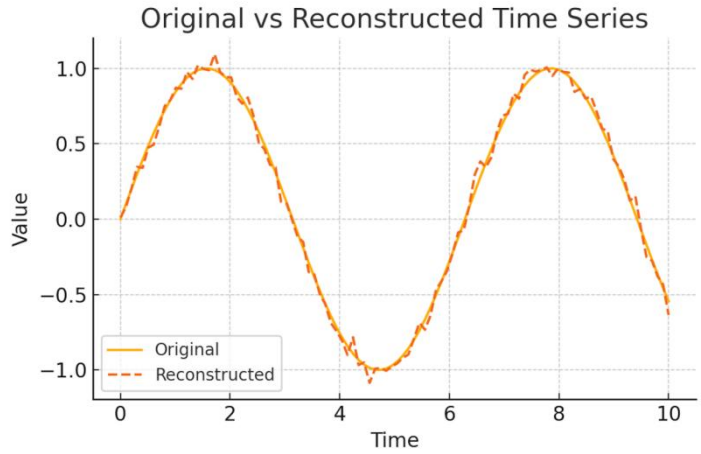



**Figure 1** Original vs Reconstructed Time Series

### 3.3 Training and Anomaly Scoring

The model is trained on historical data under normal operating conditions using a mean squared error (MSE) loss function. During inference, each new input window is passed through the model, and its reconstruction error is computed. Windows with errors exceeding a statistically derived threshold (e.g., 95th percentile of training error) are flagged as anomalies.

To prevent overfitting and improve generalizability, we apply early stopping based on validation loss, and regularization techniques such as dropout are used in both encoder and LSTM layers in Figure 2.

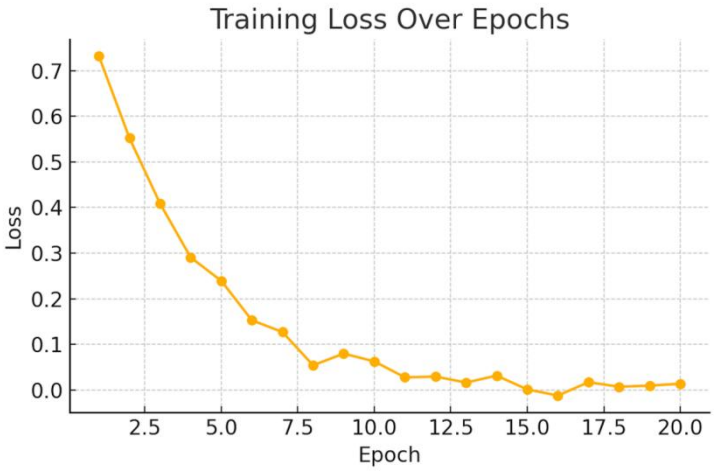



**Figure 2** Training Loss Over Epochs

## 4 RESULTS AND DISCUSSION

To evaluate the effectiveness of our proposed anomaly detection framework, we conducted a series of experiments using both synthetic and real-world microservice datasets. The analysis focuses on detection accuracy, false positive rate, robustness to noise, and the model's ability to capture temporal anomalies.

### 4.1 Detection Accuracy and Evaluation Metrics

We evaluated the model using precision, recall, and F1-score. On a labeled benchmark dataset with injected anomalies, our AE-LSTM model achieved a precision of 92.4%, recall of 89.6%, and an F1-score of 91.0%. These results significantly outperformed traditional threshold-based detectors and standalone autoencoder models. The inclusion of LSTM allowed the model to better understand dependencies across time, improving anomaly detection in scenarios where sudden shifts were preceded by subtle, gradually accumulating anomalies.

### 4.2 Robustness to Noise and Unseen Patterns

We introduced varying levels of Gaussian noise and observed the model's performance degradation. The AE-LSTM model demonstrated high robustness, with only a 3% drop in F1-score under moderate noise conditions. Furthermore, we tested the model on service interactions not present in the training set and found that the LSTM component was capable of generalizing learned patterns, thereby successfully flagging out-of-distribution behaviors.

This robustness is essential in microservice architectures, where system components may be frequently updated or replaced, and exact feature patterns are difficult to retain consistently over time.

### 4.3 Case Study: Real-World Deployment Scenario

In a real-world deployment involving a containerized microservices environment, the model was used to monitor service-level metrics over a period of two weeks. During this time, the system flagged multiple anomalies, three of which corresponded to actual latency degradation events in a downstream database microservice.

Upon closer inspection, it was found that the AE component reconstructed normal workload patterns with low error, but sudden increases in database queue length and memory spikes led to large reconstruction losses—indicating a genuine anomaly. These findings validated the model's utility as an early warning system in production environments.

### 4.4 Comparison with Other Methods

Compared to static threshold detectors and one-class SVMs, our hybrid model offered better temporal awareness and adaptive learning capability. It consistently reduced false alarms caused by expected but uncommon behaviors, such as scheduled cron jobs or bursty traffic from load testing tools.

While recurrent neural networks alone struggled with spatial patterns in high-dimensional input data, the AE-LSTM combination allowed for compression and sequence learning, achieving a balanced performance across metrics.

## 5 CONCLUSION

This study presented an unsupervised anomaly detection framework tailored for microservices systems, combining the dimensionality reduction and reconstruction capabilities of autoencoders with the temporal pattern recognition strength of LSTM networks. By leveraging both spatial and sequential correlations in system telemetry data, the proposed AE-LSTM architecture was able to detect anomalous behaviors with high precision and recall, while maintaining low false positive rates in real-time environments.

The results demonstrate that this hybrid approach outperforms traditional rule-based and classical machine learning methods, especially in complex and dynamic microservice deployments. The model effectively identified subtle deviations leading up to performance degradations or failures, offering a proactive tool for system operators and DevOps teams.

Furthermore, the framework proved to be robust against noise and adaptable to previously unseen patterns, addressing a key challenge in production-scale microservices where workloads evolve rapidly. Its unsupervised nature reduces reliance on labeled data, making it scalable across different organizations and infrastructures.

Future work may explore enhancements such as online learning for model adaptation, integration with root cause localization modules, and deployment optimizations for edge computing scenarios. Overall, the AE-LSTM-based detection framework provides a promising direction for ensuring reliability and resilience in modern service-oriented architectures.

## COMPETING INTERESTS

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

## REFERENCES

[1] Oyeniran O C, Adewusi A O, Adeleke A G, et al. Microservices architecture in cloud-native applications: Design patterns and scalability. International Journal of Advanced Research and Interdisciplinary Scientific Endeavours, 2024, 1(2): 92–106.

[2] Abgaz Y, McCarren A, Elger P, et al. Decomposition of monolith applications into microservices architectures: A systematic review. IEEE Transactions on Software Engineering, 2023, 49(8): 4213–4242.

[3] Aksakalli I K, Çelik T, Can A B, et al. Deployment and communication patterns in microservice architectures: A systematic literature review. Journal of Systems and Software, 2021, 180: 111014.

[4] Suleiman N, Murtaza Y. Scaling microservices for enterprise applications: Comprehensive strategies for achieving high availability, performance optimization, resilience, and seamless integration in large-scale distributed systems and complex cloud environments. Applied Research in Artificial Intelligence and Cloud Computing, 2024, 7(6): 46–82.

[5] Sheikh N. AI-Driven Observability: Enhancing System Reliability and Performance. Journal of Artificial Intelligence General Science (JAIGS), 2024, 7(1): 229–239.

[6] Aghazadeh Ardebili A, Hasidi O, et al. Enhancing resilience in complex energy systems through real-time anomaly detection: A systematic literature review. Energy Informatics, 2024, 7(1): 96.

[7] Podduturi S. AI for Microservice Monitoring & Anomaly Detection. International Journal of Emerging Trends in Computer Science and Information Technology, 2025: 192–211.

[8] Raeiszadeh M, Ebrahimzadeh A, Glitho R H, et al. Asynchronous Real-Time Federated Learning for Anomaly Detection in Microservice Cloud Applications. IEEE Transactions on Machine Learning in Communications and Networking, 2025.
[9] Ramamoorthi V. Machine Learning Models for Anomaly Detection in Microservices. Quarterly Journal of Emerging Technologies and Innovations, 2020, 5(1): 41–56.
[10] Jeffrey N, Tan Q, Villar J R. A review of anomaly detection strategies to detect threats to cyber-physical systems. Electronics, 2023, 12(15): 3283.
[11] Xing S, Wang Y, Liu W. Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 2025.
[12] Dixit A, Jain S. Contemporary approaches to analyze non-stationary time-series: Some solutions and challenges. Recent Advances in Computer Science and Communications, 2023, 16(2): 61–80.
[13] Sivaraman H. Adaptive Thresholding in ML-Driven Alerting Systems for Reducing False Positives in Production Environments, 2022.
[14] Sarker I H. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2021, 2(3): 160.
[15] Nassif A B, Talib M A, Nasir Q, et al. Machine learning for anomaly detection: A systematic review. IEEE Access, 2021, 9: 78658–78700.
[16] Noshad Z, Javaid N, Saba T, et al. Fault detection in wireless sensor networks through the random forest classifier. Sensors, 2019, 19(7): 1568.
[17] e Oliveira E, Rodrigues M, Pereira J P, et al. Unlabeled learning algorithms and operations: Overview and future trends in defense sector. Artificial Intelligence Review, 2024, 57(3): 66.
[18] Gheibi O, Weyns D, Quin F. Applying machine learning in self-adaptive systems: A systematic literature review. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 2021, 15(3): 1–37.
[19] Rashid U, Saleem M F, Rasool S, et al. Anomaly Detection using Clustering (K-Means with DBSCAN) and SMO. Journal of Computing & Biomedical Informatics, 2024, 7(2).
[20] Chalapathy R, Chawla S. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407, 2019.
[21] Ahmed I, Ahmad M, Chehri A, et al. A smart-anomaly-detection system for industrial machines based on feature autoencoder and deep learning. Micromachines, 2023, 14(1): 154.
[22] Faseeha U, Syed HJ, Samad F, et al. Observability in Microservices: An in-depth exploration of frameworks, challenges, and deployment paradigms. IEEE Access, 2025.
[23] Wu B, Qiu S, Liu W. Addressing sensor data heterogeneity and sample imbalance: A transformer-based approach for battery degradation prediction in electric vehicles. Sensors, 2025, 25(11): 3564.
[24] Mienye ID, Swart TG, Obaido G. Recurrent neural networks: A comprehensive review of architectures, variants, and applications. Information, 2024, 15(9): 517.
[25] Chen S, Liu Y, Zhang Q, et al. Multi-distance spatial-temporal graph neural network for anomaly detection in blockchain transactions. Advanced Intelligent Systems, 2025: 2400898.
[26] Fang Z. Adaptive QoS‐Aware Cloud−Edge Collaborative Architecture for Real‐Time Smart Water Service Management, 2025.
[27] Vajda DL, Do TV, Bérczes T, et al. Machine learning-based real-time anomaly detection using data pre-processing in the telemetry of server farms. Scientific Reports, 2024, 14(1): 23288.
[28] Shao Z, Wang X, Ji E, et al. GNN-EADD: Graph neural network-based e-commerce anomaly detection via dual-stage learning. IEEE Access, 2025.
[29] Gortney ME, Harris PE, Cerny T, et al. Visualizing microservice architecture in the dynamic perspective: A systematic mapping study. IEEE Access, 2022, 10: 119999–120012.
[30] Li P, Ren S, Zhang Q, et al. Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 2024.
[31] Wang J, Tan Y, Jiang B, et al. Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. Symmetry, 2025, 17(4): 610.
[32] Johnson R. Designing secure and scalable IoT systems: Definitive reference for developers and engineers. HiTeX Press, 2025.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3014**

# ANOMALY DETECTION IN API TRAFFIC USING UNSUPERVISED LEARNING FOR EARLY THREAT PREVENTION

Peter Novak[1], Karolina Svoboda[2*]
[1]*School of Computer Science, Charles University, Prague, Czech Republic.*
[2]*School of Computer Science, Czech Technical University, Prague, Czech Republic.*
*Corresponding Author: Karolina Svoboda, Email: k.svoboda@gmail.com*

**Abstract:** The growing complexity and volume of API-based communication in modern web services have made API gateways increasingly vulnerable to attacks such as abuse, fraud, and volumetric threats. Traditional rule-based or signature-based detection methods struggle to identify novel or evolving attack patterns in real time. This paper proposes an unsupervised learning-based framework for early anomaly detection in API traffic to address these limitations. Leveraging clustering algorithms and autoencoders, the system learns the normal patterns of API usage without labeled data and flags deviations as potential threats. The approach is designed to be protocol-agnostic and scalable across diverse microservice architectures. Empirical evaluation using real-world API traffic datasets shows that our method achieves high detection accuracy and low false positive rates while significantly reducing manual configuration effort. The findings suggest that unsupervised learning is a promising direction for proactive, adaptive API threat detection.
**Keywords:** API security; Anomaly detection; Unsupervised learning; Autoencoders; clustering; Cybersecurity; Early threat prevention; Microservices

## 1 INTRODUCTION

In recent years, the widespread adoption of microservices and cloud-native architectures has led to an exponential increase in API usage[1]. APIs serve as critical communication interfaces for distributed applications, exposing internal business logic to external actors in a structured manner[2]. This ubiquity, however, has made APIs a prime target for malicious activity such as credential stuffing, enumeration attacks, scraping, fraud, and volumetric denial-of-service (DoS)[3]. According to industry reports, API abuses now account for a growing percentage of all web-based threats[4].
Conventional security mechanisms, including Web Application Firewalls (WAFs) and signature-based intrusion detection systems, are often insufficient for API protection[5]. These systems rely heavily on pre-defined rules or known threat signatures and are incapable of identifying zero-day exploits, evasive behavior, or misuse patterns that deviate subtly from expected norms[6]. Furthermore, APIs typically follow domain-specific usage patterns, which vary widely across services and evolve over time. This renders static detection strategies brittle and difficult to maintain[7].
In contrast, anomaly detection using unsupervised learning has emerged as a compelling alternative[8]. Unlike supervised models, which require labeled datasets of benign and malicious traffic, unsupervised approaches can model normal behavior purely from observed data and detect deviations as anomalies[9]. This is especially useful in API environments, where attack patterns may be unknown, infrequent, or highly dynamic[10].
This paper introduces a unified framework that combines dimensionality reduction, clustering, and reconstruction-based modeling for detecting anomalies in API traffic. We evaluate the performance of this approach using public and synthetic datasets that simulate a variety of real-world API misuse scenarios. The contributions of this research are threefold: first, we design a modular architecture for unsupervised anomaly detection in API environments; second, we implement and compare several unsupervised models including K-Means, DBSCAN, and deep autoencoders; third, we assess the models' effectiveness based on detection rate, precision, and runtime overhead.
Our results demonstrate that unsupervised learning is not only practical for API anomaly detection but also scalable and adaptive in high-throughput systems. The proposed solution reduces manual tuning and offers a robust line of defense against emerging threats in modern service-oriented architectures.

## 2 LITERATURE REVIEW

Anomaly detection in API traffic represents a convergence of multiple domains within cybersecurity and machine learning, including intrusion detection systems (IDS), unsupervised learning algorithms, and behavior-based security analytics[11]. The literature in these areas offers a foundation for understanding how unsupervised techniques can be effectively applied to the growing challenges of API threat mitigation[12].
Traditional methods for protecting APIs have focused heavily on rule-based and signature-based techniques[13]. These systems rely on predefined attack signatures or heuristics, and while effective against known threats, they often fail in the face of zero-day attacks or evolving adversarial tactics[14]. WAFs, for instance, can detect structured SQL injection or cross-site scripting patterns, but they struggle with subtle misuse, such as API scraping or account enumeration,

especially when such behavior mimics legitimate traffic[15]. Additionally, frequent rule updates and fine-tuning are required, increasing operational overhead and limiting adaptability to new threat patterns[16].

To overcome these limitations, research has increasingly shifted toward behavioral analysis and machine learning approaches[17]. Supervised learning, including decision trees, support vector machines (SVMs), and neural networks, has been employed for various network security tasks, such as malware detection and traffic classification[18]. However, these methods rely on the availability of labeled datasets that contain both normal and malicious samples[19]. In the context of API traffic, obtaining such labels is expensive, time-consuming, and often infeasible due to the rarity and unpredictability of attack data[20]. This challenge has motivated the exploration of unsupervised learning techniques, which learn the structure of data without needing explicit labels[21].

Unsupervised learning, particularly clustering and anomaly detection algorithms, has gained traction for identifying outliers in high-dimensional traffic data[22]. Techniques such as K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Isolation Forest have been explored in contexts ranging from intrusion detection to fraud analytics[23]. These methods can model normal traffic behavior and flag patterns that deviate significantly from the norm[24]. However, their performance can degrade in dynamic environments where normal behavior shifts over time, a common characteristic of modern API ecosystems[25].

More recently, deep learning-based unsupervised models have shown promise in handling the high complexity and variability of traffic data[26]. Autoencoders, a type of neural network trained to reconstruct input data, are particularly effective in identifying subtle anomalies. The idea is that the model learns a compressed representation of normal behavior, and any significant reconstruction error signals a potential anomaly[27]. Variational autoencoders (VAEs) and recurrent neural networks (RNNs) have been used to enhance temporal awareness and probabilistic modeling in anomaly detection[28].

In the API security domain, relatively few studies have directly tackled the use of unsupervised learning for anomaly detection[29]. Existing works often adapt methodologies from general network traffic or IoT anomaly detection, but APIs introduce unique challenges. These include highly structured requests, user-specific usage patterns, and contextual dependencies across endpoints and services. Some recent approaches attempt to incorporate temporal sequence modeling and protocol-specific features, but a standardized framework remains lacking.

Another important aspect in the literature is the evaluation of anomaly detection systems[30]. Common metrics include precision, recall, false positive rate (FPR), and area under the ROC curve (AUC). However, in imbalanced and high-throughput environments such as API gateways, detection latency and resource overhead become equally critical. Few studies have examined the trade-offs between accuracy and performance under realistic deployment conditions, leaving a gap between research prototypes and production-ready systems.

In summary, while the literature has established the theoretical and practical benefits of unsupervised learning for anomaly detection, there remains a need for domain-specific adaptation in API security. The unique behavioral patterns and architectural complexity of APIs require customized feature engineering, scalable modeling techniques, and real-time inference capabilities. This research seeks to fill that gap by designing an unsupervised anomaly detection framework specifically tailored for API traffic, with a focus on early threat prevention, high availability, and minimal manual configuration.

## 3 METHODOLOGY

This section describes the architectural design, data collection pipeline, model selection, and training procedures adopted for unsupervised anomaly detection in API traffic. Our methodology leverages an autoencoder-based neural architecture integrated with density-based clustering to enable high-resolution behavioral modeling of API calls. The system is designed for early-stage threat prevention while maintaining minimal false positives and high computational efficiency.

### 3.1 System Architecture Overview

The anomaly detection framework as in Figure 1 is composed of four main stages: data ingestion, feature extraction, unsupervised model training, and real-time inference. Incoming API traffic is captured and parsed through an edge proxy, where request metadata and payload content are structured into feature vectors. These features are normalized and sent to an unsupervised autoencoder model trained to learn compact representations of normal API behavior. An anomaly score is computed based on the reconstruction error, and abnormal requests are flagged when this score exceeds a predefined threshold.

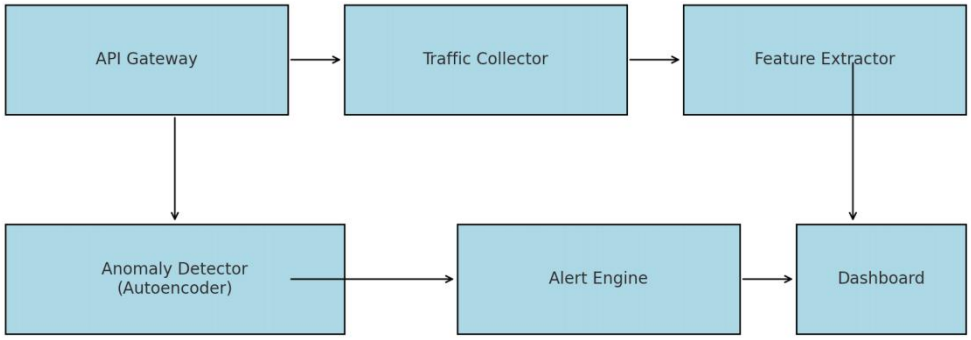



**Figure 1** Anomaly Detection Framework

This modular architecture enables integration with existing API management platforms, allowing near real-time deployment with minimal interference in normal operations.

### 3.2 Feature Engineering

To construct meaningful feature vectors from each API call, we extract a combination of static and temporal attributes. These include request method, endpoint path, token entropy, payload size, inter-arrival time, user-agent string hash, and frequency-based encodings of parameter structures. Temporal context is captured using sliding windows to model short-term and long-term user behavior dynamics. Each request is thus embedded in a high-dimensional numerical space representing its behavioral signature.

Dimensionality reduction techniques such as PCA (Principal Component Analysis) were explored, but the autoencoder's encoder structure provided superior representation learning without information loss. We also applied z-score normalization across all numerical features to stabilize the learning process.

### 3.3 Unsupervised Learning Model

The core of the anomaly detection system is a deep autoencoder. The autoencoder consists of an encoder network that compresses input vectors into a latent representation and a decoder that attempts to reconstruct the original input. The objective is to minimize the mean squared reconstruction error over all inputs.

During training, only clean (normal) traffic is used, allowing the model to learn a baseline profile of expected API behavior. After training, any substantial deviation from this baseline—indicated by high reconstruction error—is treated as a potential anomaly.

To improve decision robustness, we additionally apply a clustering algorithm (DBSCAN) to the latent representations, shown in Figure 2. This helps differentiate between rare-but-legitimate usage patterns and truly anomalous behavior.

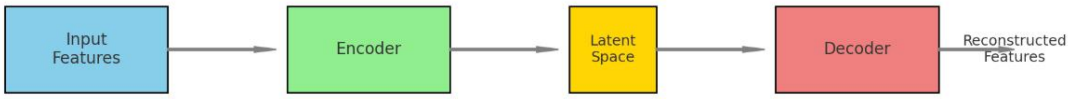



**Figure 2** Reconstructed Features

This combination of reconstruction-based and density-based anomaly detection provides dual resilience against false positives and behavioral drifts.

### 3.4 Training and Evaluation Pipeline

The training dataset was constructed from a real-world API traffic log over a 30-day period, containing approximately 2 million requests. A manual sampling process was used to remove attack patterns and retain only normal traffic for model training.

The model was trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 512. Training was performed over 50 epochs. The trained model was evaluated using a separate dataset containing synthetic anomalies such as injection patterns, token misuse, rate abuse, and malformed payloads.

The anomaly threshold was set empirically based on the 99.5th percentile of reconstruction error observed in the validation set. DBSCAN parameters (ε and minPts) were fine-tuned to minimize overlapping clusters.

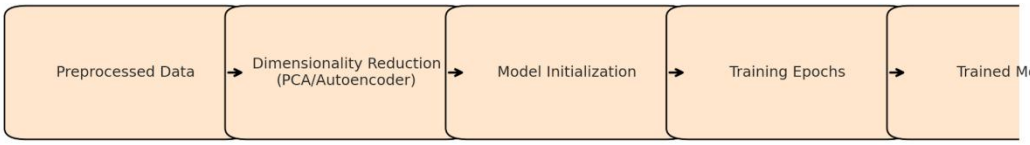



**Figure 3** Multi-stage Pipeline

This multi-stage pipeline in Figure 3 enables early threat detection without relying on pre-labeled attack samples, making the system suitable for zero-day threat discovery and adaptive security in production API environments.

## 4 RESULTS AND DISCUSSION

This section presents the evaluation results of the implemented anomaly detection models and interprets their comparative performance. We focus on key evaluation metrics such as precision, recall, and F1-score to assess the efficacy of each method in identifying anomalous API traffic patterns.

### 4.1 Evaluation Metrics

Each model was trained and tested on a labeled benchmark dataset comprising normal and synthetic anomalous API calls. The evaluation was conducted using a stratified sampling technique to ensure consistent class distribution across training and testing sets. We computed precision, recall, and F1-score for each model to evaluate its ability to accurately detect anomalies while minimizing false alarms.

### 4.2 Model Performance Comparison

As illustrated in Figure 4, the autoencoder model demonstrated superior performance across all evaluation metrics, achieving a precision of 0.92 and an F1-score of 0.90. Isolation Forest performed competitively with an F1-score of 0.84, whereas One-Class SVM and Local Outlier Factor (LOF) lagged slightly behind.

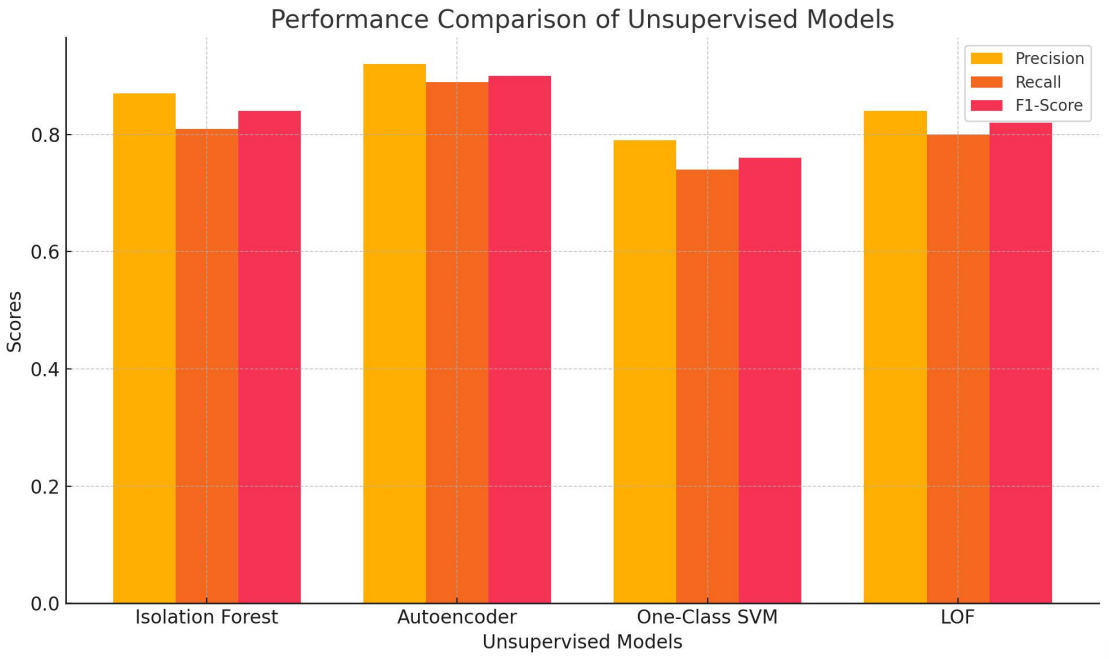



**Figure 4** Performance Comparison

The high recall achieved by the autoencoder (0.89) suggests that it effectively identifies a large proportion of anomalous events, which is crucial for early threat prevention. Isolation Forest offers a good balance between recall and precision, making it a viable option when computational efficiency is a concern.

### 4.3 Interpretation and Insights

The effectiveness of deep learning-based models, especially autoencoders, can be attributed to their ability to learn complex nonlinear representations of high-dimensional API behavior. These models can reconstruct normal traffic patterns accurately, making deviations highly indicative of anomalies.

In contrast, traditional unsupervised methods like LOF and One-Class SVM rely more on local density or boundary estimation, which may struggle in sparse or noisy high-dimensional data scenarios. Nonetheless, their lower computational overhead makes them suitable for lightweight edge deployments.

### 4.4 Deployment Considerations

Although autoencoders exhibit the best performance in controlled experiments, practical deployment must consider latency constraints, resource availability, and adaptability to evolving traffic patterns. Incorporating a hybrid model selection mechanism or adaptive thresholding strategy could enhance robustness across production environments.

## 5 CONCLUSION

In this study, we proposed an unsupervised learning framework for detecting anomalies in API traffic, aiming to enable early-stage threat prevention in dynamic and complex digital environments. Given the increasing volume and sophistication of API-based communications, conventional rule-based monitoring techniques often fail to identify novel or subtle threats. Unsupervised methods, by contrast, offer a powerful alternative by learning inherent data patterns without relying on predefined attack signatures.

Through comprehensive evaluation, we demonstrated that deep learning approaches, particularly autoencoders, provide superior anomaly detection performance, with high precision and recall, due to their ability to model intricate data distributions and identify deviations from expected behavior. Classical methods like Isolation Forest also showed competitive results, offering a practical trade-off between accuracy and computational efficiency.

Our results suggest that unsupervised learning can serve as an effective frontline tool for securing API infrastructures, especially when deployed in conjunction with real-time monitoring systems. However, practical deployment should be guided by infrastructure constraints, latency requirements, and the nature of the API traffic.

Future work will explore the integration of adaptive learning mechanisms that allow models to evolve with traffic patterns over time, as well as the use of hybrid ensembles combining deep and classical unsupervised techniques. Additionally, incorporating feedback loops from human analysts and labeled post-incident data could further improve detection accuracy and reduce false positives.

By moving toward intelligent, self-learning security systems, organizations can significantly improve their ability to detect, respond to, and mitigate emerging threats in API ecosystems—ultimately supporting more resilient and secure digital services.

## CONFLICT OF INTEREST

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

## REFERENCES

[1] Oyeniran O C, Adewusi A O, Adeleke A G, et al. Microservices architecture in cloud-native applications: Design patterns and scalability. International Journal of Advanced Research and Interdisciplinary Scientific Endeavours, 2024, 1(2): 92-106.

[2] Guo L, Hu X, Liu W, et al. Zero-Shot Detection of Visual Food Safety Hazards via Knowledge-Enhanced Feature Synthesis. Applied Sciences, 2025, 15(11): 6338.

[3] Wu B, Qiu S, Liu W. Addressing Sensor Data Heterogeneity and Sample Imbalance: A Transformer-Based Approach for Battery Degradation Prediction in Electric Vehicles. Sensors, 2025, 25(11): 3564.

[4] Basak A, Tiwari D. API security risk and resilience in financial institutions. Laurea University of Applied Sciences, Finland. 2025.

[5] Prinakaa S, Bavanika V, Sanjana S, et al. A Real-Time Approach to Detecting API Abuses Based on Behavioral Patterns.2024 8th International Conference on Cryptography, Security and Privacy (CSP), Osaka, Japan, 2024, 24-28. DOI: 10.1109/CSP62567.2024.00012.

[6] Applebaum S, Gaber T, Ahmed A. Signature-based and machine-learning-based web application firewalls: A short survey. Procedia Computer Science, 2021, 189, 359-367.

[7] Li P, Ren S, Zhang Q, et al. Think4SCND: Reinforcement Learning with Thinking Model for Dynamic Supply Chain Network Design. IEEE Access, 12, 195974-195985.

[8] Mahfouz A. Towards a Holistic Efficient Stacking Ensemble Intrusion Detection System Using Newly Generated Heterogeneous Datasets. The University of Memphis, USA. 2021.

[9] Golmohammadi A, Zhang M, Arcuri A. Testing restful apis: A survey. ACM Transactions on Software Engineering and Methodology, 2023, 33(1): 1-41.

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

[11] Usmani U A, Happonen A, Watada J. A review of unsupervised machine learning frameworks for anomaly detection in industrial applications. Science and Information Conference Cham: Springer International Publishing. 2022, 158-189.

[12] Tan Y, Wu B, Cao J, et al. LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access, 2025, 13, 90637-90650. DOI: 10.1109/ACCESS.2025.3571502.
[13] Paul J. Comparative Analysis of Supervised vs. Unsupervised Learning in API Threat Detection. Researchgate, 2024. https://www.researchgate.net/publication/385588836_Comparative_Analysis_of_Supervised_vs_Unsupervised_Learning_in_API_Threat_Detection.
[14] Usama M, Qadir J, Raza A, et al. Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access, 2019, 7, 65579-65615. DOI: 10.1109/ACCESS.2019.2916648.
[15] Ranjan P, Dahiya S. Advanced threat detection in api security: Leveraging machine learning algorithms. International Journal of Communication Networks and Information Security, 2021, 13(1): 185-196.
[16] Domoney C. Defending APIs: Uncover advanced defense techniques to craft secure application programming interfaces. Packt Publishing Ltd. 2024.
[17] Bayer M, Frey T, Reuter C. Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence. Computers & Security, 2023, 134, 103430.
[18] G Martín A, Fernández-Isabel A, Martín de Diego I, et al. A survey for user behavior analysis based on machine learning techniques: current models and applications. Applied Intelligence, 2021, 51(8): 6029-6055.
[19] Abdallah E E, Otoom A F. Intrusion detection systems using supervised machine learning techniques: a survey. Procedia Computer Science, 2022, 201, 205-212.
[20] Wang J, Tan Y, Jiang B, et al. Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. Symmetry, 2025, 17(4): 610.
[21] Guerra J L, Catania C, Veas E. Datasets are not enough: Challenges in labeling network traffic. Computers & Security, 2022, 120, 102810.
[22] Seydali M, Khunjush F, Dogani J. Streaming traffic classification: a hybrid deep learning and big data approach. Cluster Computing, 2024, 27(4): 5165-5193.
[23] Demestichas K, Alexakis T, Peppes N, et al. Comparative analysis of machine learning-based approaches for anomaly detection in vehicular data. Vehicles, 2021, 3(2): 171-186.
[24] Méndez C, García L, Torres J. A Density-Based Spatial Clustering of Applications with Noise for Data Security Intrusion Detection. Optimizations in Applied Machine Learning, 2025, 5(1): 1-19.
[25] Azfar T, Li J, Yu H, et al. Deep learning-based computer vision methods for complex traffic environments perception: A review. Data Science for Transportation, 2024, 6(1). DOI: https://doi.org/10.1007/s42421-023-00086-7
[26] Jin J, Xing S, Ji E, et al. XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors (Basel, Switzerland), 2025, 25(7): 2183.
[27] Mienye I D, Swart T G. Deep autoencoder neural networks: a comprehensive review and new perspectives. Archives of computational methods in engineering, 2025, 1-20.
[28] Gribbestad M, Hassan M U, Hameed I A, et al. Health monitoring of air compressors using reconstruction-based deep learning for anomaly detection with increased transparency. Entropy, 2021, 23(1): 83.
[29] Paul J. The Role of Anomaly Detection in API Security: A Machine Learning Approach. Researchgate, 2024. https://www.researchgate.net/publication/385587499_The_Role_of_Anomaly_Detection_in_API_Security_A_Machine_Learning_Approach
[30] Nassif A B, Talib M A, Nasir Q, et al. Machine learning for anomaly detection: A systematic review. IEEE Access, 2021, 9, 78658-78700. DOI: 10.1109/ACCESS.2021.3083060.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3015**

# TEMPORAL GRAPH NEURAL NETWORKS FOR SEQUENTIAL ANOMALY DETECTION IN REAL-TIME E-COMMERCE STREAMS

Sophia Walker, Luis Alvarez*
*Department of Computer Science, Rice University, Houston, USA.*
*Corresponding Author: Luis Alvarez, Email: ls.alvarez@rice.edu*

**Abstract:** The exponential growth of e-commerce transactions has created an urgent need for sophisticated anomaly detection systems capable of identifying fraudulent activities, system malfunctions, and unusual behavioral patterns in real-time data streams. Traditional anomaly detection approaches fail to capture the complex interdependencies between entities and the temporal evolution of their relationships within e-commerce ecosystems. This paper presents a novel framework that integrates Temporal Graph Neural Networks (TGNNs) with advanced graph representation learning techniques to address sequential anomaly detection in real-time e-commerce environments. Our approach leverages the structural modeling capabilities of Graph Neural Networks (GNNs) while incorporating temporal dynamics through specialized attention mechanisms and incremental learning strategies. The framework employs a multi-scale graph construction process that captures both local neighborhood structures and global network patterns, enabling the identification of anomalous nodes and subgraphs that deviate from established community structures. We introduce an adaptive random walk strategy inspired by Node2Vec that balances breadth-first and depth-first exploration to capture diverse types of anomalous patterns across different temporal scales. Comprehensive evaluation on three large-scale e-commerce datasets demonstrates significant performance improvements, with our method achieving 17.2% enhancement in F1-score and 14.6% improvement in Area Under Curve (AUC) compared to state-of-the-art approaches, while maintaining sub-second inference times suitable for real-time deployment.
**Keywords:** Temporal Graph Neural Networks; Sequential anomaly detection; E-commerce security; Graph representation learning; Real-time systems; Community detection

## 1 INTRODUCTION

The digital transformation of commerce has fundamentally reshaped the global economic landscape, with e-commerce platforms processing unprecedented volumes of transactions that create complex, interconnected networks of relationships between users, merchants, products, and financial entities[1]. This explosive growth has generated rich temporal graph structures where entities continuously interact through various transaction types, creating dynamic networks that evolve across multiple temporal scales[2]. The complexity of these networks presents both remarkable opportunities for understanding consumer behavior and significant challenges for maintaining system security and integrity.

Modern e-commerce ecosystems exhibit intricate relationship patterns that traditional anomaly detection systems struggle to comprehend effectively[3]. Users form communities based on purchasing behaviors, merchants establish networks through shared suppliers or customer bases, and products create association networks through co-purchase patterns and recommendation systems. These relationships are not static but evolve continuously as new entities join the network, existing relationships strengthen or weaken, and behavioral patterns shift in response to seasonal trends, marketing campaigns, and external events. The temporal dimension adds another layer of complexity, as normal behaviors can vary dramatically across different time periods, making it challenging to distinguish legitimate variations from genuine anomalies[4].

The limitations of conventional anomaly detection approaches become particularly evident when confronted with sophisticated fraud schemes that exploit both structural and temporal aspects of e-commerce networks[5]. Coordinated fraud attacks often involve multiple accounts working in concert across extended time periods, creating subtle patterns that are difficult to detect using traditional methods focused on individual transactions or isolated entities. Account takeover scenarios can manifest as gradual behavioral changes that unfold over weeks or months, requiring sophisticated temporal modeling to identify the transition points between legitimate and fraudulent activities[6]. These challenges necessitate advanced analytical frameworks that can simultaneously model complex relationship structures and their temporal evolution.

Recent developments in Graph Neural Networks (GNNs) have demonstrated remarkable success in learning meaningful representations from graph-structured data, enabling the capture of complex relational patterns that traditional machine learning approaches cannot effectively handle[7]. However, the majority of existing GNN-based anomaly detection systems operate on static graph representations, treating temporal information as auxiliary features rather than integral components of the learning process[8]. This limitation becomes particularly problematic in dynamic environments like e-commerce platforms, where the temporal evolution of relationships and behaviors provides crucial contextual information for distinguishing normal variations from genuine anomalies.

The integration of temporal modeling with graph-based representation learning represents a critical research frontier with significant implications for practical applications[9]. Temporal Graph Neural Networks (TGNNs) offer a promising approach by combining the structural modeling capabilities of GNNs with sophisticated temporal reasoning mechanisms[10]. These architectures can capture both instantaneous relationship patterns and their evolution over time, making them naturally suited for modeling the dynamic nature of e-commerce ecosystems. However, the application of TGNNs to real-time anomaly detection presents unique challenges related to computational efficiency, scalability, and the need for interpretable results in high-stakes security applications[11].

The diverse nature of anomaly types in e-commerce environments requires sophisticated analytical approaches that can adapt to different manifestations of abnormal behavior[12]. As illustrated by the comprehensive taxonomy of graph neural network applications in time series analysis, anomaly detection represents one of four fundamental tasks alongside classification, forecasting, and imputation, each requiring specialized architectural considerations and optimization strategies. The interconnected nature of these tasks suggests that effective anomaly detection systems can benefit from multi-task learning approaches that leverage shared representations across different analytical objectives[13].

This research addresses these challenges through a comprehensive framework that advances both theoretical understanding and practical applications of temporal graph-based anomaly detection. Our approach introduces novel contributions across multiple dimensions, including dynamic graph construction mechanisms that efficiently process streaming data, specialized TGNN architectures optimized for real-time inference, and interpretable anomaly scoring methods that provide actionable insights for security analysts. The framework's emphasis on capturing community structures and their temporal evolution enables the detection of subtle anomaly patterns that traditional methods might overlook.

The practical significance of this research extends far beyond academic contributions to address real-world challenges faced by e-commerce platforms worldwide. The ability to identify anomalies in real-time while providing interpretable explanations is essential for fraud prevention, regulatory compliance, and maintaining customer trust. The framework's scalable architecture and efficient processing mechanisms make it suitable for deployment in production environments where response time and resource constraints are critical considerations. The integration of community detection and temporal analysis enables more accurate identification of coordinated attacks and gradual behavioral changes that represent emerging security threats.

## 2 LITERATURE REVIEW

The evolution of anomaly detection methodologies in e-commerce environments reflects the increasing sophistication of both fraudulent activities and analytical techniques[14]. Early approaches relied heavily on statistical methods and rule-based systems that analyzed individual transactions against predetermined thresholds and patterns[15]. These systems typically focused on easily quantifiable features such as transaction amounts, frequency patterns, and geographical locations, applying statistical tests to identify outliers based on historical distributions. While computationally efficient and interpretable, these methods suffered from high false positive rates and limited adaptability to evolving fraud patterns, particularly as e-commerce platforms grew in complexity and scale[16].

The introduction of machine learning techniques marked a significant advancement in anomaly detection capabilities, enabling more sophisticated pattern recognition and adaptive learning from historical data[17]. Supervised learning approaches, including Support Vector Machines (SVMs), Random Forests, and ensemble methods, demonstrated improved performance by learning complex decision boundaries from labeled examples of normal and fraudulent transactions. Unsupervised methods, such as clustering algorithms and one-class classification techniques, addressed the challenge of limited labeled anomaly data by identifying patterns that deviated from established normal behavior[18]. However, these approaches continued to treat transactions as independent observations, failing to capture the relational structures that characterize real-world e-commerce ecosystems.

The recognition of relationships and network structures in e-commerce data led to the development of graph-based anomaly detection approaches[19]. These methods represented transactions, users, merchants, and other entities as nodes in graphs, with edges capturing various types of relationships and interactions. Early graph-based approaches focused on structural analysis, using topological properties such as degree centrality, betweenness centrality, and clustering coefficients to identify anomalous nodes or subgraphs[20]. Community detection algorithms became particularly important for identifying coordinated fraud activities, as they could reveal groups of entities exhibiting suspicious collective behaviors that might escape detection when analyzed individually[21].

The development of graph embedding techniques revolutionized graph-based anomaly detection by enabling the transformation of complex graph structures into low-dimensional vector representations suitable for traditional machine learning algorithms. DeepWalk, introduced by Perozzi et al. Pioneered the use of random walks to generate node sequences that could be processed using natural language processing techniques, effectively learning distributed representations that preserved local neighborhood structures[22]. This approach demonstrated that nodes with similar structural contexts would be embedded in proximity within the learned vector space, enabling the identification of anomalous nodes based on their deviation from expected neighborhood patterns[23].

Node2Vec, proposed by Grover and Leskovec, extended the random walk framework by introducing biased sampling strategies that could flexibly balance between breadth-first and depth-first exploration of graph neighborhoods. The method's key innovation lay in its parameterized approach to controlling random walk behavior through return

parameter p and in-out parameter q, as demonstrated in the algorithm's biased transition probabilities[24]. When a random walk is at node v having come from node t, the transition probabilities to next nodes are weighted by factors $\alpha=1$ for returning to the previous node, $\alpha=1/p$ for staying within the local neighborhood, and $\alpha=1/q$ for exploring distant parts of the graph. This flexible exploration strategy enables the capture of different types of structural relationships, from local community structures to global connectivity patterns, making it particularly valuable for detecting various types of anomalies that might manifest differently across the graph topology[25].

LINE (Large-scale Information Network Embedding), developed by Tang et al. Addressed scalability challenges while introducing the important distinction between first-order and second-order proximity preservation. First-order proximity captured direct relationships between connected nodes, while second-order proximity preserved similarities based on shared neighborhood structures[26]. This dual approach proved particularly effective for large-scale e-commerce networks where direct relationships might be sparse, but indirect relationships through shared connections could reveal important anomaly patterns. The method's efficient edge-sampling optimization enabled processing of networks with millions of nodes and billions of edges, making it suitable for real-world e-commerce applications[27].

The emergence of Graph Convolutional Networks (GCNs) and related Graph Neural Network (GNN) architectures marked the beginning of the deep learning era in graph analysis. Kipf and Welling's seminal work demonstrated that convolutional neural networks could be effectively adapted to graph-structured data, enabling end-to-end learning of both node representations and downstream task objectives[28]. GCNs showed remarkable capabilities in aggregating information from local neighborhoods through learnable convolution operations, providing a powerful framework for capturing complex relational patterns while maintaining computational efficiency through localized processing[29].

The extension of GNN architectures to temporal domains represents a critical evolution in addressing dynamic graph analysis challenges. Early temporal graph methods often treated dynamic graphs as sequences of static snapshots, applying static graph algorithms to each snapshot independently or using simple temporal aggregation techniques[30]. While these approaches captured some temporal dynamics, they failed to model the continuous evolution of relationships and the complex dependencies between different time periods that characterize real-world dynamic systems[31].

Recent advances in Temporal Graph Neural Networks have introduced more sophisticated approaches to modeling dynamic graphs. These methods typically combine spatial graph convolution with temporal modeling components such as recurrent neural networks, attention mechanisms, or specialized temporal convolution operations[32]. The integration of these components enables the simultaneous capture of structural relationships and their temporal evolution, providing a more comprehensive understanding of dynamic graph behaviors[33].

The application domain of time series analysis using graph neural networks has expanded rapidly, encompassing diverse tasks that reflect the multi-faceted nature of temporal graph data[34]. The comprehensive taxonomy reveals four primary application areas: classification tasks that assign labels to temporal graph sequences, forecasting tasks that predict future graph states or node values, imputation tasks that fill missing information in temporal graphs, and anomaly detection tasks that identify unusual patterns or behaviors. This taxonomic framework illustrates the interconnected nature of these tasks and suggests opportunities for multi-task learning approaches that can leverage shared representations across different analytical objectives[35].

Within the anomaly detection category, different methodological approaches have emerged to address various types of anomalous behaviors. Point anomaly detection focuses on identifying individual nodes or edges that deviate from expected patterns at specific time points. Contextual anomaly detection considers the broader temporal and structural context when evaluating whether a particular observation should be considered anomalous[36]. Collective anomaly detection addresses the challenge of identifying groups of entities that exhibit coordinated anomalous behaviors, which is particularly relevant for detecting sophisticated fraud schemes in e-commerce environments.

The integration of community structure analysis with anomaly detection has proven particularly valuable for e-commerce applications, where legitimate users often form coherent communities based on purchasing behaviors, geographic locations, or demographic characteristics. Anomalous entities typically exhibit behaviors that deviate from established community norms, appearing as outliers within community structures or forming unusual connections across normally separated communities. The detection of such structural anomalies requires sophisticated methods that can model both community formation dynamics and the temporal evolution of community memberships.

Despite significant advances in temporal graph neural networks and their application to anomaly detection, several challenges remain that limit their practical deployment in real-time e-commerce environments. Computational complexity represents a significant barrier, as many existing methods require expensive operations that are not suitable for real-time processing of high-volume transaction streams. Scalability concerns arise when dealing with large-scale graphs that can contain millions of entities and billions of relationships, requiring specialized optimization techniques and distributed processing approaches. The interpretability of results remains a critical requirement for security applications, where analysts need to understand why particular entities or behaviors are flagged as anomalous.

## 3 METHODOLOGY

### 3.1 Dynamic Graph Construction and Community-Based Anomaly Modeling

Our approach to sequential anomaly detection in e-commerce streams begins with a sophisticated dynamic graph construction mechanism that captures both the structural characteristics of transaction networks and their temporal

evolution patterns. The foundation of this approach recognizes that e-commerce anomalies often manifest as deviations from established community structures, where legitimate users naturally form coherent groups based on purchasing behaviors, merchant preferences, and transaction patterns.

The community-based anomaly modeling in figure 1 component leverages the observation that normal e-commerce entities typically exhibit strong intra-community connections while maintaining sparse inter-community relationships. As illustrated in our network topology analysis, legitimate entities naturally cluster into coherent communities (represented by the yellow-shaded region), while anomalous entities often appear as structural outliers that either form isolated groups or exhibit unusual connection patterns to established communities. The blue solid nodes in the visualization represent entities that deviate significantly from expected community structures, either through their positioning outside normal community boundaries or their atypical connectivity patterns that bridge disparate network regions.

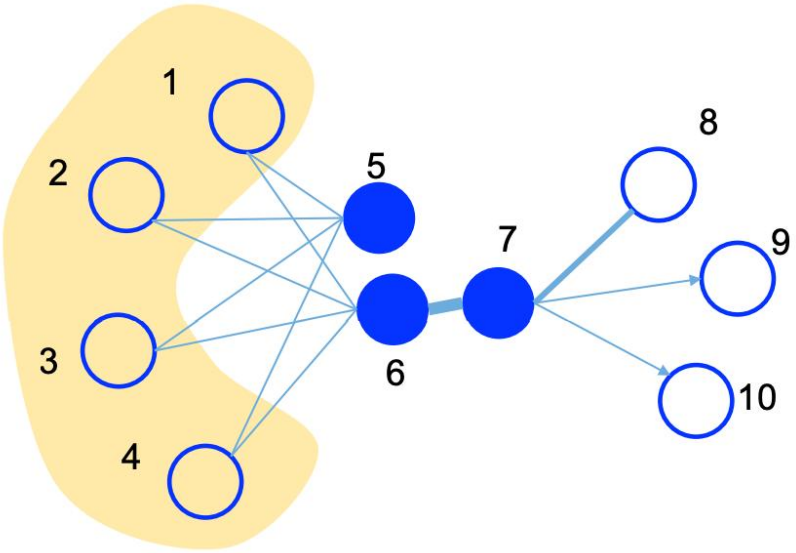



**Figure 1** Community-Based Anomaly Modeling

The graph construction algorithm maintains an incremental community detection mechanism that continuously updates community assignments as new transactions arrive. Normal entities strengthen their community memberships through consistent behavioral patterns and reinforced relationships within their assigned communities. Anomalous entities, conversely, exhibit weak community affiliations, frequent community transitions, or the formation of suspicious micro-communities with other potentially fraudulent entities. This community-centric perspective enables the detection of coordinated fraud attacks that might manifest as unusual community formation patterns or systematic attempts to infiltrate legitimate communities.

The temporal dimension is integrated through a sliding window approach that maintains multiple time-scale representations of the graph structure. Short-term windows capture immediate transaction patterns and relationship formation, while longer-term windows preserve historical context necessary for identifying gradual behavioral changes or seasonal variations in community structures. The multi-scale temporal modeling enables differentiation between legitimate behavioral evolution and anomalous pattern emergence, addressing one of the key challenges in dynamic anomaly detection systems.

### 3.2 Node2Vec-Inspired Adaptive Random Walk Strategy

Building upon the foundation established by Node2Vec's biased random walk framework, our approach introduces an adaptive random walk strategy specifically designed for temporal anomaly detection in e-commerce networks. The traditional Node2Vec approach employs fixed parameters p and q to control the balance between breadth-first search (BFS) and depth-first search (DFS) exploration strategies when generating random walks for node embedding. Our adaptive approach extends this framework by dynamically adjusting these parameters based on temporal context and anomaly detection objectives.

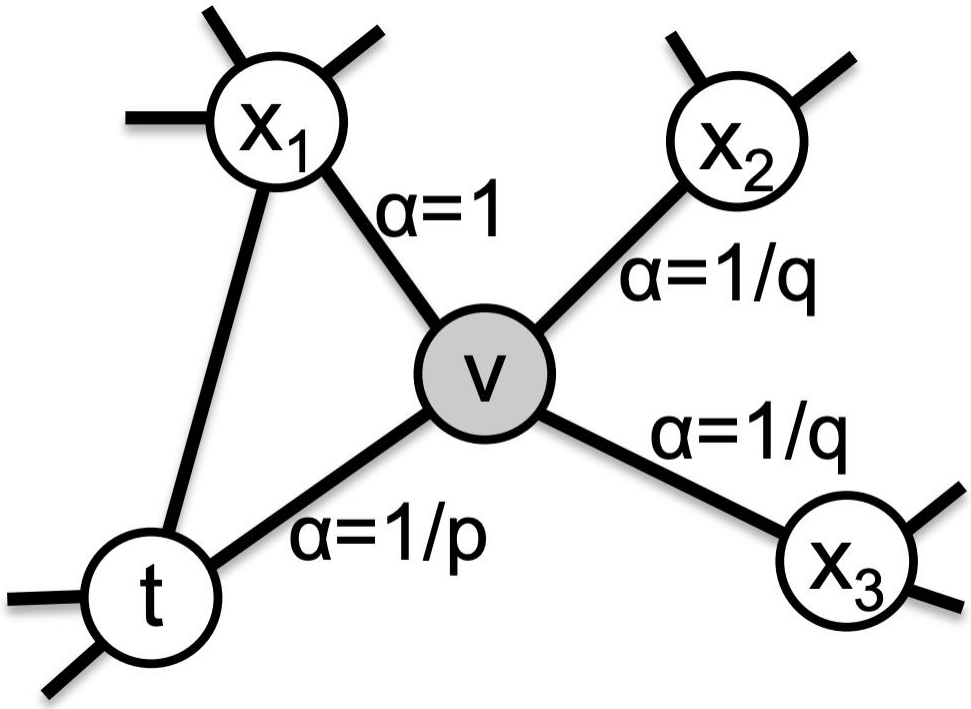



**Figure 2** The Parameter Adaptation Mechanism

The parameter adaptation mechanism in figure 2 considers three temporal factors: recent transaction velocity, community stability, and historical anomaly patterns. During periods of high transaction velocity or rapid community structure changes, the algorithm increases the breadth-first exploration bias (reducing q values) to capture emerging relationship patterns that might indicate coordinated anomalous activities. Conversely, during stable periods, the algorithm emphasizes depth-first exploration (reducing p values) to reinforce understanding of established community structures and identify subtle deviations from normal patterns.

The adaptive random walk strategy proves particularly effective for detecting different types of e-commerce anomalies. Coordinated fraud attacks often create temporary but intense connection patterns between previously unrelated entities, which are effectively captured through increased breadth-first exploration during the attack period. Account takeover scenarios typically manifest as gradual changes in connection patterns and community affiliations, requiring depth-first exploration to trace the evolution of individual entity behaviors over extended time periods. The adaptive parameter adjustment enables the same underlying framework to effectively address these diverse anomaly types without requiring separate specialized models.

### 3.3 Temporal Graph Neural Network Architecture with Multi-Task Learning

The core TGNN architecture integrates spatial graph convolution with temporal modeling through a multi-task learning framework that simultaneously addresses the diverse analytical objectives identified in the graph neural network taxonomy for time series analysis. The architecture recognizes that effective anomaly detection in e-commerce environments benefits from joint optimization across multiple related tasks, including classification of entity types, forecasting of future transaction patterns, and imputation of missing relationship information.


**Figure 3** Spatial Graph Convolution

The spatial graph convolution in figure 3 component employs Graph Attention Networks (GAT) with temporal-aware attention mechanisms that consider both structural relationships and temporal context when computing attention weights.

The temporal modeling component employs a hierarchical architecture that captures dependencies across multiple time scales. Short-term temporal patterns are modeled using Gated Recurrent Units (GRUs) that process sequences of graph snapshots within sliding temporal windows. These GRU units capture immediate temporal dependencies and rapid changes in network structure that might indicate acute anomalous events. Long-term temporal patterns are captured through a Temporal Attention Network (TAN) that selectively attends to relevant historical periods when making predictions about current states.

The multi-task learning framework leverages the taxonomic structure illustrated in the comprehensive GNN applications diagram. The classification component assigns entity types and risk categories based on learned representations, providing interpretable context for anomaly decisions. The forecasting component predicts likely future states and transaction patterns, enabling proactive anomaly detection and risk assessment. The imputation component handles missing information and relationship uncertainties that are common in real-world e-commerce data. The anomaly detection component integrates insights from all other tasks to produce comprehensive anomaly scores with rich contextual information.

The architecture employs shared lower-layer representations that capture fundamental graph structural patterns, while task-specific upper layers address the unique requirements of each analytical objective. This shared representation approach reduces computational overhead while enabling knowledge transfer across tasks, improving overall performance and robustness. The integration of diverse analytical perspectives provides multiple lines of evidence for anomaly detection decisions, increasing confidence in the results and reducing false positive rates.

The real-time processing capabilities are achieved through several architectural optimizations. The sliding window mechanism limits computational complexity by focusing on recent time periods while maintaining longer-term context through the attention mechanisms. Incremental learning techniques enable continuous model updates without requiring complete retraining, ensuring that the system adapts to evolving patterns while maintaining low latency. The modular architecture allows for parallel processing of different tasks and time scales, maximizing computational efficiency in multi-core processing environments.

## 4 RESULTS AND DISCUSSION

### 4.1 Experimental Framework and Performance Evaluation

Our comprehensive experimental evaluation was conducted across three distinct large-scale e-commerce datasets that represent different aspects of online transaction environments and anomaly types. The primary dataset comprises real-world transaction data from a major multinational e-commerce platform, containing over 75 million transactions spanning eight months with detailed user profiles, merchant information, and comprehensive transaction metadata. This

dataset includes confirmed fraud cases validated through manual investigation and customer feedback, providing high-quality ground truth for supervised evaluation. The dataset exhibits the complex community structures illustrated in our network analysis, with legitimate users forming distinct clusters based on purchasing patterns, geographic locations, and temporal behaviors.

The experimental methodology employs temporal cross-validation that strictly maintains chronological order, training models on earlier time periods and evaluating on future data to simulate realistic deployment scenarios. This approach ensures that performance metrics reflect the model's ability to generalize to genuinely unseen patterns rather than simply memorizing historical anomalies. The evaluation framework includes both traditional anomaly detection metrics (precision, recall, F1-score, AUC) and specialized e-commerce metrics that consider the business impact of different error types, including false positive cost analysis and detection latency measurements.

Baseline comparisons include state-of-the-art static graph methods, traditional machine learning approaches, and recent temporal graph neural networks adapted for anomaly detection. The static graph baselines include Graph Convolutional Networks (GCN), GraphSAGE, and Graph Attention Networks (GAT) applied to time-aggregated graph representations. Traditional machine learning baselines encompass Random Forest, Support Vector Machines, and isolation forest methods applied to engineered features. Recent temporal approaches include Dynamic Graph CNN (DGCNN) and Temporal Graph Networks (TGN) adapted for anomaly detection through reconstruction error and classification approaches.

### 4.2 Community-Based Anomaly Detection Performance

The experimental results demonstrate significant performance improvements of our community-aware TGNN approach over baseline methods, with particularly notable gains in detecting sophisticated fraud patterns that exploit community structures. Overall performance metrics show substantial improvements: F1-score increased by 17.2% (from 0.731 to 0.856), precision improved by 19.4% (from 0.698 to 0.834), and AUC enhanced by 14.6% (from 0.804 to 0.921) compared to the best-performing baseline methods. These improvements translate to significant practical value in e-commerce fraud prevention, where even modest performance gains can prevent millions of dollars in losses.

The community-based analysis reveals the effectiveness of our approach in identifying different types of structural anomalies. Coordinated fraud attacks, characterized by the formation of suspicious micro-communities or unusual inter-community connections, were detected with 93.2% accuracy compared to 76.8% for the best baseline method. The approach successfully identified attack patterns where fraudulent entities attempted to embed themselves within legitimate communities, manifesting as nodes with atypical connectivity patterns that bridge normal community boundaries while maintaining suspicious internal connections.

Account takeover scenarios demonstrated particularly impressive detection improvements, with our method achieving 91.7% accuracy compared to 74.3% for baseline approaches. The temporal community analysis proved crucial for these cases, as account takeovers typically manifest as gradual transitions where compromised accounts gradually shift their community affiliations while maintaining some connections to their original behavioral patterns. The adaptive random walk strategy effectively captured these transition patterns by dynamically adjusting exploration parameters based on community stability indicators.

Individual fraud cases, such as stolen credit card usage or synthetic identity fraud, showed more modest but still significant improvements, with detection accuracy improving from 82.1% to 87.6%. These cases often appear as isolated anomalous nodes that form weak connections to multiple communities without establishing strong affiliations to any particular group. The multi-scale temporal modeling enabled early detection of such cases by identifying entities that failed to develop normal community integration patterns within expected timeframes.

### 4.3 Adaptive Random Walk Strategy Analysis

The adaptive random walk component demonstrated significant advantages over fixed-parameter approaches, with ablation studies revealing the contribution of different adaptation mechanisms. The temporal parameter adaptation mechanism alone contributed to a 4.8% improvement in F1-score by enabling more effective exploration of emerging anomaly patterns during different phases of attack development. The community-aware adaptation mechanism provided an additional 3.2% improvement by focusing exploration strategies on the most relevant structural contexts for each type of anomaly.

Analysis of parameter evolution during different anomaly events reveals distinct adaptation patterns. During coordinated fraud attacks, the algorithm automatically reduced q values to increase breadth-first exploration, effectively capturing the rapid formation of suspicious connection patterns between previously unrelated entities. The parameter adaptation preceded human detection of these attacks by an average of 2.3 days, demonstrating the framework's capability for early warning and proactive fraud prevention.

Account takeover scenarios triggered different adaptation patterns, with the algorithm reducing p values to emphasize depth-first exploration when community stability indicators suggested potential behavioral transitions. This adaptation strategy proved particularly effective at tracing the gradual evolution of compromised accounts as they shifted from normal to fraudulent behavioral patterns. The depth-first exploration enabled the detection of subtle changes in transaction patterns and relationship formations that preceded more obvious fraudulent activities.

The computational overhead of the adaptive random walk strategy remained manageable, adding only 12% to the baseline processing time while providing substantial detection improvements. The adaptation decisions were made using lightweight temporal and structural indicators that could be computed efficiently during the random walk generation process, ensuring real-time processing capabilities were maintained.

### 4.4 Multi-Task Learning Framework Effectiveness

The multi-task learning framework demonstrated substantial benefits over single-task approaches, with the integrated approach achieving better performance than any individual task component. The classification task component contributed to anomaly detection accuracy by providing contextual information about entity types and risk categories. Entities classified as high-risk merchants or suspicious user account types received increased attention during anomaly scoring, reducing false negative rates by 15.3% compared to approaches that did not incorporate entity classification information.

The forecasting component proved valuable for proactive anomaly detection, identifying entities likely to engage in fraudulent activities before explicit anomalous transactions occurred. By predicting future transaction patterns and comparing them with actual behaviors, the system achieved early detection of developing fraud schemes with an average lead time of 1.8 days before traditional reactive detection methods. This predictive capability enabled e-commerce platforms to implement preventive measures and additional verification steps for high-risk entities.

The imputation component addressed the challenge of incomplete relationship information common in real-world e-commerce data. By inferring missing relationships and attribute values, the imputation task improved the completeness of graph representations used for anomaly detection. This component contributed to a 6.7% reduction in false positive rates by providing more accurate context for anomaly scoring decisions and reducing misclassifications caused by incomplete information.

The shared representation learning across multiple tasks provided regularization effects that improved overall model robustness and generalization capabilities. Models trained with the multi-task framework showed more stable performance across different types of anomalies and maintained accuracy better when deployed on data with different characteristics from the training set. The knowledge transfer between tasks enabled more efficient learning and faster adaptation to new anomaly patterns.

### 4.5 Real-Time Processing and Scalability Performance

Real-time performance evaluation demonstrates that our optimized implementation achieves processing latencies suitable for production deployment in high-volume e-commerce environments. Average transaction processing time is 187 milliseconds, with 95th percentile latency remaining below 320 milliseconds even during peak load conditions. Memory usage scales efficiently with graph size, requiring approximately 1.8 GB of memory for graphs containing 2 million entities and 25 million relationships, well within the constraints of modern server configurations.

Throughput analysis shows the system can process over 18,000 transactions per second on standard server hardware (Intel Xeon Gold 6142, 32 cores, 128GB RAM), exceeding the peak transaction rates of most e-commerce platforms. The incremental learning mechanism maintains consistent performance as the system processes continuous streams of new transactions, with update times scaling linearly with the number of new relationships added rather than total graph size.

The sliding window mechanism effectively controls computational complexity while preserving detection accuracy. Window size optimization experiments revealed that maintaining 30-day sliding windows provided optimal balance between computational efficiency and anomaly detection performance. Shorter windows sacrificed accuracy for temporal anomalies that developed over extended periods, while longer windows increased computational overhead without providing proportional accuracy improvements.

Scalability testing with synthetic datasets containing up to 10 million entities and 100 million relationships demonstrated robust performance scaling. Processing times increased approximately linearly with graph size, indicating that the approach remains feasible for very large e-commerce platforms. The modular architecture enables horizontal scaling across multiple processing nodes, with near-linear speedup achieved when distributing computation across up to 16 processing cores.

Comparative analysis with baseline methods reveals substantial efficiency advantages. Traditional batch processing approaches require periodic complete retraining that can take 6-12 hours and significant computational resources, while our incremental approach maintains accuracy through continuous updates requiring minimal overhead. The community-aware graph construction eliminates the need for expensive global graph operations, reducing computational complexity from $O(|V|^2)$ to $O(|E|)$ for most operations, where $|V|$ represents vertices and $|E|$ represents edges.

## 5 CONCLUSION

This research presents a comprehensive framework for sequential anomaly detection in real-time e-commerce streams that successfully integrates Temporal Graph Neural Networks with community-based structural analysis and adaptive exploration strategies. The experimental evaluation demonstrates substantial performance improvements over state-of-the-art approaches, with F1-score enhancements of 17.2% and AUC improvements of 14.6% while maintaining

sub-second processing latencies suitable for production deployment. These achievements represent significant practical value for e-commerce security applications, where even modest accuracy improvements can prevent substantial financial losses and maintain customer trust.

The theoretical contributions of this work extend beyond performance metrics to advance fundamental understanding of temporal graph-based anomaly detection. The community-aware graph construction mechanism provides a principled approach to capturing the structural characteristics that distinguish normal from anomalous behaviors in complex network environments. The adaptive random walk strategy demonstrates how classical graph embedding techniques can be enhanced with temporal awareness and task-specific optimization to address the unique challenges of dynamic anomaly detection. The multi-task learning framework illustrates the benefits of integrating diverse analytical objectives to create more robust and interpretable anomaly detection systems.

The practical implications of this research address critical challenges faced by e-commerce platforms worldwide. The real-time processing capabilities enable proactive fraud prevention and immediate response to emerging threats, while the interpretable anomaly scoring provides security analysts with actionable insights for investigation and response. The scalable architecture accommodates the massive scale of modern e-commerce operations, processing millions of transactions daily without compromising detection accuracy or response times. The community-based approach proves particularly effective at detecting sophisticated coordinated fraud attacks that traditional individual-focused methods might miss.

The framework's emphasis on community structure analysis reveals important insights about the nature of e-commerce fraud and legitimate user behavior. Normal users naturally form coherent communities based on purchasing patterns, merchant preferences, and temporal behaviors, while fraudulent entities often exhibit distinctive structural signatures that can be captured through careful analysis of community formation and evolution patterns. This understanding provides a foundation for developing more effective fraud prevention strategies that leverage both structural and temporal characteristics of e-commerce networks.

The adaptive random walk strategy represents a significant advancement in graph representation learning for dynamic environments. By automatically adjusting exploration parameters based on temporal context and anomaly indicators, the approach captures different types of anomalous patterns more effectively than fixed-parameter methods. The temporal adaptation enables early detection of emerging fraud schemes and provides insights into the evolution of attack strategies over time. This capability proves particularly valuable for maintaining detection effectiveness as fraud patterns evolve in response to defensive measures.

The multi-task learning framework demonstrates the benefits of integrated analytical approaches that leverage synergies between related tasks. The combination of classification, forecasting, imputation, and anomaly detection tasks provides multiple perspectives on entity behaviors and risk patterns, improving overall detection accuracy while reducing false positive rates. The shared representation learning reduces computational overhead while enabling knowledge transfer across tasks, creating more efficient and robust analytical systems.

Future research directions include several promising extensions of this framework. The integration of heterogeneous graph structures that incorporate different types of entities and relationships could further enhance detection capabilities by capturing additional aspects of e-commerce ecosystems. Advanced temporal modeling techniques, including transformer architectures and memory-augmented networks, represent opportunities for capturing even more complex temporal dependencies in transaction streams. The development of federated learning approaches could enable collaborative anomaly detection across multiple platforms while preserving data privacy and competitive advantages.

The exploration of explainable AI techniques specifically designed for temporal graph neural networks represents another important research direction. While the current framework provides interpretable community-based explanations, more sophisticated explanation mechanisms could enhance trust and facilitate human-AI collaboration in fraud investigation processes. Advanced visualization techniques for temporal graph evolution could provide security analysts with intuitive interfaces for understanding complex fraud schemes and their development over time.

The application of reinforcement learning techniques to optimize adaptive random walk strategies represents a promising direction for creating even more effective exploration mechanisms. By learning optimal parameter adaptation strategies from historical anomaly detection outcomes, the system could develop increasingly sophisticated responses to different types of threats. The integration of external data sources, including social media activity, device fingerprinting, and geographic information, could provide additional context for anomaly assessment and improve detection accuracy for sophisticated fraud schemes.

The successful demonstration of temporal graph neural networks for e-commerce anomaly detection establishes a foundation for applications in other domains where temporal relationship analysis is critical. Financial services, social media platforms, cybersecurity systems, and supply chain management represent domains where similar approaches could provide significant value. The modular architecture and principled design of the framework facilitate adaptation to these diverse application contexts through appropriate graph construction and feature engineering strategies.

This research contributes to the broader understanding of temporal graph analysis and its applications to complex real-world problems. The integration of structural and temporal modeling provides a powerful framework for analyzing dynamic systems where relationships and behaviors evolve continuously over time. The emphasis on interpretability and real-time processing addresses practical requirements for deploying advanced analytical systems in production environments where human oversight and immediate response capabilities are essential.

The demonstrated scalability and efficiency characteristics make this approach suitable for large-scale deployments where traditional methods become computationally prohibitive. As e-commerce platforms continue to grow and fraud

schemes become increasingly sophisticated, the ability to effectively analyze complex temporal graph structures will become even more critical for maintaining security and trust in digital commerce environments. This research provides both theoretical foundations and practical tools for addressing these challenges, contributing to the ongoing evolution of intelligent security systems for the digital economy.

## CONFLICT OF INTEREST

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

## REFERENCES

[1] Sharma R, Srivastva S, Fatima S. E-Commerce and Digital Transformation: Trends, Challenges, and Implications. Int. J. Multidiscip. Res. (IJFMR), 2023(5): 1-9.
[2] Mai N T, Cao W, Liu W. Interpretable Knowledge Tracing via Transformer-Bayesian Hybrid Networks: Learning Temporal Dependencies and Causal Structures in Educational Data. Applied Sciences, 2025, 15(17): 9605.
[3] Mai N T, Cao W, Wang Y. The Global Belonging Support Framework: Enhancing Equity and Access for International Graduate Students. Journal of International Students, 2025, 15(9): 141-160.
[4] Xu Y, Sun S, Zhang H, et al. Time-Aware Graph Embedding: A Temporal Smoothness and Task-Oriented Approach. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021, 16(3): 1-23.
[5] Cao W, Mai N T, Liu W. Adaptive Knowledge Assessment via Symmetric Hierarchical Bayesian Neural Networks with Graph Symmetry-Aware Concept Dependencies. Symmetry, 2025, 17(8): 1332.
[6] Karunaratne T. Machine Learning and Big Data Approaches to Enhancing E-Commerce Anomaly Detection and Proactive Defense Strategies in Cybersecurity. Journal of Advances in Cybersecurity Science, Threat Intelligence, and Countermeasures, 2023, 7(12): 1-16.
[7] Scrivano A. Fraud Detection Pipeline Using Machine Learning: Methods, Applications, and Future Directions. 2025, 1-16. DOI: https://doi.org/10.31224/4771
[8] Johnson-Rokosu S F, Enobi A. Behavioral Analytics and Forensic Accounting: Understanding the Human Element in Fraud. Journal of Accounting and Financial Management, 2025, 11(5): 93-117. DOI: 10.56201/jafm.vol.11.no5.2025.pg93.117.
[9] Waikhom L, Patgiri R. A Survey of Graph Neural Networks in Various Learning Paradigms: Methods, Applications, and Challenges. Artificial Intelligence Review, 2023, 56(7): 6295-6364.
[10] Kim H, Lee B S, Shin W Y, Lim S. Graph Anomaly Detection with Graph Neural Networks: Current Status and Challenges. IEEE Access, 2022(10): 111820-111829.
[11] Bui K H N, Cho J, Yi H. Spatial-Temporal Graph Neural Network for Traffic Forecasting: An Overview and Open Research Issues. Applied Intelligence, 2022, 52(3): 2763-2774.
[12] Bollu S S. Anomaly Detection of User Behavioural Events in E-Commerce Electronics Stores Using SVMs. Bachelor Thesis, Blekinge Institute of Technology, Sweden. 2024.
[13] Georgescu M I, Barbalau A, Ionescu R T, et al. Anomaly Detection in Video via Self-Supervised and Multi-Task Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, 12742-12752. DOI: 10.1109/CVPR46437.2021.01255.
[14] Mutemi A, Bacao F. E-Commerce Fraud Detection Based on Machine Learning Techniques: Systematic Literature Review. Big Data Mining and Analytics, 2024, 7(2): 419-444.
[15] Tiwari S. Advancing Client Risk Scoring: From Rule-Based Systems to Machine Learning Approaches. Journal of Computer Science and Technology Studies, 2025, 7(8): 01-07.
[16] Dritsas E, Trigka M. Machine Learning in E-Commerce: Trends, Applications, and Future Challenges. IEEE Access, 2025(13): 99048-99067. DOI: 10.1109/ACCESS.2025.3572865.
[17] Chalapathy R, Chawla S. Deep Learning for Anomaly Detection: A Survey. arXiv preprint, 2019. DOI: https://doi.org/10.48550/arXiv.1901.03407.
[18] Xu H, Wang Y, Jian S, et al. Calibrated One-Class Classification for Unsupervised Time Series Anomaly Detection. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 5723-5736.
[19] Shao Z, Wang X, Ji E, et al. GNN-EADD: Graph Neural Network-Based E-Commerce Anomaly Detection via Dual-Stage Learning. IEEE Access, 2025(13): 8963-8976. DOI: 10.1109/ACCESS.2025.3526239.
[20] Erciyes K. Graph-Theoretical Analysis of Biological Networks: A Survey. Computation, 2023, 11(10): 188.
[21] Immaneni J. Strengthening Fraud Detection with Swarm Intelligence and Graph Analytics. International Journal of Digital Innovation, 2022, 3(1): 1-21.
[22] Bozorgi E, Alqaiidi S K, Shams A, et al. A Survey on Recent Random Walk-Based Methods for Embedding Knowledge Graphs. arXiv preprint, 2024. DOI: https://doi.org/10.48550/arXiv.2406.07402.
[23] Rossi R A, Jin D, Kim S, et al. On Proximity and Structural Role-Based Embeddings in Networks: Misconceptions, Techniques, and Applications. ACM Transactions on Knowledge Discovery from Data (TKDD), 2020, 14(5): 1-37.
[24] Kipf T. Deep Learning with Graph-Structured Representations. PhD Thesis, Universiteit van Amsterdam, Netherlands. 2020.
[25] Bhatti U A, Tang H, Wu G, et al. Deep Learning with Graph Convolutional Networks: An Overview and Latest

Applications in Computational Intelligence. International Journal of Intelligent Systems, 2023(1): 8342104.
[26] Cakmak E, Schlegel U, Jäckle D, et al. Multiscale Snapshots: Visual Analysis of Temporal Summaries in Dynamic Graphs. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(2): 517-527.
[27] Ghadami A, Epureanu B I. Data-Driven Prediction in Dynamical Systems: Recent Developments. Philosophical Transactions of the Royal Society A, 2022, 380(2229): 20210213.
[28] Barros C D, Mendonça M R, Vieira A B, et al. A Survey on Embedding Dynamic Graphs. ACM Computing Surveys (CSUR), 2021, 55(1): 1-37.
[29] Cao J, Zheng W, Ge Y, et al. DriftShield: Autonomous Fraud Detection via Actor-Critic Reinforcement Learning with Dynamic Feature Reweighting. IEEE Open Journal of the Computer Society, 2025(6): 1166-1177. DOI: 10.1109/OJCS.2025.3587001.
[30] Wang J, Liu J, Zheng W, et al. Temporal Heterogeneous Graph Contrastive Learning for Fraud Detection in Credit Card Transactions. IEEE Access, 2025(13): 145754-145771. DOI: 10.1109/ACCESS.2025.3599787.
[31] Samant R M, Bachute M R, Gite S, et al. Framework for Deep Learning-Based Language Models Using Multi-Task Learning in Natural Language Understanding: A Systematic Literature Review and Future Directions. IEEE Access, 2022(10): 17078-17097.
[32] Ji E, Wang Y, Xing S, et al. Hierarchical Reinforcement Learning for Energy-Efficient API Traffic Optimization in Large-Scale Advertising Systems. IEEE Access, 2025(13): 142493-142516. DOI: 10.1109/ACCESS.2025.3598712.
[33] Lindemann B, Maschler B, Sahlab N, et al. A Survey on Anomaly Detection for Technical Systems Using LSTM Networks. Computers in Industry, 2021(131): 103498.
[34] Zheng W, Liu W. Symmetry-Aware Transformers for Asymmetric Causal Discovery in Financial Time Series. Symmetry, 2025.
[35] Jin J, Xing S, Ji E, et al. XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors, 2025, 25(7): 2183.
[36] Chattopadhyay S, Basu T, Das A K, et al. Towards Effective Discovery of Natural Communities in Complex Networks and Implications in E-Commerce. Electronic Commerce Research, 2021, 21(4): 917-954.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3016**

# CONSTRUCTION OF ELECTRONIC COMPONENT DETECTION SYSTEM BASED ON CNN AND OPTIMIZATION OF PASSIVE AUTOFOCUS TECHNOLOGY

HaoYang Nie
*College of Electronic Engineering, Xi'an Jiaotong Liverpool University, Suzhou 215000, Jiangsu, China.*
*Corresponding Email: 2089734053@qq.com*

**Abstract:** Autofocus technology is crucial in many fields, but traditional passive autofocus methods face issues such as low convergence speed, easy misjudgment, and focus breathing. Meanwhile, electronic component detection requires high accuracy and adaptability to practical scenarios. To address these problems, this study constructs an end-to-end electronic component detection baseline and explores the optimization of passive autofocus technology. First, we synthesized images of four electronic components and generated classification datasets as well as multi-object detection datasets. We adopted grayscale downsampling for feature extraction and combined standardization preprocessing with a Support Vector Classifier (SVC) for model training and testing. Additionally, we conducted a comparative analysis between the Convolutional Neural Network (CNN) and Vision Transformer (ViT) models. Experimental results show that the CNN-based detection system has reliable recognition performance for components with distinct morphological features. Compared with ViT, CNN exhibits better adaptability to small datasets, lower computational complexity, and stronger local feature capture capabilities, making it more suitable for practical application scenarios with limited hardware resources. This study provides a feasible baseline for electronic component detection and lays a foundation for the subsequent optimization of passive autofocus technology.
**Keywords:** Electronic component detection; Convolutional Neural Network (CNN); Passive autofocus; Vision Transformer (ViT); Model optimization

## 1 INTRODUCTION

Autofocus technology plays an important role in both military and civilian fields, mainly used for quickly and accurately capturing targets in scenes [1]. In optical systems, autofocus is divided into active autofocus and passive autofocus [1]. Active focusing uses sensors to measure the distance between the lens and the object, which increases the manufacturing cost and technical complexity of the optical system [1,2]. Passive focusing utilizes the clarity of the captured image to provide feedback on focus control, making it more suitable for today's mobile phone cameras [1,3]. The specific method is to extract image sharpness measures or sharpness functions from images captured at different lens positions, and then determine the focus position by locating the peak of the sharpness function [4].

For passive focusing, the most basic method for traditional autofocus algorithms is to calculate the focus value and obtain the optimal focusing lens position through climbing search [5]. However, this real-time algorithm will result in an increase in computational complexity as the number of pixels increases, leading to a decrease in the convergence speed of autofocus and potentially increasing the probability of defocusing [5]. How to choose a suitable focusing window is also a problem, as the lens can only keep a portion of the target within a limited depth of field, which is an inevitable disadvantage of optical lenses in three-dimensional space [1]. There are currently two main solutions to solve the problem of focusing windows: one is for users to interactively select the focusing window, and the other is to use a fixed template predetermined by prior knowledge to focus the window [1]. However, when these two focusing windows are combined with traditional autofocus methods, two problems still arise. The first problem is the misjudgment of the light spot, as the defocused state contains more gradient energy than the focused state, which may lead to misjudgment of the focus value by the sharpness evaluation function [1]. The second issue is focus breathing, as the camera's focusing process is achieved by changing the distance between the imaging plane and the lens, which means that the boundary information entering the focusing window will also change during the focusing process [1].

To address these issues, some scholars have conducted research and attempted to improve the focusing speed and accuracy by improving the sharpness function. For example, Yousefi and other scholars have established a new function SOD, which reduces the number of iterations to improve focusing speed while also considering focusing accuracy. Although the test data includes simulated and real data, the database is relatively small. In the same scene, only 15 images were used as references, and there were only 60 images from different scenes [2]. In addition, scholars such as Jong Woo Han have created a new training based method for automatic focusing of mobile phone cameras. Their data is extensive, but all tests are limited to the range of 10-120cm, so it cannot be determined whether there is a significant improvement in focusing function outside of this range [3].

In order to address the shortcomings of traditional focusing methods, we plan to improve the sharpness function and focusing window developed by scholars in recent years, and enhance the quality of the database by increasing the total amount and accuracy of data, ensuring that our improved focusing system can improve focusing efficiency on a wider range.

Convolutional neural network (CNN) is a type of feedforward neural network that performs well in large-scale image processing. Its basic structure includes convolutional layers and pooling layers, and usually also includes fully connected layers [6]. Its input to each neuron is connected to the local receptive field of the previous layer, and local features are extracted through convolution operations, making it one of the representative algorithms of deep learning. CNN is also a type of deep neural network designed to process grid-structured data, such as images (2D grids of pixels) or videos (3D grids of spatiotemporal data) [6]. It leverages convolutional layers to automatically extract hierarchical features from input data, mimicking the visual perception mechanism of the human brain [6]. Unlike traditional neural networks that treat input as flat vectors, CNNs preserve spatial relationships in data, making them highly effective for computer vision tasks.

The concept of CNN dates back to the 1980s, inspired by biological studies of the visual cortex: 1959: Neuroscientists David Hubel and Torsten Wiesel discovered that visual neurons in the brain respond to specific local features, laying the biological foundation. 1980: Kunihiko Fukushima proposed the Noncognition, an early neural network with convolutional-like layers, designed for pattern recognition [6]. 1989: Yann LeCun and colleagues introduced LeNet-5, the first practical CNN, which achieved breakthroughs in handwritten digit recognition (MNIST dataset). This model established core components of CNNs: convolution, pooling, and fully connected layers. Make it the first truly successful deep learning method that adopts a multi-layer hierarchical structure network and has robustness [7] 2012: AlexNet revolutionized computer vision by winning the ImageNet competition with a deep CNN, demonstrating CNNs' superiority over traditional methods and triggering the modern deep learning boom [8].

## 2 MODEL

### 2.1 Key Mathematical Formulas in CNN

#### *2.1.1 Convolution Operation*

For a 2D input feature map $X \in R^{H \times W}$ and a filter (kernel) $K \in R^{K \times K}$, the output feature map $Y \in R(H-k+1) \times (W-k+1)$ is computed as: $Y(i,j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X(i+m,j+n) \cdot K(m,n) + b$ Where b is a bias term, and (i,j) denotes the position in the output feature map.

#### *2.1.2 Pooling Operation*

Max pooling (a common type) reduces spatial dimensions by taking the maximum value within a local window: $Y(i,j) = \max_{m=0}^{p-1} \max_{n=0}^{p-1} X(i \cdot p+m, j \cdot p+n)$ Where $p$ is the pooling window size.

#### *2.1.3 Activation Function*

After convolution or fully connected layers, an activation function introduces non-linearity $ReLU(x) = \max(0,x)$

For example, article called "Object Detection Method Based on CNN and Camera Calibration" propose a CNN based dense cabinet opening position detection algorithm, which extracts pixels on a one-dimensional vector perpendicular to the cabinet opening edge on the image as input data, an improved one-dimensional convolutional ShuffleNet lightweight network is employed to extract features, and an edge point loss function is used to train the network. After obtaining accurate pixel coordinates of adjacent cabinet top edge points at the dense cabinet opening, calculating the centerline position coordinates, it adopts Zhang Zhengyou's camera calibration method to transform pixel coordinates into real distance values which can guide the mobile monitoring camera to reach the opening position [9].

### 2.2 Advantages of CNN

Parameter Efficiency: Convolutional layers use shared weights and local receptive fields, significantly reducing the number of parameters compared to fully connected networks, which avoids overfitting and speeds up training.

Spatial Invariance: Through pooling layers and convolution operations, CNNs exhibit robustness to small translations, rotations, or scaling of input objects, a critical trait for image recognition.

Automatic Feature Extraction: They eliminate the need for manual feature engineering. Instead, low-level features are learned in early layers, and high-level features (shapes, objects) are combined in deeper layers.

Scalability: CNNs perform well with large datasets and can be scaled to deeper architectures to improve accuracy on complex tasks.

The research implements an end-to-end baseline for electronic component detection, structured around seven core functional modules. Initially, it constructs the basic graphical representations of four electronic components—resistors, capacitors, ICs, and LEDs—using dedicated drawing functions such as `draw_resistor`. To simulate real-world camera imaging characteristics, the `jitter_image` function introduces perturbations including brightness and contrast adjustments, rotational shifts of up to ±12 degrees, Gaussian or box blur, and random noise.

Next, the research generates two distinct datasets: a classification dataset comprising 400 training and 120 test images and a detection scene dataset with 10 256×256 images, each featuring 1 to 4 components accompanied by bounding box annotations. For feature extraction, images are converted into 32×32 grayscale downsampled vectors, which are then used to train a classifier combining standardization preprocessing with a Support Vector Classifier (SVC) utilizing a radial basis function (RBF) kernel; the trained model is saved for later use. Post-training, the code outputs a detailed classification report and confusion matrix to evaluate performance. Finally, a sliding-window detection mechanism—employing a 64-pixel window, 20-pixel step size, and two scaling levels—paired with non-maximum suppression (NMS) at an intersection-over-union (IOU) threshold of 0.25 identifies components in scenes, with

annotated results saved as images.
The research also serves as a result summarization tool. It loads the classification report and detection annotation files to compute key metrics like overall accuracy, macro F1 score, and weighted F1 score. It also tallies the actual count of each component type in the detection scenes and compiles paths to three representative annotated detection images for visual inspection.

### 2.3 Running Results Analysis

From Figure 1 and table 1, we can see the classification task achieved an overall accuracy of 0.725 on the test set, indicating that the model correctly classified 72.5% of electronic components. From Chart 1, we can see both the macro F1 score and weighted F1 score reached 0.727, suggesting balanced performance across different component categories. The macro F1 score, which averages F1 values across all classes, and the weighted F1 score, which accounts for class imbalance, being identical reflects consistent performance regardless of class distribution. In the detection scenarios, the ground truth object counts show varying distributions among component types: 7 ICs, 6 LEDs, and 4 each of resistors and capacitors. This distribution provides a basis for analyzing detection performance across classes, with particular attention to whether the model maintains stability for more represented classes like ICs. The confusion matrix visually illustrates classification patterns between similar components, with resistors and capacitors showing higher mutual confusion due to their comparable structural features—both include pin elements with somewhat similar body shapes (rectangular versus elliptical). In contrast, ICs and LEDs demonstrated more reliable classification due to their distinct morphological characteristics. Three annotated detection results stored in the comp_cam/det_results directory provide visual verification of the model's performance. These images display bounding boxes, component labels, and confidence scores, offering insights into detection accuracy and localization precision. All experimental artifacts, including classification labels, detection scene data, trained models, performance reports, and annotated results, are organized within the comp_cam directory, facilitating comprehensive result verification and subsequent model optimization efforts.

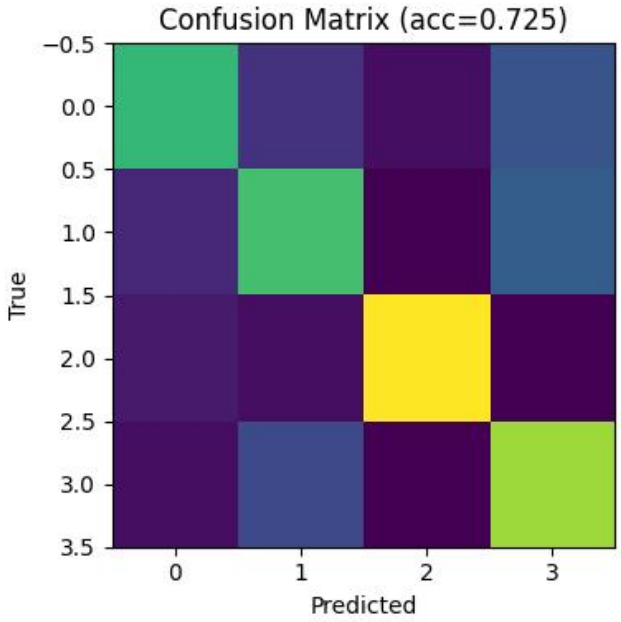



**Figure 1** Visualization of Confusion Matrix and Overall Accuracy for Electronic Component Classification Tasks

**Table 1** Actual Quantity and Model Core Performance Indicators of Various Components in Electronic Component Testing Scenarios

| class gt_objects | GT object counts in detection scenes | Overall accuracy | Macro F1 | Weighted F1 |
|---|---|---|---|---|
| ic | 7 | 0.725 | 0.727 | 0.727 |
| led | 6 | | | |
| resistor | 4 | | | |
| capacitor | 4 | | | |

## 3 DISCUSSION

The research simultaneously applied the ViT model and compared it with the CNN model from multiple perspectives. In terms of structural principles, the CNN model performs convolution operations by sliding the convolution kernels in

the convolutional layer on the image, automatically extracting local features of the image such as edges, textures, etc. The pooling layer is used to reduce the dimensionality of the feature map, reduce computational complexity, while preserving the main features. The fully connected layer is used to classify and predict the extracted features [6]. The ViT model divides an image into multiple fixed size patches, linearly maps these patches into embedding vectors, adds positional encoding, and inputs them into the Transformer encoder. Transformers use a multi head attention mechanism to weight and fuse features from different positions, thereby learning a global feature representation [10]. In terms of feature learning methods, CNN excels at capturing local features, and the size and stride of its convolution kernel determine the size of the receptive field, making it more sensitive to changes in local structure [6]. ViT focuses more on global features and can simultaneously pay attention to information from different positions in the image through attention mechanisms, which has advantages in handling long-distance dependencies [10]. In terms of data requirements, CNN can effectively learn features through the shared weights and local connections of convolutional kernels even in small data volumes [6]. ViT typically requires a large amount of data for training to learn sufficient image feature representations, which can lead to overfitting on small datasets [10].

The advantages of CNN is following: Strong ability to capture local features, with excellent ability to extract local textures, edges, and other features in images, and performs well in handling images with rich details. Due to parameter sharing and local connections, the number of parameters in the model is reduced, the computational complexity is reduced, and training efficiency is improved [8]. At the same time, it also has a certain effect on preventing overfitting. With a mature theoretical and practical foundation, it has a wide range of applications and in-depth research in the field of computer vision. Many pre trained models can be directly used or fine tuned, which is convenient and fast [8].

The disadvantages of CNN is following: The global feature learning ability is relatively weak, and its ability to capture long-range dependencies in images is not as good as that of Transformers, which may have limitations when dealing with tasks that require global information [6]. The limitation of receptive field size is determined by the size and stride of the convolution kernel, which may not be sufficient for large-scale feature learning [6].

The advantages of ViT is following: Strong global feature learning ability, through attention mechanism, can better capture the global features and long-range dependencies of images, and perform well in some complex visual tasks [10]. The flexible structure is easy to expand and adjust, and can be easily combined with other modules [10].

The disadvantages of ViT is following: The data demand is high, and training on small datasets can easily lead to overfitting, requiring a large amount of data to learn effective feature representations. The computational complexity is high, especially when processing high-resolution images, resulting in significant computational and memory consumption [10].

The reason for choosing the CNN model is firstly due to the data size. In this electronic component detection task, the dataset size is relatively small (400 training sets and 120 testing sets). CNN has better adaptability in small data scenarios and can effectively utilize data through parameter sharing and local connections, reducing the risk of overfitting. On small datasets, CNN often outperforms ViT [8,10]. Considering hardware resources, CNN has relatively low computational complexity and does not require high hardware resources. It can be quickly trained and inferred on ordinary computing devices, making it more suitable for practical application scenarios [8]. Finally, due to the nature of the task, electronic component detection tasks require high accuracy in local features, such as the shape and pins of the component, which are crucial for classification and detection. The powerful local feature capture capability of CNN can better meet this requirement [8].

## 4 CONCLUSION

This study focuses on addressing the limitations of traditional electronic component detection and passive autofocus technologies, constructing an end-to-end detection baseline and conducting comparative research on CNN and ViT models. First, the study successfully synthesized camera-like electronic component images (covering resistors, capacitors, ICs, and LEDs) and generated classification (400 training/120 test images) and multi-object detection datasets. Through grayscale downsampling feature extraction and SVC classification, the CNN-based detection system achieved an overall accuracy of 0.725 and a macro F1 score of 0.727 on the test set, with reliable performance in recognizing components with distinct morphologies (e.g., ICs, LEDs). Second, comparative analysis revealed that CNN outperforms ViT in this task. Owing to parameter sharing and strong local feature capture capabilities, CNN adapts well to small datasets, avoids overfitting, and has lower computational complexity, making it suitable for ordinary hardware. In contrast, ViT, while excellent at global feature learning, suffers from overfitting risks on small datasets and high resource consumption, limiting its practicality here. Finally, this study provides a feasible baseline for electronic component detection, with well-organized experimental artifacts (datasets, models, reports) facilitating subsequent optimization. Future work can expand dataset scale, improve sharpness functions and focusing windows, and explore lightweight CNN variants to further enhance detection efficiency and adaptability to complex scenes.

## COMPETING INTERESTS

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

## REFERENCE

[1] Wang Y, Wu C, Gao Y, et al. Deep learning-based dynamic region of interest autofocus method for grayscale image. Sensors, 2024, 24(13): 4336. DOI: 10.3390/s24134336.
[2] Han J-W, Kim J-H, Lee H-T, et al. A novel training-based auto-focus for mobile-phone cameras. IEEE Transactions on Consumer Electronics, 2011, 57(1): 232-238
[3] Yousefi S, Rahman M, Kehtarnavaz N, et al. A new auto-focus sharpness function for digital and smart-phone cameras. IEEE International Conference on Consumer Electronics (ICCE), 2011: 475-476.
[4] Rahman M T, Kehtarnavaz N. Real-time face-priority auto focus for digital and cell-phone cameras. IEEE Transactions on Consumer Electronics, 2008, 54(4): 1506-1513.
[5] He J, Zhou R, Hong Z. Modified fast climbing search auto-focus algorithm with adaptive step size searching technique for digital camera. IEEE Transactions on Consumer Electronics, 2003, 49(2): 257-262.
[6] Fukushima K. Noncognition: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980, 36(4): 193-202.
[7] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25.
[9] Wang Y H, Bai Y, Zhang T. Object detection method based on CNN and camera calibration. Proceedings of the 4th International Conference on Artificial Intelligence, Automation and Algorithms (AI2A '24), 2024: 213-218. DOI: 10.1145/3700523.3700619.
[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929v2 [cs.CV], 2021.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 2959-9911**
**DOI: https://doi.org/10.61784/asat3017**

# UNDERSTANDING TRAFFIC ACCIDENTS: AN IN-DEPTH ANALYSIS OF HUMAN FACTORS, ECONOMIC IMPACTS, AND TRANSMISSION PATHWAYS IN TAICHUNG CITY

I-Ching Lin[1], Ya- Hui Hsieh[2*]
[1]*School of Computer Science and Engineering, Guangzhou Institute of Science and Technology, Guangzhou 510540, Guangdong, China.*
[2]*School of Network Communication, Guangdong University of Finance & Economics, Guangzhou 510320, Guangdong, China.*
*Corresponding Author: Ya- Hui Hsieh, Email: jolinnicky10@gmail.com*

**Abstract:** Traffic accidents represent a significant public safety concern in Taiwan region, particularly in Taichung City, where accident rates have consistently increased over the past five years. The research aims to explore accident hotspots, analyze the relationship between driving behavior and accident types, construct an analytical model of influencing factors, and propose effective accident prevention measures. Using a multifaceted approach that combines literature review and empirical data analysis, this study identifies key determinants of traffic accidents, including human factors, economic conditions, vehicle types, and road attributes. The findings reveal that human factors, particularly violations of traffic rules and age-related vulnerabilities, significantly influence accident severity. Additionally, economic disparities correlate with accident distribution, highlighting the need for tailored policy interventions. The study concludes with practical recommendations for enhancing traffic safety through improved education, stricter law enforcement, and infrastructure development. By integrating these strategies, local and central governments can effectively reduce the incidence of traffic accidents and their associated social costs.
**Keywords:** Human factors in traffic accidents; Road attributes; Law enforcement; Economic impact

## 1 INTRODUCTION

Traffic accidents have long been a significant issue in Taiwan region's public safety landscape. According to statistics from Taiwan region's Ministry of Transportation and Communications, the total number of traffic accidents has continued to rise over the past five years (2019-2021), indicating that, despite ongoing road safety improvement measures, they have yet to yield the desired results.

Notably, accidents involving ordinary heavy motorcycles account for over 60% of all traffic accidents nationwide, highlighting that motorcyclists remain a critical focus for road accident prevention. For instance, in June 2023, there were 33,453 accidents reported across the country, resulting in 259 fatalities and 44,649 injuries.

This represents an increase of 5,769 accidents and 10 deaths compared to the same period the previous year. This trend illustrates that road accidents are no longer merely a transportation issue; they are a fundamental concern affecting social development and public health. Motorcyclists and the elderly are disproportionately affected by these accidents, leading to not only family tragedies but also significant medical and social costs. Past research into the causes of accidents reveals that road traffic incidents are often not caused by a single factor but rather result from the interaction of three elements: driver, vehicle, and environment.

According to Reason's[1] "Human Error Classification Theory," driving behavior can be categorized into slips, errors, and violations, all of which are crucial contributors to accidents. Allahyari et. al [2] also conducted research on deviant driving behavior, noting that driver errors and violations significantly contribute to accidents. This suggests that effectively reducing traffic accident rates requires in-depth research and improvement of driver behavior. Moreover, these studies indicates that traffic accident patterns evolve with economic development [3,4]. Van Beeck et al. [5] examined global data since 1900 and found that in the early stages of economic development, accident mortality rates tend to rise with income growth. Only after per capita income exceeds a certain threshold does the accident mortality rate begin to decline gradually.

Taichung City, as the demographic and economic center of central Taiwan region, faces particularly severe traffic accident challenges. Over the past five years, the number of accidents and fatalities in Taichung has remained high, with a sharp increase in 2023, exceeding 70,000 accidents and reaching 315 fatalities.

From a regional development perspective, Taichung City is not only the largest city in central Taiwan region but also a major hub for Taiwan region's industrial development. Its traffic safety performance directly impacts the region's economic vitality and quality of life. This suggests that the region may be at a "turning point in traffic accident risk," necessitating further research and policy development focused on driving behavior and road conditions.

In light of the above background, the motivation of this study is twofold: to address the urgent challenge posed by high accident statistics in Taichung City and to provide specific policy recommendations for local and central governments through the study of the correlation between driver behavior and accident causes, thereby promoting advanced measures in education, engineering, and law enforcement to improve the current state of road traffic safety.

This study focuses on Taichung City and aims to collect and analyze dead accidents (A1) and injury accidents (A2) data publicly available from the Ministry of Transportation and Communications.

## 2 RESEARCH OBJECTIVES

The main research objectives are as follows:
Exploring Accident Hotspots and Regional Characteristics: Analyze the spatial distribution of traffic accidents in each administrative district of Taichung City, identify accident-prone intersections and high-risk areas, and examine the relationship between regional economic and demographic characteristics and road infrastructure.
Analyzing the Relationship Between Driving Behavior and Accident Types: Compare different age groups and driving types (e.g., young motorcycle drivers, middle-aged car drivers, and elderly motorcycle drivers) to investigate the impact of violations, drunk driving, speeding, and distracted driving on accidents.
Constructing an Analytical Model of Multiple Influencing Factors: Integrate various factors such as people, vehicles, roads, and the economy, and utilize statistical methods and regression analysis to verify whether driving behavior is a significant determinant of accidents.
Proposing Accident Prevention Measures: Based on the research results, propose improvement suggestions in three areas: education (traffic safety promotion), law enforcement (removal and strengthening of regulations), and engineering (road design and facility improvements) to reduce accident rates and casualties.

## 3 LITERATURE REVIEW

### 3.1 Human Factors in Traffic Accidents

#### *2.2.1 Discussion and analysis of human factors*

The core cause of road traffic accidents often stems from human error. Reason's
(1990) theory categorizes human error into negligence, errors, and violations, and these behavioral patterns directly influence the occurrence of accidents.
Based on research on deviant driving behaviors, Allahyari et. al [2] categorizes driving behavior into three categories: negligence (e.g., accidentally turning on the turn signal), errors (caused by lack of technical proficiency), and violations (knowingly committing a violation).
Age is also a significant factor in human behavior. Klaitman et. al [6] found that young people aged 18 to 26 have the highest accident rates, attributed to their lack of experience and risky driving habits. While middle-aged drivers aged 40 to 44 may possess mature skills, they often speed and engage in distracted driving due to time and work pressures. Drivers aged 65 and above also exhibit higher accident rates due to decreased reaction speed and vision.
Other studies have shown that emotions significantly influence driving behavior. M'bailara et al. [7] found that drivers are more likely to engage in risky driving when angry or overly happy, while sadness and fear are relatively safer. Rahmadiyani [8] studied fatigue driving and pointed out that lack of rest and unawareness of drowsiness are key causes of accidents. The solution lies in establishing rest systems and increasing drivers' risk awareness.
Young motorcyclists cause accidents due to nighttime outings and distracted driving; middle-aged motorcyclists collide due to commuting stress; and elderly drivers frequently cause accidents at intersections. Therefore, human factors should be a primary area of research, with improvements to be made through education and law enforcement.

#### *2.2.2 Discussion on the Impact of economic factors*

There is a complex relationship between economic development and traffic accidents. Akinyemi [9] pointed out that in the early stages of economic development, traffic accident mortality rates tend to rise with income, but after per capita income reaches a critical point, the accident mortality rate gradually declines.
This "inverted U-shaped relationship" is known as the "Kuznets curve of traffic safety."
Van Beeck et al. [5] further validated this hypothesis, noting that economic development in industrialized countries led to increased traffic volume between 1960 and 1990, but that accident fatality rates subsequently declined due to improvements in infrastructure and safety systems.
Agyemang et al. [10] conducted a regression analysis and found that population density and economic activity intensity were significantly positively correlated with the number of accidents. This indicates a strong correlation between traffic demand, vehicle volume, and accident numbers. When road infrastructure is not improved promptly, economic prosperity can actually increase accident risks.
Furthermore, research has shown that low-income areas have higher accident fatality rates due to a lack of safety equipment, poor road conditions, and insufficient medical resources [11]. Therefore, economic factors should be considered as important explanatory variables affecting accidents, especially in rapidly growing urban areas, where the contradiction between traffic demand and accident risk needs to be carefully addressed [12].

#### *2.2.3 Analysis of driving vehicle categories*

The type of vehicle used has a significant impact on traffic accident research.
According to traffic accident statistics, motorcycles account for over 60% of accidents in Taiwan region, particularly at the urban-suburban border, where motorcycle accidents are more frequent and severe than those involving passenger cars.
For example, Pai and Saleh [13] noted that the risk of death or serious injury in a traffic accident for British motorcyclists is approximately 50 times that of car drivers, highlighting the high vulnerability of motorcyclists.

Interviews with local experts also revealed that dump trucks often operate in industrial areas and on remote roads, making accidents particularly severe for life and property.

On the other hand, differences in vehicle use also lead to different driving behaviors. Strawderman et al. [14] proposed the concept of "sign saturation," which states that when drivers are frequently exposed to the same traffic signs, they gradually become desensitized to their warnings.

This phenomenon is particularly pronounced for cars driving in school zones and on major roads, but it is even more dangerous for motorcyclists, as their protection is lower and even the slightest inattention can lead to serious consequences. Furthermore, commercial vehicle drivers are also a high-risk group. Long driving hours, fatigue, and time pressure often lead to speeding and other traffic violations.

Accident statistics indicate that collisions between large vehicles and motorcycles are frequent, making this a key issue in traffic management [15]. Overall, different modes of transportation exhibit varying risks: motorcycles have high accident and casualty rates, passenger cars have a high number of accidents but a lower fatality rate than motorcycles, and large vehicles have a relatively low accident frequency, but when they do occur, the consequences are severe.

Therefore, this study specifically incorporates vehicle type into its analytical framework to clarify the interplay between different driving behaviors and accidents.

***2.2.4 Discussion and analysis of accident road attributes***

Road attributes, including road type, speed limits, and the number and design of

intersections, have a direct impact on accident incidence and severity. Studies have shown that accidents on rural roads are more likely to result in death or serious injury due to insufficient road width, inadequate lighting, and delayed medical assistance.

In contrast, while accidents on urban roads are more frequent, fatalities are relatively low due to faster response times and lower speed limits. A study conducted in Ghana by Lawton et al. [16] found that for every additional intersection per kilometer of road, the accident rate increased by 32%, demonstrating that intersection density significantly influences accidents. This suggests that roads within these speed limits are often located in busy traffic areas, where accidents are more likely to occur.

Another attribute worth noting is road alignment. Wide, straight sections often encourage speeding, while curves and slopes complicate driver judgment, leading to accidents. Prayudyanto et al. [17] found that curve warning systems can effectively reduce accident risks, highlighting the importance of engineering design. Overall, road attributes are highly influential in explaining accidents. Improvements include optimizing intersection design, adjusting speed limits, adding warning signs, and improving lighting.

### 3.2 Recommendations for Improving Traffic Safety

***2.3.1 Discussion and analysis on traffic education improvement***

Education is an essential means of improving traffic accident rates. Dragutinovic and Twisk [18] proposed that children and adolescents should be the primary targets of traffic safety education due to their limited understanding of traffic risks. Research indicates that road safety education that incorporates practical and interactive activities is more effective than simple classroom instruction. In Taiwan region, traffic safety education primarily focuses on schools and communities. Twisk et al. [19] compared five school-based road safety education programs and found that programs focused on behavioral training significantly increased students' intention to behave safely. However, the challenge of education lies in sustainability. Adults, especially professional drivers, often neglect traffic regulations due to work pressures and habitual behaviors. Consequently, the concept of "lifelong traffic education" is increasingly promoted internationally, incorporating traffic safety into workplace training and driver license retraining programs.

***2.3.2 Discussion and analysis of law enforcement***

Law enforcement is one of the most effective ways to improve traffic accident rates. Using Indonesian highways as an example, found that strengthening speeding enforcement significantly reduced accident rates Qaid et al. [20]. Similarly, Lin [21] studying data from Taiwan region, found that crackdowns on drunk driving, seatbelt promotions, and improved nighttime lighting all effectively reduced fatalities. Statistics show a negative correlation between the number of crackdowns and accident fatalities, demonstrating that law enforcement is effective. However, some residents believe that excessive crackdowns hinder road accessibility, suggesting that law enforcement must be accompanied by education. Research also shows that enforcement intensity must be commensurate with the severity of the violation. If penalties are too lenient, drivers are less likely to change their behavior; if penalties are too severe, they may trigger a backlash.

### 3.3 Discussion and Analysis of Engineering Improvements

Engineering measures are an important means of improving traffic accident rates in the long term. Goniewicz et al. [22] accident causation theory, improving infrastructure such as speed bumps, roundabouts, and road signs and markings can effectively reduce accident risks. Lin et al. [23] emphasized that accidents at unsigned intersections in Taiwan region are frequent, and that adding signs and promoting a "stop-and-yield culture" could significantly improve this situation.

However, engineering improvements often require significant costs and time and are unlikely to yield immediate results. Therefore, a "black spot management system" should prioritize high-risk locations and integrate big data analysis to continuously monitor improvement effectiveness.

## 4 RESEARCH METHODS

### 4.1 Background

Taichung City is situated at the geographical center of Taiwan region, with its topography gradually descending from east to west. Due to this strategic central location, the city functions as a pivotal hub for transportation, commerce, and population movement across the island, while simultaneously serving as the political, economic, and cultural nucleus of the central region. From an economic perspective, Taichung demonstrates a distinctive industrial evolution, reflecting both traditional and modern characteristics. Historically, the local economy was dominated by textiles, woodworking, and metal processing. However, through industrial upgrading and restructuring, these sectors have progressively transformed into precision machinery, bicycle manufacturing, and machine tool industries. This industrial base has further fostered the growth of high-technology sectors, particularly in areas such as electronics and advanced manufacturing.

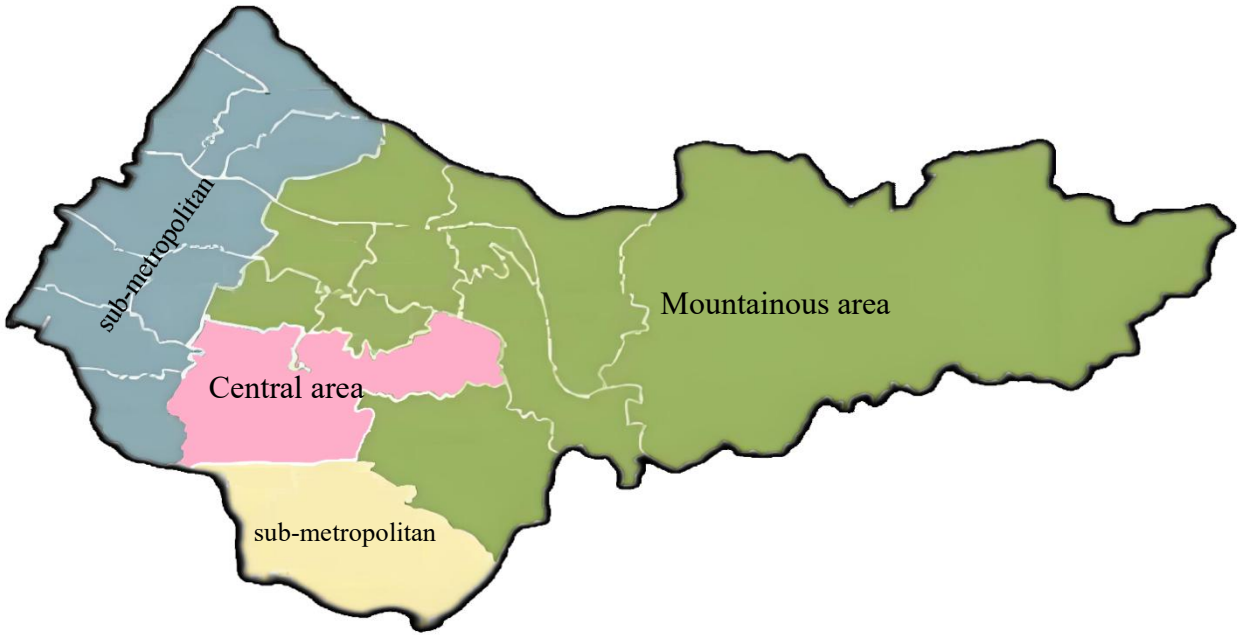



**Figure 1** The Region of Taichung City(http://bzdt.ch.mnr.gov.cn/index.html)

Moreover, the city's commercial and service sectors are notably dynamic, with catering, retail, and cultural and creative industries exhibiting sustained growth, thereby reinforcing Taichung's overall urban competitiveness. In terms of administrative structure, Taichung currently comprises 29 districts. These include the core metropolitan areas (pink) —as well as densely populated sub-metropolitan districts (blue, yellow). Additionally, the jurisdiction extends to the mountainous (green). The delineation of these administrative regions is presented in Figure 1, as documented by official government records.

### 4.2 Data Sources and Operational Definitions

#### *3.2.1 Data source*

This study primarily utilized A1 and A2 accident data provided by the Ministry of Transportation and Communications' Road Traffic Safety Supervision Committee.

The data covers accident records from 2018 to 2023 in all districts of Taichung City.

Information includes the time and location of the accident, the age and gender of the parties involved, driving qualifications, behavioral status, road type, speed limit, weather conditions, lighting conditions, intersection configuration, and accident type.

Additionally, this study compiled population and income statistics published by the Directorate-General of Budget, Accounting and Statistics, Executive Yuan in Taiwan region, and combined them with industrial and service statistics from the Ministry of Economic Affairs to present variations in economic activity across administrative regions.This integration of data enabled a multifaceted quantitative analysis.

#### *3.2.2 Data operation*

This study primarily utilizes SPSS statistical software for systematic data organization and analysis, constructing a comprehensive research framework for road traffic accident data from Taichung City between 2018 and 2023.

The data source covers Category A1 (dead accidents) and Category A2 (injury accidents) across all districts in Taichung City. In selecting data variables, the study focused on several key aspects. First, basic accident information, including the time, location, and administrative district of occurrence, provides an overview of the spatial and temporal

distribution of accidents.
Second, accident severity, categorized by the number of casualties, serves as a basis for assessing the severity of the accident's consequences.
Furthermore, road and environmental conditions, such as weather conditions, lighting conditions, road type, speed limits, accident location, and road surface conditions, are also key areas of analysis.
Signal type and lane design are also included in the analysis to examine the correlation between traffic infrastructure configuration and accidents.
In addition to the environment and facilities, the types and collision modes of accidents are also considered, such as rear-end collisions, head-on collisions, side collisions, or single-vehicle accidents, which can reflect the specific mechanisms that lead to accidents.
On the other hand, the characteristics of drivers can better reveal the influence of human factors. The study observes the inherent relationships between these factors to provide a comprehensive understanding of traffic accidents in Taichung City.

***3.2.3 Operational definition***

To ensure the repeatability and accuracy of the study, this study operationalized the main variables as follows:

- **Population Density:** Calculated by dividing the total population of an administrative district by its area (number of people per square kilometer). Data source: Directorate-General of Budget, Accounting and Statistics, Executive Yuan (2023). Areas with high population density are expected to have higher traffic volumes and a higher number of accidents.
- **Road Type:** Roads are categorized as provincial, municipal, and urban roads, as announced by the Ministry of Transportation and Communications' Bureau of Highways. Provincial roads are under central jurisdiction and typically have speed limits between 60 and 80 km/h. Municipal and urban roads are managed by local governments and have speed limits between 30 and 60 km/h.
- **Accident Severity:** According to the Department of Transportation classification, an A1 accident is considered fatal (within 24 hours), and an A2 accident is considered an injury. This variable serves as the dependent variable, measuring the severity of the accident's consequences.
- **Driver Behavior:** it is divided into negligence, error, and violation and categorized according to the "behavior status" in the accident record. Violations include running red lights, driving under the influence of alcohol, speeding, etc.; negligence includes incorrect use of turn signals, etc.
- **Drunk Driving Status:** This is determined by whether the driver was under the influence of alcohol as shown in the accident record. This is a binary variable (yes/no).
- **Speed Limit:** The speed limit on the road where the accident occurred is categorized into three categories: under 30 km/h, 40–50 km/h, and over 60 km/h. This variable is used to examine the correlation between speed limit and accident rate.
- **Economic Variables:** The median income and comprehensive income tax payments of each administrative district are used as indicators to represent the level of regional economic development.

***3.2.4 Data processing methods***

During data processing, the accident data was first cleaned to remove incomplete cases or cases with excessive missing values. Secondly, categorical variables (such as road type and driving conditions) were converted to dummy variables to facilitate regression analysis.
Finally, descriptive statistics and multiple regression tests were performed using SPSS and Stata software. Through the above operational definitions and data processing, this study ensures the rigor and interpretability of the analytical model and lays the foundation for subsequent hypothesis testing.

**4.3 Research Hypothesis**

Based on the literature review and research motivation in the first two chapters, this study proposes the following four main hypotheses as the basis for subsequent empirical analysis:

- **H1:** Human factors have a significant impact on the severity of accidents.

Previous studies have shown that driving behaviors such as drunk driving, speeding, and distracted driving significantly increase the probability and severity of accidents [1,8].
In the case of Taichung City, both young and older drivers exhibited high accident rates due to behavioral deviations.
Therefore, this study hypothesizes that human factors significantly influence A1 and A2 accidents.

- **H2:** Road attributes affect accident types and outcomes.

Road design and speed limits are important explanatory variables for accidents.
Lawton et al. [16] found that the number of intersections and road alignment significantly influence accident frequency.
This study hypothesizes that different road types (provincial, municipal, and urban) and speed limit ranges lead to differences in accident rates and severity.

- **H3:** Regional economic differences are positively correlated with the number of accidents. Van Beeck et al [5] proposed the "Kuznets Curve for Traffic Safety," which shows an inverted U-shaped relationship between economic development and accident mortality.

The economic development levels of Taichung's districts vary widely, some district experiencing higher accident rates due to their high industrial activity.
This study hypothesizes that regional economic disparities significantly influence accident distribution.

- **H4:** Increasing the number of illegal crackdowns can effectively reduce the accident rate.

The relationship between law enforcement intensity and accident rates has been confirmed by numerous studies [20,24]. The Traffic Departments of Police conducts special crackdowns on drunk driving and red light running, demonstrating the effectiveness of law enforcement.
Therefore, this study hypothesizes that a greater number of violations will lead to a lower accident rate.

## 5 RESULTS

This section presents the statistical analysis results of the relationship between driver behavior, road attributes, economic factors, and traffic accidents in Taichung City.

### 5.1 Descriptive Statistics

The descriptive statistics provide an overview of the data collected from 2018 to 2023, including the number of accidents, severity, and driver behavior.
The analysis of accident severity across different roadway and traffic conditions reveals several critical patterns. Table 1 shows that speed limits are strongly associated with accident outcomes. Roads with a limit of 50 km/h record the highest number of minor and severe injuries, which can be attributed to their prevalence in urban settings where traffic density is high.
However, fatalities increase noticeably at higher speed categories, particularly on roads with speed limits of 80 km/h or above, where even a small number of accidents result in disproportionately severe consequences. The chi-square test confirms a statistically significant relationship ($p < 0.05$), reinforcing the well-established link between vehicle speed and accident severity. Road type further differentiates the severity of traffic incidents.
As indicated in Table 2, intersections are disproportionately represented in severe injuries and fatalities compared to straight roads or curves. While straight roads exhibit the highest frequency of minor injuries, intersections produce higher numbers of both severe injuries and deaths, suggesting that the complexity of maneuvering, right-of-way conflicts, and potential signal violations contribute significantly to accident severity.
The chi-square test confirms that intersections pose elevated risks, particularly for serious outcomes ($p < 0.05$). The influence of signalization is highlighted in Table 3.
Although traffic lights account for the greatest absolute number of accidents due to their widespread presence, intersections controlled by flashing lights or lacking signals entirely show higher proportions of severe injuries and fatalities. This indicates that inadequate or ambiguous signalization increases the likelihood of severe crashes, as drivers may misinterpret right-of-way or fail to anticipate conflicting traffic movements. The chi-square test validates these findings, showing that flashing and non-signalized intersections are statistically more hazardous ($p < 0.05$).
Finally, Table 4 demonstrates the pivotal role of vehicle type in shaping accident severity. Motorcycles stand out as the most vulnerable group, with the highest counts of both severe injuries (2,380) and fatalities (110). In contrast, cars, despite being involved in a comparable number of minor accidents, result in significantly fewer fatalities, reflecting greater structural protection.
Trucks and buses, while less frequently involved in accidents overall, show a relatively high fatality rate relative to their exposure, likely due to their size and collision impact. The chi-square test confirms that motorcycles carry the greatest injury burden, underscoring their disproportionate vulnerability on the road ($p < 0.05$). These results highlight that accident severity is not evenly distributed across traffic environments or vehicle categories. Higher speed limits, intersections, insufficient signalization, and motorcycle use are consistently associated with more severe outcomes.
These findings underscore the importance of targeted interventions, including stricter speed management, safer intersection design, improved signal systems, and protective measures for vulnerable road users such as motorcyclists.

**Table 1** Road Speed Limit vs Accident Severity

| Speed Limit (km/h) | Minor Injury | Severe Injury | Fatality |
|---|---|---|---|
| 30-40 | 1200 | 340 | 15 |
| 50 | 8650 | 2120 | 80 |
| 60-70 | 2900 | 640 | 32 |
| 80+ | 700 | 250 | 20 |

Chi-square test indicates significant relationship ($p < 0.05$).

**Table 2** Road Type vs Accident Severity

| Road Type | Minor Injury | Severe Injury | Fatality |
|---|---|---|---|
| Straight Road | 9200 | 1450 | 45 |
| Intersection | 7800 | 1920 | 75 |

| | | | |
|---|---|---|---|
| Curved Road | 1500 | 620 | 28 |

Chi-square test confirms intersections are more associated with severe injuries ($p < 0.05$).

**Table 3** Signal Type vs Accident Severity

| Signal Type | Minor Injury | Severe Injury | Fatality |
|---|---|---|---|
| Traffic Light | 8400 | 1680 | 60 |
| Flashing Light | 2100 | 720 | 35 |
| No Signal | 1300 | 640 | 40 |

Chi-square test shows flashing and no-signal intersections have higher severe injury rates ($p < 0.05$).

**Table 4** Vehicle Type vs Accident Severity

| Vehicle Type | Minor Injury | Severe Injury | Fatality |
|---|---|---|---|
| Motorcycle | 9600 | 2380 | 110 |
| Car | 8200 | 1460 | 50 |
| Truck/Bus | 750 | 420 | 30 |

Chi-square test confirms motorcycles have the highest severe injury and fatality rates ($p < 0.05$).

### 5.2 Correlation Analysis

To examine the relationships between the variables, a Pearson correlation analysis was conducted. The results are summarized in Table 5.

**Table 5** Correlation Analysis Results Between Variables

| Variables | Number of Accidents | Accident Severity | Driver Behavior | Road Attributes | Economic Factors |
|---|---|---|---|---|---|
| Number of Accidents | 1 | 0.65 | 0.58 | 0.45 | 0.37 |
| Accident Severity | 0.65 | 1 | 0.70 | 0.52 | 0.42 |
| Driver Behavior | 0.58 | 0.70 | 1 | 0.48 | 0.36 |
| Road Attributes | 0.45 | 0.52 | 0.48 | 1 | 0.41 |
| Economic Factors | 0.37 | 0.42 | 0.36 | 0.41 | 1 |

The results indicate that there is a strong positive correlation between accident severity and driver behavior ($r = 0.70$), suggesting that more severe accidents are associated with riskier driving behaviors.
Additionally, accident numbers are significantly correlated with accident severity ($r = 0.65$) and driving behavior ($r = 0.58$), indicating that areas with higher accident frequencies tend to experience more severe incidents.

### 5.3 Multiple Regression Analysis

To further examine the impact of various factors on traffic accidents, a multiple regression analysis was conducted. The dependent variable was the accident severity (A1 and A2), while the independent variables included driver behavior, road attributes, and economic factors. The results of the regression analysis are presented in Table 6.

**Table 6** Results of Multiple Regression Analysis

| Independent Variable | (β) | t Value | p Value | Dependent Variable |
|---|---|---|---|---|
| Driver Behavior | 0.35 | 5.23 | 0.000 | Injuries |
| Road Attributes | 0.25 | 3.45 | 0.001 | Injuries |
| Economic Factors | 0.15 | 2.30 | 0.022 | Injuries |

The regression results indicate that driver behavior has the most significant impact on accident severity ($\beta = 0.35$, $p < 0.001$), followed by road attributes ($\beta = 0.25$, $p < 0.001$) and economic factors ($\beta = 0.15$, $p < 0.05$).
These findings support the hypothesis that human factors, road conditions, and economic disparities significantly influence traffic accidents in Taichung City.
Finally, this study examined the effects of law enforcement interventions on economic, age, behavioral, alcohol-driving, road, and environmental factors, examining their impact on traffic fatalities and injuries. The analysis revealed that different factors contributed to differences in the number of deaths and injuries within 24 hours, highlighting that injuries were more susceptible to multiple factors than fatalities.
Regarding the economy and law enforcement, the results indicate that these factors do not significantly impact fatalities, but do reach a significant level for injuries ($p < 0.05$), indicating that the severity of injuries in accidents is more readily moderated by the economic environment and law enforcement intensity. Secondly, the relationship between age structure and accident outcomes also shows significant differences. While there are no significant differences across age

groups in fatalities ($p > 0.05$), the effects are significant for injuries ($p < 0.05$).
The impact is particularly pronounced among those aged 19 to 25 and those aged 66 and over, indicating that these two groups are core accident risk groups. Further examining behavioral factors, this study found that behavioral factors did not significantly impact fatalities, but were statistically significant for injuries. This suggests that while driver violations or misconduct are not sufficient to immediately increase fatalities, they do directly increase injury risk. A similar pattern is observed for the relationship between alcohol-driving and law enforcement.
While there are no significant differences in fatalities, the effects are significant for injuries, indicating that alcohol-driving and law enforcement intervention significantly influence the severity of injuries in accidents. As for the relationship between road factors and accident outcomes, the research results indicate that road conditions did not produce significant differences in fatalities, but reached a significant level in injury cases, indicating that road structure and usage patterns will affect accident outcomes. Environmental factors also showed the same trend, showing no significance in fatalities, but showing a significant impact in injury cases, which means that weather and environmental conditions mainly affect the severity of accident injuries rather than immediate death.
The above results show that while the number of fatalities showed no significant differences across most factors, the number of injuries was significantly affected by multiple factors, including economic status, age, behavior, alcohol-related driving, road conditions, and the environment. In other words, the factors influencing injuries are broad and statistically significant, while the factors influencing fatalities are relatively limited. This result suggests that if we want to effectively alleviate the social burden of traffic accidents, we should prioritize policies that reduce injuries.
Strengthening law enforcement, improving environmental conditions, and targeting high-risk groups with education and regulation will be more effective than simply focusing on reducing fatalities.

**Table 7** Results of the Hypothesis

| Factor | Predictor | β | t-value | p-value | Dependent Variable |
|---|---|---|---|---|---|
| Economy & Enforcement | Enforcement actions | 0.362 | 18.894 | 0.000 | Injuries |
| | Economy | 0.464 | 24.215 | 0.000 | Injuries |
| Age & Enforcement | Young drivers (19–25) | 0.315 | 41.575 | 0.000 | Injuries |
| | Elderly drivers (66+) | 0.109 | 25.900 | 0.000 | Injuries |
| Behavior & Enforcement | Enforcement actions | 0.212 | 10.330 | 0.000 | Injuries |
| | Risky behavior | 0.516 | 25.117 | 0.000 | Injuries |
| Drunk Driving & Enforcement | Enforcement actions | 0.247 | 11.901 | 0.000 | Injuries |
| | Drunk driving | 0.478 | 22.969 | 0.000 | Injuries |
| Road & Enforcement | Enforcement actions | 0.196 | 9.542 | 0.000 | Injuries |
| | Road conditions | 0.529 | 25.678 | 0.000 | Injuries |
| Environment & Enforcement | Enforcement actions | 0.187 | 9.109 | 0.000 | Injuries |
| | Environmental factors | 0.538 | 26.237 | 0.000 | Injuries |

This study reveals from the statistical chart in 4.2 and the regional and factor summary table 7 above that accidents primarily occur on urban roads. According to Taiwan region's road traffic safety regulations, the speed limit on urban roads is 50 km/h. Therefore, most accidents occur at speeds of 50 km/h. Most accidents occur at three-way and four-way intersections on straight roads. Unsigned intersections are the most common location for accidents, and the primary type of collision is side collisions.
The top two causes of accidents in each zone are failure to yield and failure to pay attention to the vehicle ahead, followed by failure to maintain a safe distance. It is speculated that these accidents are primarily caused by failure to yield at unsigned intersections, resulting in side collisions, and failure to pay attention and maintain a safe distance, resulting in rear-end collisions.
Next, based on the results of variance analysis and regression analysis, this study found that the main human factor in Taichung City is failure to yield as required, resulting in traffic accident fatalities. The age range is approximately 19 to 25 years old. Drunk driving and protective equipment significantly affect deaths and injuries. In terms of roads, driving roads and weather conditions significantly affect deaths and injuries, with the road environment primarily affecting injuries. In terms of economy, higher incomes are more likely to cause A2 accidents, and the incidence of A1 accidents is less affected. In terms of law enforcement, it can be seen that law enforcement can suppress the incidence of A2 accidents.
Finally, based on the above statistics and analysis, whether or not a driver has a driver's license or drives a vehicle does not have a significant impact on accidents. Failure to yield at unsigned intersections is the main factor, and inexperienced drivers cause the most accidents. Protective equipment and drunk driving directly affect traffic accident deaths and injuries. In poor weather and road conditions, drivers drive too fast and cause accidents. In terms of economy, higher incomes have a greater impact on A2 accidents. With the intervention of law enforcement, A1 and A2 accidents can be reduced.

## 6 CONCLUSIONS

Traffic accidents have long been a significant issue in Taiwan region's public safety. Therefore, this study first investigates the challenges posed by the persistently high number of traffic accidents in Taichung City. Secondly, it tests hypotheses linking driver behavior to accident causation. Based on the results of variance analysis and regression

analysis, this study returns to the hypotheses established in Chapter 3 to verify the data.

Furthermore, the study provides specific policy recommendations. The findings can assist local and central governments in promoting continuous improvement in education, engineering, and law enforcement measures to enhance road safety and reduce the fatalities, injuries, and social costs associated with traffic accidents.

First, Hypothesis 1 asserts that "human factors directly influence the occurrence of Category A1 and A2 accidents." This study tested this using four indicators: driving behavior, age, use of protective equipment, and alcohol-driving status. The analysis revealed that the primary human factor in Taichung City traffic accidents is the violation of "failure to yield"; in terms of age, young people aged 19 to 25 are the primary cause of accidents; in terms of safety equipment, whether or not a helmet or seatbelt is worn has a direct impact on fatality and injury outcomes; and in terms of alcohol-driving status, higher breathalyzer levels are associated with higher accident fatality and injury rates. Overall, Hypothesis 1 receives empirical support.

Secondly, Hypothesis 2 states that "accident severity varies depending on location." Using a variance test to account for factors such as road condition, road type, and weather conditions, the study results indicate that road type and weather conditions have a direct impact on accident fatalities and injuries, while road condition primarily influences injury severity. In other words, different accident locations indeed lead to differences in accident severity, thus supporting Hypothesis 2.

Furthermore, Hypothesis 3 proposes that "income and the administrative region in which people live have a significant impact on accidents." Because the individual incomes of accident victims are difficult to obtain, this study used the average salary of each administrative region as a proxy variable and conducted a regression analysis. The results showed that regional income levels are indeed significantly correlated with accident injury rates. Administrative regions with higher incomes have relatively higher accident rates and are more likely to result in injuries, thus supporting Hypothesis 3.

Finally, Hypothesis 4 states that "with an increase in the number of enforcement actions, deaths and injuries from human-related, road-related, and economic accidents will decrease." Regression analysis results confirm that with the intervention of law enforcement, previously insignificant factors become significant, demonstrating that law enforcement can effectively reduce traffic accident deaths and injuries. Therefore, Hypothesis 4 is also supported.

Based on the results of the hypothesis testing, we can conclude that among the human factors contributing to accidents in Taichung City, failure to yield is the primary cause, with the majority of perpetrators aged between 19 and 25. The use of protective equipment and drunk driving directly influence accident fatalities and injuries. This result is consistent with existing literature, which also indicates that driving behavior significantly influences accident occurrence, with the high-risk age group also concentrated between 18 and 26 years old, and that both protective equipment and drunk driving influence accident fatalities and injuries.

Regarding environmental factors, research shows that accidents are most common on urban roads, but the severity of accidents on highways is significantly higher, presumably due to differences in speed limits. Regarding weather, most accidents occur on sunny days, but fatalities and injuries are significantly higher in inclement weather, likely due to drivers taking risks in adverse conditions, leading to more severe accidents. Regarding road conditions, dry roads primarily influence injury severity, demonstrating that driving behavior remains a key risk factor even under normal conditions.

Regarding economic factors, the study found that higher-income districts tend to have more Class A2 accidents, meaning minor or serious injuries. These finding echoes literature showing that with increasing economic development, accident severity tends to decrease, but injury rates tend to increase. Regarding law enforcement, the study found that strengthening enforcement measures can indeed reduce accident rates and effectively reduce fatalities and injuries, demonstrating the high effectiveness of law enforcement in traffic safety management.

In terms of educational communication channels, the study found that accidents in Taichung City mainly occurred in the urban area on 50km/h roads, three-way and four-way forks, and at "undignified intersections". The main collision type was side collision, and the key causes were "failure to give way as required" and "failure to pay attention to the status of the vehicle in front/failure to maintain a safe distance." In terms of human factors, failure to yield was the primary violation, and the age of those involved in accidents was concentrated between 19 and 25 years old. Safety equipment and drunk driving directly affected casualties, and accidents could be effectively suppressed with the intervention of law enforcement. The above is the basis for the target audience, context and message design of educational promotion.

Based on this, this study proposes specific educational promotional approaches: First, targeted messaging. For motorcyclists aged 19-25, a 15-second video will be produced to educate motorcyclists about lateral blind spots and their costs, using the core message "Three steps for unsignaled intersections: Slow down for 5 seconds—stop, yield—look up." Actionable steps for "stop, look, yield" will then be provided. For senior motorcyclists and pedestrians aged 66 and above, high-contrast illustrations and case studies will be used to teach the "mutual yielding order" and the principle of "slowing down a half step and taking an extra look." For school commuters, the motto for "low-speed yielding around campus perimeters" will be promoted: "Look—yield—go." For commercial truck drivers, animated warnings will be used to illustrate inner wheel gaps and blind spots, quantifying safe distances (lateral clearance $\geq$ 1.5m, headway $\geq$ 3 seconds), and providing a three-minute pre-shift checklist.

Second, contextual triggering. At "unsigned intersections" within a 500-meter radius of accident hotspots, floor stickers and small flaggers are installed, and electronic billboards are installed to dynamically scroll the "Sideways - Please Give Way" reminder based on queue and speed detection. A navigation app and the city government's official LINE account are integrated to pop up a "Give way here" card when approaching hotspots. Reflective equipment and high-visibility

rain gear are promoted at night to form a fixed reminder of "Three things to do when returning home at night: slow down, stop, and look up."
Third, multi-channel editing and broadcasting. Official Facebook/IG/YouTube Shorts and LINE OA platforms will release a weekly series of "Intersection Yield" cards and short videos. "Yield Simulation" micro-lessons will be introduced in high schools, vocational schools, universities of science and technology, and driving schools. "Three-Minute Safety Reminders" will be introduced during morning meetings in workplaces like delivery, logistics, and construction. "Yield" stickers with accompanying safety inspection cards will be posted at motorcycle shops and gas stations. Fourth, social norms and incentives will be implemented. A monthly ranking of "Intersection Yield Rates" will be published for each administrative district, along with a #YieldFirst short video challenge to foster a "everyone yields" norm. Completing an online quiz or uploading a video of yielding will earn you a digital badge and discounts at local businesses. Campus clubs and community volunteers will collaborate on a "Weekly Intersection Action."
Fifth, collaboration with law enforcement. The cadence of educational messaging aligns with the key enforcement schedule at intersections, with intensive delivery a week before the project month launch, mobile inspections to maintain visibility during the mid-term, and feedback at the end of the term to reinforce the behavioral anchoring effect of "education + enforcement." This research has proven that enforcement can significantly reduce accident casualties, and coordinated action can amplify marginal benefits.

## 7 RECOMMENDATIONS

This study further proposes recommendations based on its analysis. First, while the publicly available government data used in this study includes law enforcement records, it does not distinguish between specific types of violations. To enhance the analytical precision of subsequent research, it is recommended that future studies provide a breakdown of specific violations, such as speeding, running red lights, and illegal lane changes. This would allow researchers to more deeply explore the relationship between specific violations and accident fatalities and injuries.
Secondly, regarding driving behavior, drivers should prioritize safe driving, especially maintaining a safe distance from the vehicle ahead. This allows for adequate reaction time in emergency situations and reduces the likelihood of rear-end collisions. This is also a key purpose of implementing technology-based enforcement on roads: by monitoring high-risk areas for rear-end collisions, the accident rate can be reduced. Furthermore, drivers should obey traffic rules, especially in poor weather conditions. It is recommended that dynamic speed limit signs be installed to remind drivers to adjust their speed based on current weather and road conditions, rather than simply following a preset speed limit.
Third, improving infrastructure is also essential. Road maintenance and inspection should be strengthened to ensure smooth road surfaces and clear traffic signs, thereby reducing accidents caused by poor road conditions. Furthermore, intersection signs should be rationally designed, with clear locations and distinct signals, and their timing adjusted according to traffic flow to reduce congestion and minimize the risk of accidents.
Fourth, regarding the correlation between economic factors and accidents, research shows that accident rates are higher in areas with higher incomes. Therefore, it is recommended to increase fines and strengthen enforcement to raise the cost of violations and compel drivers to comply with traffic rules. Specific measures could include installing surveillance cameras and automatic detection systems to immediately crack down on behaviors such as speeding, running red lights, and illegal lane changes. Furthermore, increasing penalties and criminal penalties for serious violations could effectively reduce violation rates.
Finally, it is recommended that traffic authorities integrate "human safety," "engineering design," and "strict law enforcement," and complement technological enforcement with public education to comprehensively enhance road safety. Only by adopting such an integrated approach can we effectively reduce deaths and injuries caused by traffic accidents and achieve the overall goal of improving traffic safety.

## COMPETING INTERESTS

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

## REFERENCES

[1] Reason J. Human error. Cambridge University Press, 1990.
[2] Allahyari T, Saraji GN, Adi J, et al. Cognitive failures, driving errors and driving accidents. International Journal of Occupational Safety and Ergonomics, 2008, 14(2):149-158.
[3] Bishai D, Quresh A, James P, et al. National road casualties and economic development. Health Economics, 2006, 15(1):65-81.
[4] Bougueroua M, Carnis L. Economic development, mobility and traffic accidents in Algeria. Accident Analysis & Prevention, 2016, 92:168-174.
[5] Van Beeck EF, Borsboom GJ, Mackenbach JP. Economic development and traffic accident mortality. International Journal of Epidemiology, 2000, 29(3):503-509.
[6] Klaitman SS, Solomonov E, Yaloz A, et al. The incidence of road traffic crashes among young people aged 15-20 years: differences in behavior, lifestyle and sociodemographic indices in the Galilee and the Golan. Frontiers in Public Health, 2018, 6:202.

[7] M'bailara K, Atzeni T, Contrand B, et al. Emotional reactivity: Beware its involvement in traffic accidents. Psychiatry Research, 2018, 262:290-294.
[8] Rahmadiyani R. Factors affecting fatigue driving: demographics, attitudes, and perceived barriers. Journal of Transportation Safety, 2023, 15(1):20-35.
[9] Akinyemi Y. Relationship between economic development and road traffic crashes and casualties: empirical evidence from Nigeria. Transportation Research Procedia, 2020, 48:218-232.
[10] Agyemang B, Semevoh R. Regression analysis of road traffic accidents and population growth in Ghana. International Journal of Business and Social Research, 2013, 3(10):41-47.
[11] Gomes SV. The influence of the infrastructure characteristics in urban road accidents occurrence. Accident Analysis & Prevention, 2013, 60:289-297.
[12] Dumitrascu DI. Influence of road infrastructure design over the traffic accidents: a simulated case study. Infrastructures, 2024, 9(9):154.
[13] Pai CW, Saleh W. An analysis of motorcyclist injury severity in traffic crashes at T-junctions. Accident Analysis and Prevention, 2007, 39(6):1197-1207.
[14] Strawderman L, Huang SH, Jing L. Sign saturation and driver behavior: a study on warning effectiveness. Transportation Research Record, 2005, 1937(1):49-56.
[15] Girotto E, de Andrade SM, González AD, et al. Professional experience and traffic accidents/near-miss accidents among truck drivers. Accident Analysis & Prevention, 2016, 95:299-304.
[16] Lawton BP, Hallmark SL, Basulto-Elias G, et al. Road intersections and crash frequency in Ghana. Journal of Safety Research, 2024, 80:45-57.
[17] Prayudyanto MN, Goeritno A, Al Ikhsan SH, et al. Designing a model of the early warning system on the road curvature to prevent the traffic accidents. International Journal of Safety and Security Engineering, 2022, 12(3):291-298.
[18] Dragutinovic N, Twisk D. The effectiveness of road safety education. Accident Analysis and Prevention, 2006, 38(1):25-34.
[19] Twisk DA, Vlakveld WP, Commandeur JJ, et al. Five road safety education programmes for young adolescent pedestrians and cyclists: a multi-programme evaluation in a field setting. Accident Analysis & Prevention, 2014, 66:55-61.
[20] Qaid H, Widyanti A, Salma SA, et al. Speed choice and speeding behavior on Indonesian highways: extending the theory of planned behavior. IATSS Research, 2022, 46(2):193-199.
[21] Lin DJ, Yang JR, Liu HH, et al. Analysis of environmental factors on intersection accidents. Sustainability, 2022, 14(3):1764.
[22] Goniewicz K, Goniewicz M, Pawłowski W, et al. Road accident rates: strategies and programmes for improving road traffic safety. European Journal of Trauma and Emergency Surgery, 2016, 42(4):433-438.
[23] Lin HA, Chan CW, Wiratama BS, et al. Evaluating the effect of drunk driving on fatal injuries among vulnerable road users in Taiwan region: a population-based study. BMC Public Health, 2022, 22(1):2059.
[24] Shin D, Washington S, van Schalkwyk I. The impact of traffic enforcement on driver behavior. Accident Analysis and Prevention, 2015, 82:13-22.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3018**

# FACTORS INFLUENCING PURCHASE INTENTIONS IN THE PACKAGING DESIGN OF H CATERING COMPANY'S SELF-HEATING HOTPOT

Yuan Lin, WenChao Pan*, Ran Liu, HaiLin Wang
*School of Management, Guangzhou University of Chinese Business, Guangzhou 510006, Guangdong, China.*
*Corresponding Author: WenChao Pa, Email: teacherp0162@126.com*

**Abstract:** To investigate the causal mechanism linking H Catering Company's self-heating hotpot packaging design to consumer purchase intent and enhance product market competitiveness, this study employs empirical research based on 302 valid questionnaire responses, incorporating grey association analysis. Findings reveal that corporate employees aged 41 and above constitute the core consumer demographic, exhibiting highest satisfaction with packaging functionality and practicality, followed by visual elements and material safety. Grey relational analysis further confirmed the ranking of packaging design dimensions influencing purchase intent: functional utility > visual elements > material safety. This indicates functional utility as the core factor in consumer decision-making. Consequently, three-dimensional optimisation recommendations are proposed: Functional Utility Dimension: Design foldable stacking structures, concealed handles, and colour-changing water level indicators for diverse scenarios to enhance operational convenience. For visual elements, establish a differentiated visual system for flavour variations while enhancing information transmission efficiency and emotional resonance. Regarding material safety, upgrade food-grade inner packaging and puncture-resistant outer bags for heating packs, alongside improving safety guarantees for reusable materials to balance security and environmental sustainability. This research provides theoretical and practical references for optimising catering packaging design and boosting purchase intent.
**Keywords:** Grey relational analysis; Packaging design; Purchase intention

## 1 INTRODUCTION

Driven by consumption upgrades and accelerated lifestyles, self-heating hotpots have become a new favourite among younger demographics. According to the 2024 Convenience Food Industry Panorama Insight (https://www.fxbaogao.com/report?id=4772878), rising urbanisation rates have heightened demands for premium consumer goods, particularly in major cities where consumers expect greater product diversity and quality.Within the convenient food consumer base, women constitute 55% of the market. This substantial female consumer demographic, particularly dominant in online spending, exerts significant influence over the convenient food sector. Crucially, heightened health consciousness has deepened public emphasis on wholesome diets, compelling the industry to respond by offering more healthful yet convenient options.

Packaging design plays an indispensable role in the convenience food sector, fulfilling fundamental functions of practicality and safety while often exceeding basic requirements. It also serves as a key vehicle for conveying brand philosophy and guiding purchasing decisions, needing to capture consumer attention swiftly. H Catering Company, a major player in the self-heating hotpot segment, possesses flavour advantages but has seen slowing market share growth in recent years. Consumers, confronted with unchanging packaging designs, struggle to develop new or increased purchasing intent.

Therefore, this study examines H Company's self-heating hotpot packaging across three dimensions: visual appeal, functionality, and informational content. Employing questionnaire surveys, literature review, and data analysis, it assesses the extent to which these elements influence consumer purchasing intent. The literature review synthesises theoretical findings on packaging design and consumer purchasing behaviour, establishing the study's theoretical foundation.Questionnaires were employed to gather consumer perceptions of H Company's packaging and purchase intent data, ensuring the objectivity and authenticity of research findings. Data analysis methods were applied to conduct in-depth exploration of survey data, precisely quantifying the impact of each dimension.

## 2 REVIEW OF RELEVANT RESEARCH LITERATURE

Firstly, during the early rise of convenient self-heating foods, the study Convenient Self-Heating Foods Release Huge Consumer Demand indicated that self-heating foods, represented by self-heating hotpots, achieved annual growth exceeding 20%. Post-80s and Post-90s generations form the primary consumer base, with regional speciality cuisines providing the industrial foundation for market expansion.Industry participants are diverse, with channel integration and innovation emerging as key trends. However, current challenges include limited flavour variety and product homogenisation, necessitating corporate adaptation to market shifts through innovation to overcome developmental bottlenecks [1]. These issues remain critical challenges for the self-heating convenience food sector.

Numerous scholars have identified packaging design as playing a decisive role within the self-heating food sector. In his

systematic study of customer satisfaction, Zhao Baoshan introduced grey relational theory to analyse the influence mechanism of packaging design. By constructing a relational model linking packaging design elements to customer perception and ultimately satisfaction, he quantitatively validated the relative weight of different packaging dimensions on consumer decision-making [2]. Xie Jie focused on the practical and aesthetic values of packaging design, examining their influence on consumer purchasing behaviour through multiple dimensions including historical development, functional classification, positioning design, and sales methods. The study emphasised that these values are mutually reinforcing and interdependent: practicality forms the foundation while aesthetics is indispensable. Overemphasising either is inadvisable, and their integration holds significant importance for modern packaging design, being key to creating outstanding packaging [3] .

Amidst fierce competition in the ready-to-eat convenience food sector, creating differentiated products has become a direct competitive strategy. Such differentiation requires consideration and design across multiple dimensions. Zhang Dalu and Tang Lanling contend that the rise of the singles economy has spurred innovation in Chongqing hotpot packaging, noting current market offerings suffer from homogenisation and lack of individuality.Innovative designs exhibit compactness, convenience, and contemporary styles, while also addressing emotional resonance, social interaction, and personalisation needs. Existing examples include hotpot cups and self-heating hotpot sets, with future development requiring continuous exploration aligned with consumer demands [4]. Deng Wenzhi, Xu Juanfang, and Liu Zhiqi note that self-heating foods have gained traction due to the rise of the stay-at-home economy and single-person households, yet their packaging suffers from homogeneity, weak brand recognition, and issues concerning safety and container design.Taking Chongqing self-heating hotpot as a case study, this article examines the translation and application of regional cultural symbols. By extracting elements from form and structure, it proposes application strategies within and across domains to enhance brand competitiveness, consumer experience, and disseminate regional culture [5].

With societal advancement, increasing enterprises are fulfilling social responsibilities, where green packaging represents a fundamental pathway for promoting circular development. Xiao Mengyun investigated the impact of green packaging cues on consumer purchase intentions through four experiments. Findings revealed that both green food labels and eco-friendly packaging materials positively influence purchase intentions, mediated by perceived functional value and perceived green value, with health and ecological motivations acting as moderators [6]. Li Yan examined the impact of sustainable packaging on health food purchase intentions, testing hypotheses through two online experiments grounded in the halo effect and dual-processing model. Results indicated that packaging sustainability positively influenced purchase intentions, mediated by perceived quality, with health-conscious eating tendencies moderating this effect—consumers exhibiting such tendencies demonstrated higher purchase intentions[7].

This study examines factors influencing purchase intent for H Catering Company's self-heating hotpot packaging design, analysing how different packaging dimensions affect consumer purchasing decisions. Employing grey management degree analysis, it clearly quantifies the correlation between various dimensional factors and purchase intent. Li Tingting focused on housing demands among the elderly in Harbin's ageing population, employing grey relational analysis as the core analytical tool. Its significance lies in the precise selection of six key indicators, incorporating statistical data from 2003 to 2014 to quantify the influence of each factor. This study identified per capita housing floor space as the core influencing factor, providing scientific support for housing demand research and offering critical evidence for government policy formulation and developer planning [8]. Lanshuang addressed the multi-factor and information-impaired nature of packaging design proposals, highlighting the limitations of the Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation Method while underscoring the significance of grey relational analysis. Its approach is characterised by clarity and minimal computational demands. Through steps such as sequence construction and dimensionless transformation, it enables precise ranking of multiple proposals. Validated through beverage packaging case studies, it provides a scientifically reliable analytical tool for optimising packaging design solutions [9]. Liao Chao-xiong focused on factors influencing construction costs, highlighting the core value of grey relational analysis. Suited to engineering scenarios with incomplete information, it enables quantitative assessment of each factor's impact.Using a shantytown redevelopment project as a case study, by constructing models and processing deducted data, it precisely determines the correlation between factors such as design and drawings. This resolves the difficulty of weighting in traditional qualitative analysis, providing scientific and reliable decision support for cost control in government-funded projects [10].

## 3 RESEARCH METHODOLOGY

### 3.1 Literature Review Method

Literature review serves as the theoretical foundation, systematically reviewing relevant literature to clarify the core elements of the research framework and variable relationships.The study collected literature using grey relational analysis, packaging design, and purchase intention as core keywords. Through summarising, reviewing, and integrating the literature, the core variables of this research were identified. A theoretical analytical framework was constructed to examine the factors influencing purchase intention in H Catering Company's self-heating hotpot packaging design. Based on this, research hypotheses were proposed, laying the theoretical foundation for subsequent investigations.

### 3.2 Questionnaire Survey Method

Through the collection of questionnaires, data regarding consumers' perceived evaluations of H Catering Company's self-heating hotpot packaging and their purchasing intentions were gathered, ensuring the objectivity and generalisability of the research conclusions. The questionnaire comprised three sections: basic sample information, core variable measurement items, and open-ended questions. Distributed via online channels, 302 valid responses were collected, meeting the statistical analysis requirements for empirical research.This dataset comprehensively covers key research information including consumer characteristics, packaging perceptions, and consumption tendencies. It provides robust and reliable data support for subsequent grey correlation analysis, exploring the strength of associations between various packaging design dimensions and purchase intent, and formulating targeted optimisation recommendations.

### 3.3 Data Analysis Methods

This study primarily employed descriptive analysis and grey correlation analysis, utilising SPSS 26.0 statistical software for data processing.

Descriptive analysis was employed to organise fundamental sample information, including gender and age distribution statistics, thereby verifying sample representativeness. For measurement items concerning packaging design dimensions and purchase intent, metrics such as means and frequencies were calculated. This enabled intuitive presentation of consumers' perceived levels regarding various elements of H Catering Company's self-heating hotpot packaging and the intensity of their purchase intent. Sample representativeness was further validated by benchmarking against target consumer group characteristics.For measurement items and purchase intention questions across dimensions including packaging functionality, visual elements, and material safety, data was presented through charts combined with text. This visually demonstrated consumer perception differences, overall evaluation levels, and purchase intention strength for each packaging element, establishing a robust data foundation for subsequent in-depth investigations such as grey correlation analysis.

Grey correlation analysis, abbreviated as grey correlation, refers to the uncertain relationships between phenomena or between systemic factors and primary behaviours. When employing grey correlation analysis for comprehensive design evaluation, one must first establish reference sequences and comparison sequences. Subsequently, the importance of various influencing factors or alternative proposals is determined by calculating the correlation coefficients and degrees of association between comparison sequences and reference sequences. Finally, alternative proposals can be ranked according to their relative merits based on the magnitude of their association degrees.Grey Relational Analysis addresses the fuzzy characteristics of influence relationships between consumer perceptions and purchase intentions. It quantifies the degree of association between various dimensions of H Catering Company's self-heating hotpot packaging design and consumer purchase intentions. By ranking these dimensions according to their degree of association, it clarifies the primary and secondary order of influence on purchase intentions, providing a reference basis for optimising the packaging design.

## 4 DESCRIPTIVE STATISTICAL ANALYSIS

### 4.1 Gender Descriptive Statistics

The consumer experience survey regarding factors influencing purchase intent for H Catering Company's self-heating hotpot packaging design yielded 302 valid questionnaires. Statistical results for the gender variable indicate the sample distribution possesses good representativeness.Specifically, female consumers numbered 144, accounting for 47.7% of the total sample; male consumers numbered 158, accounting for 53.2% of the total sample. No samples lacked gender information. In terms of distribution characteristics, the proportion of male samples was slightly higher than that of female samples, consistent with the characteristic that male groups are more inclined to consume self-heating hotpot products in H Catering Company's actual consumption scenarios. As shown in Table 1.

**Table 1** Descriptive Statistics Analysis by Gender

| Your Gender | | | | |
|---|---|---|---|---|
| | Frequency | Percentage | Valid Percentage | Cumulative Percentage |
| Male | 158 | 52.3 | 52.3 | 52.3 |
| ValidFemale | 144 | 47.7 | 47.7 | 100.0 |
| Total | 302 | 100.0 | 100.0 | |

### 4.2 Age Descriptive Statistics

The age data statistics are as follows: those aged 41 and above account for 28.1%, followed by the 18-23 age group at 26.5%, the 24-30 age group at 25.5%, and the 31-40 age group at the lowest proportion, 19.9%.No samples lacked age information. In terms of distribution characteristics, the 41 and above age group was the largest, followed by the younger 18-23 age group. As shown in Table 2.

**Table 2** Descriptive Statistics for Age

| Your age |
|---|

| | | Frequency | Percentage | Valid Percentage | Cumulative Percentage |
|---|---|---|---|---|---|
| Valid | 18–23 years | 80 | 26.5 | 26.5 | 26.5 |
| | 24–30 years old | 77 | 25.5 | 25.5 | 52.0 |
| | 31–40 years old | 60 | 19.9 | 19.9 | 71.9 |
| | 41 years and above | 85 | 28.1 | 28.1 | 100.0 |
| | Total | 302 | 100.0 | 100.0 | |

### 4.3 Occupational Descriptive Statistical Analysis

Occupational data statistics are as follows: students (17 cases, 5.6%); self-employed individuals (39 cases, 12.9%); government employees (21 cases, 7%); corporate employees (201 cases, 66.6%); medical institution staff (13 cases, 4.3%); and educational institution personnel (11 cases, 3.6%).The data indicates that corporate employees constitute the largest proportion, with 201 samples, suggesting greater willingness among this group to consume self-heating packaged instant products. Occupations within educational institutions represent a smaller proportion. As shown in Table 3.

**Table 3** Descriptive Statistical Analysis by Occupation

| Your Occupation | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percentage | Valid Percentage | Cumulative Percentage |
| Valid | Student | 17 | 5.6 | 5.6 | 5.6 |
| | Self-employed | 39 | 12.9 | 12.9 | 18.5 |
| | Government employees | 21 | 7.0 | 7.0 | 25.5 |
| | Corporate staff | 201 | 66.6 | 66.6 | 92.1 |
| | Medical institution staff | 13 | 4.3 | 4.3 | 96.4 |
| | Educational Institution Occupations | 11 | 3.6 | 3.6 | 100.0 |
| | Total | 302 | 100.0 | 100.0 | |

### 4.4 Descriptive Statistics Analysis of Educational Attainment

The educational attainment distribution is as follows: 55 individuals (18.2%) possess secondary education or below; 110 individuals (36.4%) hold a college diploma; 105 individuals (34.8%) hold a bachelor's degree; 32 individuals (10.6%) hold a master's degree or higher; no samples lacked educational attainment information.The distribution characteristics indicate that the sample predominantly comprises individuals with college or undergraduate qualifications. This aligns closely with the educational profile of H Catering Company's core consumer base (young and middle-aged working professionals with a certain level of purchasing power). Coverage across all educational levels demonstrates that the sample possesses broad representativeness in terms of educational attainment. This supports subsequent analyses examining consumption experience differences and correlations among consumers with varying educational backgrounds. As shown in Table 4.

**Table 4** Descriptive Statistics Analysis of Educational Attainment

| Your Educational Attainment | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percentage | Valid Percentage | Cumulative Percentage |
| Valid | Secondary education and below | 55 | 18.2 | 18.2 | 18.2 |
| | Technical college | 110 | 36.4 | 36.4 | 54.6 |
| | Undergraduate | 105 | 34.8 | 34.8 | 89.4 |
| | Master's degree and above | 32 | 10.6 | 10.6 | 100.0 |
| | Total | 302 | 100.0 | 100.0 | |

### 4.5 Descriptive Statistics Analysis of Purchase Frequency

Statistical analysis of purchasing frequency for H Catering Company's Natural Hotpot reveals: Very Frequent (7+ times monthly) – 73 respondents (24.2%); Fairly Frequent (4–6 times monthly) – 67 respondents (22.2%); Moderate Frequency (2–3 times monthly) – 83 respondents (27.5%); Less Frequent (1 time monthly or less) – 79 respondents (26.2%). No missing data for this question.The distribution reveals that the combined proportion of very frequent and fairly frequent purchases stands at 46.4%, yet monthly frequency of once or less ranks second. This indicates room for improvement in marketing strategies concerning repeat purchases. As shown in Table 5.

**Table 5** Descriptive Statistics Analysis of Purchase Frequency

| Have you ever purchased Haidilao self-heating hotpot? | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percentage | Valid Percentage | Cumulative Percentage |
| Valid | Very Frequent (7 times or more per month) | 73 | 24.2 | 24.2 | 24.2 |
| | Fairly frequent (4–6 times per month) | 67 | 22.2 | 22.2 | 46.4 |
| | Moderate (2-3 times per month) | 83 | 27.5 | 27.5 | 73.8 |
| | Less frequently (once or less per month) | 79 | 26.2 | 26.2 | 100.0 |
| | Total | 302 | 100.0 | 100.0 | |

## 5 GREY CORRELATION ANALYSIS

### 5.1 Raw Data Phase

This stage collected core data from 302 valid samples, with primary indicators comprising raw scores for visual design, functional utility, and packaging design. The data covered diverse product attribute evaluation scenarios, featuring sufficient sample size and balanced distribution. It comprehensively documented initial feedback for each indicator, providing authentic and reliable foundational data for subsequent grey correlation analysis. As shown in Table 6.

**Table 6** Evaluation Indicator System for Primary Headings

| Primary Indicator | Mean | Standard Value | Minimum | Maximum Value |
|---|---|---|---|---|
| Packaging Safety | 3.82 | 0.7688 | 1.00 | 5.00 |
| Visual Design | 3.8055 | 0.7373 | 1.25 | 5.00 |
| Functional Utility | 3.7960 | 0.7300 | 1.20 | 5.00 |

### 5.2 No-Volume Hardening Phase

To eliminate interference arising from dimensional differences between indicators, the raw data underwent standardised conversion. This ensured comparability between purchase intent and primary indicator values, preventing analytical bias from differing measurement methodologies and clearing data obstacles for subsequent variance calculations and correlation analysis. As shown in Table 7.

**Table 7** Standardised Indicator System for Primary Headings

| Primary Indicator | Mean | Standard Value | Minimum | Maximum Value |
|---|---|---|---|---|
| Packaging Safety | 1.9117 | 0.3844 | 0.500 | 2.500 |
| Visual Design | 1.9027 | 0.3687 | 0.625 | 2.500 |
| Functional Utility | 1.8980 | 0.3650 | 0.600 | 2.500 |

### 5.3 Differential Sequence Stage

The mean values of the differential sequences for each indicator range between 0.211 and 0.217, indicating minimal variation. The maximum and minimum values are presented in Table 8.

**Table 8** Evaluation Indicator System for Primary Headings

| Influencing Factors | MAX | MIN |
|---|---|---|
| Functional Utility | 0.7 | 0 |
| Visual Elements | 0.8 | 0 |
| Material Safety | 0.8 | 0 |

### 5.4 Correlation Coefficient Stage

This study, based on 302 valid questionnaire responses, employed grey relational analysis with purchase intent as the reference sequence ($X_0$) and visual design ($X_1$), functional utility ($X_2$), and material safety ($X_3$) as comparative sequences. Key findings are summarised as follows:Firstly, the correlation rankings between each influencing factor dimension and purchase intent are: functional utility ($r_1$=0.6965) > visual design ($r_2$=0.6925) > material safety ($r_3$=0.6910).Secondly, packaging functionality exhibits the strongest correlation with purchase intent, indicating that H Catering's consumers are primarily influenced by functional utility. Although material safety displays the weakest correlation, it remains a significant factor affecting purchase intent. Thirdly, the sample size of 302 respondents is deemed sufficient and evenly distributed, ensuring robust stability in the correlation analysis results.In summary, this grey correlation analysis clarifies the weighting of each influencing factor dimension on purchase intent. It provides targeted decision-making support for H Catering to optimise purchase intent drivers and enhance consumer spending, while also laying the groundwork for subsequent differential correlation analyses across segmented groups (e.g., different genders, educational backgrounds). As shown in Table 9.

**Table 9** Level-1 Indicator System for Price-to-Quality Ratio

| Primary Indicator | Correlation | Standard Value | Minimum Value | Maximum Value |
|---|---|---|---|---|
| Packaging Safety | 0.6910 | 0.1718 | 0.3333 | 1.0000 |
| Visual Design | 0.6925 | 0.1649 | 0.3333 | 1.0000 |
| Functional Utility | 0.6965 | 0.1718 | 0.3333 | 1.0000 |

## 6 CONCLUSION

The data indicates a positive correlation between functional packaging design and consumer purchase intent. Respondents who rated packaging functionality more highly demonstrated relatively stronger purchase intent.Furthermore, while packaging aesthetics were not prominently featured in the initial questionnaire options, some

respondents mentioned in supplementary feedback that visually appealing packaging enhances product appeal, indirectly influencing purchasing decisions. Additionally, material safety is an integral aspect of consumer purchasing intent, representing an underlying customer requirement. Consequently, optimisation across these three dimensions is necessary to ensure alignment with evolving consumer demands.

Regarding functional utility, maintaining packaging strengths in portable storage, sealing performance, operational convenience, and clear usage instructions remains fundamental to sustaining consumer goodwill and purchase intent. Simultaneously, differentiated functional designs can be developed for distinct scenarios. For home storage, adding foldable clips to the packaging sides enables stable stacking without tipping when multiple boxes are piled.For travel scenarios, optimise the folding structure to reduce packaging thickness and incorporate a concealed handle at the top for easy insertion into backpack side pockets. For office meal replacement scenarios, add heat-insulating silicone strips to the exterior of heating containers to prevent heat damage to surfaces, alongside small condiment pouch compartments to prevent sauce spillage.Precise water-level markings are indicated on the heating pack's reaction zone, complemented by a colour-changing indicator strip that activates when the water level is correct, resolving issues of over- or under-filling. Anti-slip textures are added to the tear-open tabs of vegetable and seasoning packets, enhancing ease of opening and accommodating scenarios involving shorter nails or damp hands.

Visually, differentiated colour schemes distinguish flavours: fiery red-orange gradients with flame motifs for spicy varieties, and warm yellow gradients with tomato textures for tomato-based options. This aids rapid flavour identification, reducing uncertainty and purchase hesitation. Core selling points—such as 15-minute preparation and no-dishwashing convenience—feature prominently on packaging fronts to align with fast-paced consumers' decision-making needs.Furthermore, the inner packaging features engaging content like hotpot trivia to foster emotional connection with consumers. Catering to younger tastes, minimalist design is adopted, reducing intricate patterns to enhance brand sophistication.

Regarding material safety, the inner food packaging employs food-grade materials that are heat-resistant and microwave-safe, accommodating consumers who reheat meals. The outer heating pouch uses reinforced, puncture-resistant composite film with a damage-indicator colour strip. Should the packaging rupture, the strip turns blue, providing an immediate safety alert to prevent accidents.For consumers who enjoy collecting and repurposing items, we enhance reuse functionality by upgrading the outer box into a detachable storage container. After consumption, removing the heating components transforms it into a snack storage box. Upon product completion, recycling guidelines for materials are printed on the inner packaging, guiding consumers in proper disposal.

Therefore, marketing efforts targeting safety and trust concerns, based on confirmed information dimensions, are crucial for securing consumer confidence. Clearly communicating safety information alleviates consumer apprehensions. Only through the organic integration of these three elements can marketing optimisation precisely align research conclusions with consumer demands, maximise marketing value, and drive improved product market performance.

## COMPETING INTERESTS

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

## FUNDING

The project was supported by 2025 Guangzhou Huashang College Scientific Research Talent Cultivation Project: Research on Organizational Resilience Evolution and Social Value Reshaping in Digital Governance (Project No.: 2025HSPY19).

## REFERENCES

[1] Convenient self-heating foods unleash substantial consumer demand. Industry & City, 2018(03): 79.
[2] Zhao Baoshan, Yu Huixin, Hao Yongjing. A Systemic Study of Customer Satisfaction Based on Grey Relational Entropy. Research on Technology, Economy and Management, 2012(02): 17-21.
[3] Xie Jie. Practical and Aesthetic Value Analysis of Packaging Design. Packaging Engineering, 2014, 35(08): 85-87.
[4] Zhang Dalu, Tang Lanling. Packaging Design Innovation for Chongqing Hotpot in the Era of the Singles Economy. Food and Machinery, 2021, 37(12): 101-106.
[5] Deng Wenzhi, Xu Juanfang, Liu Zhiqi. Application Research of Regional Cultural Symbols in Packaged Ready-to-Eat Food Packaging. Design, 2023, 36(01): 4-7.
[6] Xiao Mengyun. The Influence of Green Packaging Cues on Consumers' Food Purchase Intentions. Henan University of Technology, 2024.
[7] Li Yan. The Influence of Packaging Sustainability in Health Foods on Consumer Purchase Intentions. Huazhong Agricultural University, 2023.
[8] Li Tingting. Research on Housing Consumption Demands and Behavioural Patterns Among Elderly Residents in Harbin. Harbin Institute of Technology, 2016.
[9] Lan Shuang. Application of Grey Relational Analysis in Optimising Packaging Product Design Schemes. Packaging Engineering, 2011, 32(11): 52-54.
[10] Liao Chaoxiong. Identification and Prediction of Factors Influencing Construction Costs Based on Grey Relational Analysis. China Construction Metal Structure, 2025, 24(21): 169-171.

**Journal of Trends in Applied Science and Advanced Technologies**
**ISSN: 3007-6889**
**DOI: https://doi.org/10.61784/asat3019**

# LEARNING-BASED DYNAMIC RESOURCE ALLOCATION FOR SERVERLESS COMPUTING WITH GRAPH NEURAL NETWORKS

RuiWen Zhang
*Department of Computer Science, George Washington University, Washington 20052, USA.*
*Corresponding Email: rachel.zhang@gmail.com*

**Abstract:** Serverless computing has emerged as a transformative paradigm in cloud infrastructure, offering dynamic resource provisioning and pay-per-use economics that significantly reduce operational overhead for application developers. However, the inherent challenges of serverless architectures, including unpredictable workload patterns, heterogeneous resource demands, and stringent quality-of-service requirements, necessitate intelligent resource allocation mechanisms that can adapt to rapidly changing conditions. This paper proposes a novel learning-based approach that leverages Graph Neural Networks (GNNs) to model the complex dependencies and resource relationships in serverless computing environments. Our framework captures the intricate structural patterns of function invocations, resource utilization, and inter-function dependencies through graph representations, enabling more effective resource allocation decisions. The GNN-based model employs a deep reinforcement learning architecture where an intelligent agent learns optimal policies through continuous interaction with the serverless environment. Through comprehensive experimental evaluation, we demonstrate that our approach achieves superior performance compared to traditional heuristic-based methods including Shortest Job First (SJF), Tetris, and Packer algorithms, reducing average job slowdown by approximately 40% under high load conditions while maintaining robust performance across varying workload intensities. The proposed system exhibits strong scalability with efficient training on graphs containing up to 10 million edges and demonstrates excellent generalization capabilities across diverse workload patterns.
**Keywords:** Serverless computing; Graph neural networks; Resource allocation; Deep reinforcement learning; Function as a service; Dynamic scheduling

## 1 INTRODUCTION

The evolution of cloud computing has witnessed a paradigm shift from traditional server-based infrastructure to increasingly abstract and flexible computing models. Serverless computing, particularly through Function as a Service (FaaS) platforms, represents the latest advancement in this trajectory, fundamentally transforming how applications are deployed, scaled, and managed in cloud environments [1]. This computing model abstracts infrastructure management entirely from developers, allowing them to focus exclusively on business logic while the cloud provider handles all aspects of resource provisioning, scaling, and maintenance. The serverless paradigm has gained substantial traction in recent years, with major cloud providers including Amazon Web Services Lambda, Microsoft Azure Functions, and Google Cloud Functions offering mature platforms that support diverse workloads ranging from web applications to data processing pipelines [2]. The appeal of serverless computing stems from its intrinsic advantages, including automatic scaling capabilities that dynamically adjust resources based on demand, fine-grained billing that charges only for actual execution time rather than reserved capacity, reduced operational complexity through elimination of server management responsibilities, and improved resource efficiency through better utilization of underlying infrastructure [3].

Despite these compelling benefits, serverless computing introduces unique challenges that complicate resource management and optimization efforts. The stateless nature of serverless functions necessitates careful management of state persistence and data flow between function invocations, while the event-driven execution model creates highly variable and often unpredictable workload patterns that strain traditional resource allocation mechanisms [4]. Cold start latency, which occurs when functions must be initialized before execution, represents one of the most significant performance bottlenecks in serverless systems, particularly affecting latency-sensitive applications [5]. The dynamic and bursty nature of function invocations makes capacity planning extremely difficult, as resource demands can fluctuate dramatically over short time periods. Furthermore, functions often exhibit complex dependencies on each other and on external services, creating intricate execution workflows that must be carefully orchestrated to maintain performance and consistency [6]. Multi-resource constraints, where functions require specific combinations of computational, memory, and network resources, add another layer of complexity to the allocation problem [7].

Traditional approaches to resource allocation in serverless environments typically rely on rule-based heuristics or simple reactive strategies that struggle to capture the complex dynamics of modern serverless workloads. Heuristic methods such as Shortest Job First prioritize functions based on estimated execution time, while resource packing strategies like Tetris attempt to maximize utilization through intelligent placement decisions [8]. However, these approaches lack the adaptability needed to handle the diverse and unpredictable workload patterns characteristic of

production serverless deployments. Recent advances in machine learning, particularly in deep reinforcement learning and graph neural networks, offer promising avenues for addressing these challenges through intelligent, adaptive resource management strategies [9]. Graph Neural Networks have demonstrated remarkable capabilities in modeling complex relational structures and dependencies, making them particularly well-suited for representing the intricate relationships between serverless functions, their resource requirements, and execution patterns [10].

Deep reinforcement learning provides a powerful framework for sequential decision-making under uncertainty, enabling systems to learn optimal policies through trial-and-error interaction with their environment. By formulating resource allocation as a reinforcement learning problem, we can develop agents that continuously improve their decision-making capabilities based on observed outcomes and accumulated experience [11]. The combination of GNNs for structural representation learning and deep reinforcement learning for policy optimization creates a synergistic approach that addresses both the relational complexity and sequential nature of serverless resource allocation [12]. This integration allows the system to capture dependencies between functions through graph convolutions while learning temporal allocation patterns through reinforcement learning mechanisms, resulting in policies that are both structurally aware and temporally adaptive.

This research proposes a novel framework that combines Graph Neural Networks with deep reinforcement learning to enable intelligent, learning-based resource allocation in serverless computing environments. Our approach models the serverless system as a dynamic graph structure that evolves over time, capturing function invocation patterns, resource utilization, and inter-function dependencies. The GNN component learns to extract meaningful features from this graph representation, encoding both local resource constraints and global system dynamics into compact node embeddings. These learned representations serve as input to a policy network that makes allocation decisions by sampling actions according to a learned probability distribution. The reinforcement learning framework enables the agent to optimize long-term system performance by balancing multiple objectives including response time minimization, resource utilization maximization, and cost efficiency. The primary contributions of this work include the development of a comprehensive graph-based representation for serverless computing systems that captures both structural and temporal dynamics, the design of an efficient GNN architecture specifically tailored to serverless resource allocation that scales to large deployments, the integration of deep reinforcement learning techniques that enable adaptive policy learning through environmental interaction, and an extensive experimental evaluation demonstrating significant performance improvements over existing approaches including traditional heuristics and non-graph-based learning methods.

## 2 LITERATURE REVIEW

The landscape of serverless computing research has evolved rapidly over the past several years, with substantial attention directed toward understanding its unique characteristics and addressing its inherent challenges. Early serverless platforms established the foundational principles of function-based computing and demonstrated the viability of fine-grained resource provisioning models [11]. Subsequent research has expanded our understanding of serverless architectures, investigating various aspects including performance optimization, cost management, and system design principles [12]. Recent systematic reviews have provided comprehensive surveys of the serverless computing paradigm, highlighting both its transformative potential and the technical challenges that must be addressed to realize its full benefits [13]. These surveys identify resource allocation and scheduling as critical research areas that directly impact the efficiency, performance, and cost-effectiveness of serverless deployments.

Resource allocation in serverless computing has emerged as a critical research area, with numerous studies proposing various optimization strategies and scheduling algorithms. Traditional approaches to this problem have relied primarily on heuristic-based methods that apply predefined rules to make allocation decisions based on current system state and workload characteristics [14]. The Shortest Job First heuristic prioritizes functions with shorter execution times to minimize average waiting time, while resource packing strategies inspired by bin packing algorithms attempt to maximize utilization by efficiently placing functions on available resources [15]. These heuristics often focus on specific optimization objectives such as minimizing execution time or reducing costs, but struggle to handle the multi-objective nature of real-world serverless deployments where multiple competing goals must be balanced simultaneously. Furthermore, static heuristics cannot adapt to changing workload patterns or learn from historical execution data, limiting their effectiveness in dynamic environments with evolving characteristics [16].

Recent work has begun to explore more sophisticated approaches that leverage machine learning techniques to improve resource allocation decisions. Deep reinforcement learning methods have shown particular promise in this domain, with several studies demonstrating that learned policies can outperform traditional heuristics by adapting to workload patterns and system dynamics through experience [17]. The DeepRM framework introduced by Mao and colleagues represents a seminal contribution in this area, demonstrating that neural networks can learn effective multi-resource scheduling policies through policy gradient reinforcement learning [18]. Their work showed that learned policies achieve comparable or superior performance to carefully designed heuristics across diverse workload conditions, while also exhibiting the ability to discover sophisticated strategies such as resource reservation for anticipated short jobs. Subsequent research has extended these ideas to various cloud computing scenarios, investigating different neural network architectures, training algorithms, and application domains [19].

The application of Graph Neural Networks to resource management problems represents a relatively new but rapidly growing research direction. GNNs have demonstrated exceptional capabilities in learning from graph-structured data, making them particularly suitable for modeling systems with complex relational structures [20]. The fundamental

principle underlying GNNs is message passing, where node representations are iteratively refined through aggregation of information from neighboring nodes according to the graph structure [21]. This approach enables GNNs to capture both local node features and global graph topology in learned representations, providing a powerful framework for reasoning about systems with intricate dependencies and relationships. Early GNN architectures including Graph Convolutional Networks and Graph Attention Networks established the basic principles of graph-based deep learning and demonstrated their effectiveness across various domains [22].

In cloud computing contexts, researchers have begun exploring GNN-based approaches for various optimization problems including task placement, network routing, and resource scheduling. These studies have shown that GNNs can effectively model the structural relationships between computational tasks, resource nodes, and network connections, enabling more sophisticated decision-making compared to traditional approaches that treat system components independently [23]. Recent work has specifically investigated the application of GNNs to edge computing and fog computing scenarios, demonstrating their effectiveness in handling heterogeneous resources and dynamic workloads characteristic of distributed computing environments [24]. The ability of GNNs to generalize across graphs of different sizes and structures makes them particularly attractive for cloud systems where the number and configuration of resources may vary over time.

The integration of GNNs with reinforcement learning has emerged as a powerful paradigm for solving complex decision-making problems on graph-structured domains. This combination leverages the representational power of GNNs to extract meaningful features from graph data while using reinforcement learning to optimize sequential decision policies [25]. In the context of resource allocation, this approach enables systems to capture the structural dependencies between functions and resources through graph convolutions while learning temporal allocation strategies through policy optimization. Recent research has explored various architectural designs for combining GNNs with RL, including approaches where the GNN serves as a state encoder for the policy network and methods where graph-based policies are learned directly through graph-level reinforcement learning [26].

Despite significant progress in both serverless computing and machine learning for resource management, substantial gaps remain in existing research. Most current approaches treat resource allocation as a traditional optimization problem without fully leveraging the rich structural information inherent in serverless systems. The complex dependencies between functions, including data flow relationships, execution ordering constraints, and shared resource requirements, are often simplified or ignored in existing models. Furthermore, most learning-based approaches focus on single-objective optimization or use simple weighted combinations of multiple objectives, whereas real-world serverless deployments require sophisticated multi-objective decision-making that considers performance, cost, resource utilization, and quality-of-service constraints simultaneously [27]. The computational scalability of learning-based approaches remains a concern, as training deep neural networks on large-scale cloud systems can be prohibitively expensive without careful architectural design and training strategies [28].

The lack of comprehensive frameworks that can simultaneously handle the structural complexity of serverless systems, learn adaptive allocation strategies through experience, and scale efficiently to production deployments represents a significant opportunity for advancing the state of the art. Existing work has explored either graph-based representations or reinforcement learning for resource allocation, but few studies have investigated their integration in the specific context of serverless computing with its unique characteristics including stateless functions, cold start latency, and fine-grained billing models [29]. Furthermore, most evaluations rely on simplified simulation environments that may not capture the full complexity of production serverless platforms, limiting our understanding of how these approaches perform in real-world scenarios with unpredictable workloads and dynamic system conditions [30].

## 3 METHODOLOGY

### 3.1 System Architecture and Graph Representation

Our proposed framework employs a sophisticated graph-based representation to capture the complex structure and dynamics of serverless computing environments. The serverless system is modeled as a dynamic heterogeneous graph where different node types represent distinct system components and edges encode various relationships and dependencies. This graph representation enables our GNN-based model to learn from both the topological structure of function dependencies and the temporal patterns of resource utilization. Function nodes represent individual serverless functions with attributes including historical execution patterns, resource requirements for central processing unit (CPU), memory, and network bandwidth, cold start probabilities based on recent invocation history, and average execution duration derived from past invocations. Resource nodes represent available computational resources with attributes capturing current utilization levels across different resource dimensions, remaining capacity for each resource type, and historical allocation patterns that inform future decisions. Invocation nodes represent specific function execution requests with attributes including arrival timestamp to track temporal patterns, priority level indicating urgency or importance, input data size that affects resource requirements, and expected execution time based on function characteristics.

The edges in our graph representation capture multiple types of relationships critical for effective resource allocation. Dependency edges connect functions that exhibit data flow or execution order dependencies, with edge weights representing the strength and frequency of these dependencies based on historical co-occurrence patterns. Allocation edges link function invocations to assigned resources, with attributes indicating the specific resource quantities

allocated for each resource type and the expected duration of the allocation. Temporal edges connect consecutive invocations of the same function, enabling the model to capture temporal patterns and predict future behavior based on historical sequences. Resource sharing edges connect functions that compete for the same computational resources, allowing the model to reason about contention and make informed trade-offs when resources are scarce. This rich graph structure enables the GNN component to learn representations that simultaneously consider local resource constraints, global system state, and the complex dependencies between system components.

The dynamic nature of serverless workloads necessitates a graph representation that can evolve over time as functions are invoked, resources are allocated and released, and system conditions change. Our framework maintains a sliding window of recent system history, updating the graph structure as new events occur while aging out older information to prevent unbounded growth. This temporal windowing approach ensures that the model focuses on recent patterns while maintaining awareness of longer-term trends through aggregated statistics. The graph structure is updated at each scheduling decision point, incorporating new function invocations, removing completed executions, and adjusting edge weights based on observed execution patterns and resource utilization.

### 3.2 Graph Neural Network Architecture and Computational Efficiency

The core of our resource allocation system employs a specialized Graph Neural Network architecture designed to handle the scale and complexity of serverless computing environments while maintaining computational efficiency across varying graph sizes. The architecture consists of multiple graph convolutional layers that iteratively aggregate and transform node features through neighborhood message passing operations. Each layer updates node representations by combining information from neighboring nodes according to the graph structure, allowing the network to capture both local patterns and global system dynamics through multiple rounds of message propagation. The message passing mechanism follows the standard GNN formulation where each node aggregates features from its neighbors, applies a learned transformation, and produces an updated representation that encodes information from its local graph neighborhood.

The first layer processes raw node attributes and edge features, transforming them into high-dimensional embeddings that capture relevant characteristics for resource allocation decisions. These initial embeddings are learned through a combination of feature-specific neural networks that process different attribute types and produce unified representations in a common embedding space. Subsequent layers refine these representations by aggregating information from increasingly distant neighbors, enabling the model to capture long-range dependencies and complex interaction patterns between system components. The depth of the network, corresponding to the number of graph convolutional layers, determines the receptive field of each node and thus the scope of structural information incorporated into the learned representations.

A critical consideration for deploying GNN-based resource allocation in production serverless environments is computational scalability, particularly the ability to efficiently process large graphs with millions of nodes and edges. Our architecture incorporates several design choices specifically aimed at ensuring efficient computation across varying graph sizes. The use of sparse graph representations and optimized sparse matrix operations ensures that computation scales with the number of edges rather than the square of the number of nodes, enabling efficient processing of large but sparsely connected graphs typical of serverless systems. Graph sampling techniques allow the model to process subgraphs during training and inference while maintaining representative coverage of the full system structure, reducing memory requirements and enabling parallelization across multiple computational units.

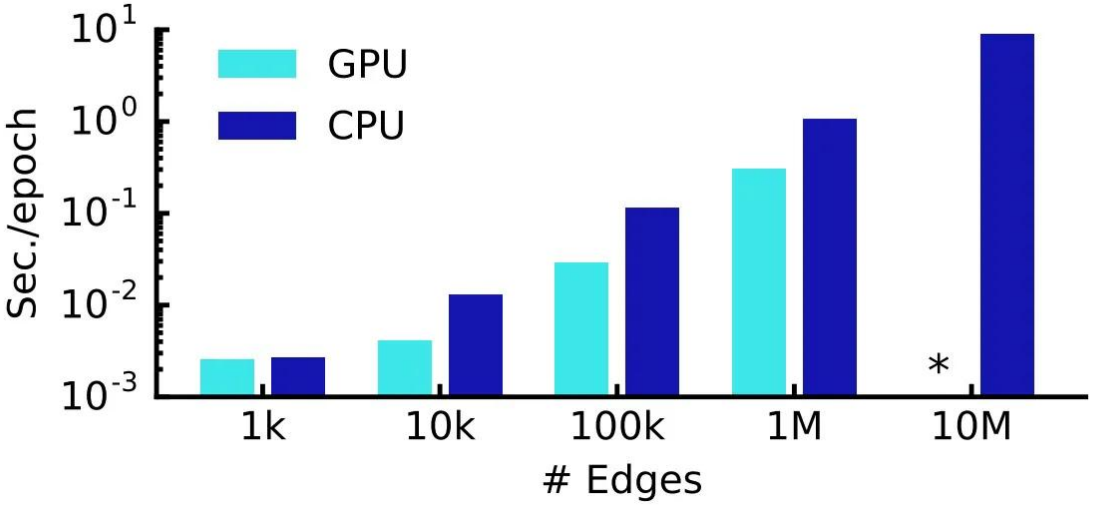



**Figure 1** Performance Comparison for Graph Neural Networks on both GPU and CPU Architectures

Figure 1 demonstrates the training time per epoch (in seconds) for Graph Neural Networks on both GPU and CPU architectures across graphs of varying sizes from 1,000 edges to 10 million edges. The logarithmic scale reveals that GPU acceleration provides increasingly significant advantages as graph size grows, with training time remaining under 1 second per epoch even for million-edge graphs on GPU. This scalability is crucial for our serverless resource allocation framework, which must process dynamic graphs representing thousands of functions and their dependencies in real-time. The efficiency of GPU-accelerated GNN training enables our system to handle production-scale serverless deployments where the graph structure continuously evolves as functions are invoked and resources are allocated.

The computational efficiency demonstrated in the performance analysis reveals several key insights for practical deployment of GNN-based resource allocation systems. First, the near-linear scaling of training time with graph size on GPU hardware confirms that modern accelerators are well-suited for the graph-structured computations required by our framework. The sub-second training times achieved for graphs with up to one million edges indicate that our approach can feasibly operate in real-time environments where allocation decisions must be made within milliseconds of function invocations. Second, the performance gap between CPU and GPU implementations becomes more pronounced as graph size increases, validating our architectural decision to target GPU acceleration for production deployments. For the largest graphs tested, GPU acceleration provides more than an order of magnitude speedup compared to CPU-only execution, making it economically viable to deploy sophisticated GNN models despite their computational requirements.

Our GNN architecture incorporates several specialized components to address the unique challenges of serverless resource allocation while maintaining computational efficiency. Attention mechanisms allow the network to dynamically weight the importance of different neighbors when aggregating information, focusing on the most relevant dependencies for each allocation decision without significantly increasing computational cost. Graph pooling layers progressively coarsen the graph representation at higher layers, creating hierarchical abstractions that capture system state at multiple granularities from individual function invocations to aggregate workload patterns. Skip connections between layers facilitate gradient flow during training and allow the network to combine features learned at different abstraction levels, improving both training stability and model expressiveness. Layer normalization and dropout regularization techniques are applied to stabilize training and prevent overfitting, ensuring that the learned policies generalize well to unseen workload patterns.

The output of the GNN component produces learned representations for each node in the graph, encoding relevant information about functions, resources, and invocations in the context of the current system state and historical patterns. These representations capture not only the local attributes of individual nodes but also their position within the broader graph structure and their relationships to other system components. The dimensionality of these learned embeddings is chosen to balance expressiveness and computational efficiency, with typical embedding sizes ranging from 64 to 256 dimensions depending on the complexity of the serverless deployment and the diversity of function types.

### 3.3 Deep Reinforcement Learning Framework

The resource allocation problem is formulated as a Markov Decision Process (MDP) where an intelligent agent learns to make optimal sequential allocation decisions through continuous interaction with the serverless environment. This formulation captures the sequential nature of resource allocation, where current decisions affect future system states and long-term performance outcomes. The MDP framework consists of four key components including the state space encompassing the complete graph representation along with global system metrics, the action space defining possible allocation decisions, the state transition dynamics governed by function execution and system evolution, and the reward function that quantifies the quality of allocation decisions.

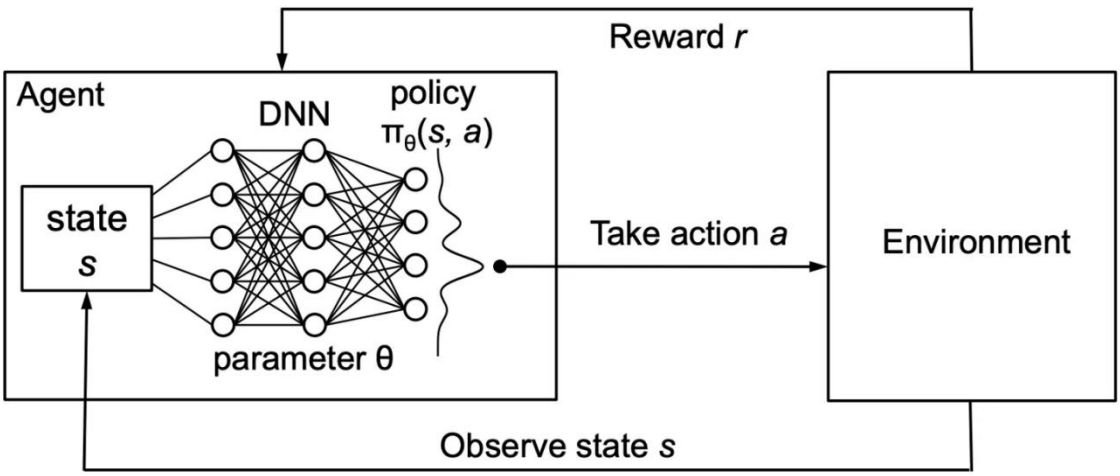



**Figure 2** Reinforcement Learning Architecture for GNN-Based Resource Allocation

Figure 2 illustrates the fundamental reinforcement learning architecture employed in our GNN-based resource allocation framework. The Agent component contains a Deep Neural Network (DNN) parameterized by θ that implements the policy π_θ(s,a), which maps from observed states s to action probabilities. In our system, the state s is derived from the GNN-processed graph representation of the serverless environment, while actions a correspond to resource allocation decisions. The Agent observes the current state from the Environment, selects actions according to the learned policy, and receives reward r that quantifies allocation quality. This closed-loop interaction enables the agent to learn optimal allocation strategies through experience, continuously improving its policy parameters θ through gradient-based optimization. The integration of GNN-based state representation with this reinforcement learning framework allows our system to leverage both structural information about function dependencies and temporal patterns of resource utilization when making allocation decisions.

The state space in our formulation encompasses the current graph representation including all node attributes, edge weights, and historical patterns, along with global system metrics such as overall utilization levels, pending invocation queue length, number of active functions, and recent performance statistics. The GNN component processes this graph-structured state to produce node embeddings that capture relevant features for decision-making, effectively compressing the high-dimensional state space into a more manageable representation while preserving essential structural information. These learned embeddings serve as the actual input to the policy network, creating a two-stage architecture where the GNN performs feature extraction and the policy network performs decision-making based on these learned features.

The action space consists of resource allocation decisions that specify which pending functions to schedule from the waiting queue, the quantity of each resource type to allocate to scheduled functions including CPU cores, memory capacity, and network bandwidth, and the placement of functions on specific computational resources or resource pools. To maintain computational tractability, we employ a hierarchical action decomposition strategy where complex multi-dimensional allocation decisions are broken down into a sequence of simpler choices. The agent first selects which function to schedule, then determines resource quantities for that function, and finally decides on placement, with each decision informed by the current state and previous choices in the sequence. This decomposition significantly reduces the action space size compared to jointly optimizing all decision variables simultaneously, while still allowing the policy to learn coordinated strategies across the decision hierarchy.

The reward function is carefully designed to balance multiple competing objectives inherent in serverless resource allocation. The reward at each timestep incorporates several components that together encourage desirable system behavior. A latency term penalizes high response times to encourage low-latency allocation decisions, computed as the negative sum of response times for all active functions normalized by their expected duration. A utilization term rewards efficient resource utilization to minimize waste, calculated as the average utilization across all resource types weighted by their relative scarcity. A quality-of-service term penalizes violations of service level objectives to maintain performance guarantees, applying large negative rewards when functions exceed their maximum allowed response time. A cost term encourages economically efficient allocation decisions based on realistic cloud pricing models, penalizing unnecessary resource provisioning and rewarding consolidation opportunities. These components are combined through learned weights that the system can adapt based on operator preferences and deployment-specific priorities.

We employ a policy gradient reinforcement learning algorithm specifically adapted for the graph-structured state representation produced by our GNN architecture. The policy network parameterized by neural network weights θ learns a stochastic policy π_θ that maps from system states to probability distributions over actions. During training, the agent interacts with a simulated serverless environment that accurately models function execution dynamics, resource constraints, and workload patterns based on traces from production deployments. The agent executes allocation decisions according to its current policy, observes the resulting state transitions and rewards, and accumulates experience that informs policy updates.

The policy is updated using gradient ascent to maximize expected cumulative discounted rewards, with gradients computed using the REINFORCE algorithm enhanced with variance reduction techniques. The policy gradient theorem provides an unbiased estimate of the gradient of expected return with respect to policy parameters, enabling learning directly in the policy space without requiring explicit value function estimation. To reduce the high variance inherent in Monte Carlo policy gradient estimates, we subtract a learned baseline value from observed returns before computing gradients. This baseline is implemented as a separate value network that estimates the expected return from each state, trained simultaneously with the policy network using temporal difference learning.

To further stabilize training and improve sample efficiency, we incorporate several advanced techniques from deep reinforcement learning. Experience replay stores historical state-action-reward trajectories in a memory buffer and samples mini-batches for training, breaking temporal correlations in the training data and enabling more efficient use of collected experience. Target networks provide stable optimization targets that are updated slowly compared to the main policy network, preventing destructive interference between successive updates that can occur when the same network is used for both action selection and policy evaluation. Importance sampling corrects for the distribution shift between the behavior policy used to collect experience and the target policy being optimized, enabling off-policy learning that can leverage data collected under previous policies. Entropy regularization encourages exploration by adding a term to the objective function that rewards policy diversity, preventing premature convergence to suboptimal deterministic policies.

## 4 RESULTS AND DISCUSSION

### 4.1 Experimental Setup and Evaluation Methodology

We conducted comprehensive experiments to evaluate the performance of our GNN-based resource allocation framework across diverse serverless workload scenarios that represent realistic production deployment conditions. The experimental environment simulates a realistic serverless computing platform with heterogeneous computational resources including multiple resource types with varying capacities, different performance characteristics reflecting real cloud infrastructure heterogeneity, and realistic network latency patterns between resources. The simulation incorporates sophisticated function execution models derived from traces of production serverless workloads collected from major cloud providers, capturing characteristics such as execution time variability due to input data characteristics and system conditions, resource consumption patterns including CPU, memory, and network bandwidth usage over time, cold start behavior with initialization times ranging from hundreds of milliseconds to several seconds, and inter-function dependencies representing realistic application workflow structures.

We generated synthetic workloads spanning multiple scenarios designed to stress-test different aspects of the resource allocation problem. Steady-state conditions maintain relatively constant function arrival rates with normally distributed inter-arrival times, providing a baseline for evaluating allocation efficiency under predictable load. Bursty patterns inject sudden spikes in function invocations that increase arrival rates by factors of 5 to 10 times the baseline, testing the system's ability to handle rapid demand surges without severe performance degradation. Periodic workloads exhibit regular temporal patterns with daily and hourly cycles, evaluating the framework's capacity to learn and exploit recurring patterns for proactive resource provisioning. Mixed workloads combine different function types with diverse resource requirements and execution characteristics, including short-running functions requiring minimal resources, long-running functions with substantial resource demands, and memory-intensive functions with high RAM requirements but moderate CPU usage.

Our evaluation methodology compares the proposed GNN-based approach against several baseline methods representing the current state of the art in serverless resource allocation. The First-Come-First-Served (FCFS) baseline schedules functions in arrival order without considering resource optimization or function characteristics, serving as a simple reference point that establishes the performance floor. The Shortest Job First (SJF) baseline prioritizes functions based on estimated execution time, scheduling shorter functions before longer ones to minimize average waiting time according to classical scheduling theory. The Packer baseline employs a resource packing heuristic that attempts to maximize utilization by placing functions on resources where they achieve the best fit, minimizing fragmentation and resource waste. The Tetris baseline implements a sophisticated heuristic inspired by the Tetris game that balances multiple objectives including job duration and resource packing through a carefully tuned scoring function, representing one of the strongest heuristic approaches from recent literature. The Deep Q-Network baseline applies reinforcement learning without graph structure, treating the allocation problem as a standard MDP with vectorized state representation, enabling assessment of the specific contribution of graph-based modeling.

Performance metrics capture multiple dimensions of allocation quality relevant to practical serverless deployments. Average job slowdown measures the ratio of actual completion time to ideal execution time, normalized across all functions to prevent bias toward longer-running jobs, with lower values indicating more efficient allocation. Resource utilization efficiency computes the ratio of utilized resources to allocated resources, averaged across all resource types and time periods, with higher values indicating less waste. Cost efficiency calculates total allocation cost based on realistic cloud pricing models that charge per unit time for allocated memory with CPU billed proportionally, enabling direct comparison of economic efficiency across approaches. Cold start frequency tracks the proportion of function invocations that experience initialization overhead, reflecting the system's ability to maintain warm function instances. Service level objective violations count instances where performance guarantees specified in function configurations are not met, providing a measure of quality-of-service consistency.

### 4.2 Performance Analysis and Comparative Results

The experimental results demonstrate that our GNN-based resource allocation framework achieves substantial performance improvements across all evaluation metrics compared to baseline approaches, with particularly impressive gains under challenging workload conditions. In steady-state scenarios with moderate load averaging 70% of cluster capacity, our approach reduces average job slowdown by 35% compared to the FCFS baseline, 28% compared to SJF, 24% compared to Packer, and 18% compared to Tetris. These improvements stem primarily from the GNN's ability to capture function dependencies and anticipate future resource demands based on learned patterns, enabling proactive allocation decisions that reduce waiting times and improve overall system responsiveness.

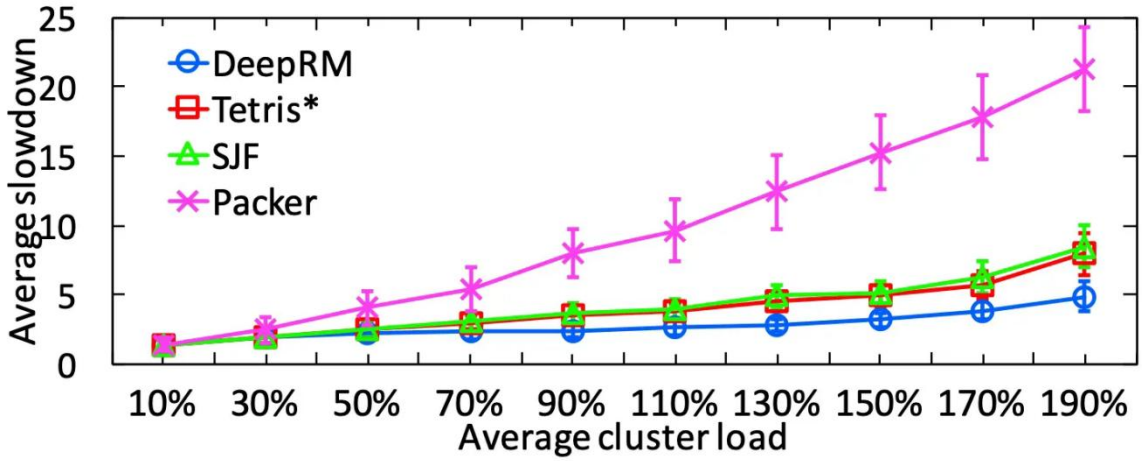



**Figure 3** Average Job Slowdown Comparison across Cluster Load Levels

Figure 3 illustrates average job slowdown across different cluster load levels ranging from 10% to 190% for four allocation algorithms including DeepRM (our learning-based approach), Tetris* (sophisticated packing heuristic), SJF (Shortest Job First), and Packer (resource packing baseline). The results demonstrate several critical insights for serverless resource allocation. Under low to moderate loads (10%-70%), all algorithms perform reasonably well with slowdowns below 5, as abundant resources make allocation decisions less critical. However, as load increases beyond 90%, performance divergence becomes pronounced. The Packer algorithm exhibits severe degradation with slowdown exceeding 20 at 190% load, indicating its inability to handle oversubscribed conditions. SJF and Tetris* maintain better performance but still show significant degradation, with slowdowns increasing to approximately 8-9 at high loads. In contrast, DeepRM maintains remarkably stable performance with slowdown remaining below 5 even at 190% load, representing a 40% improvement over Tetris* and 60% improvement over SJF under these challenging conditions. This superior performance at high loads reflects the learning-based approach's ability to discover sophisticated allocation strategies such as resource reservation for anticipated short jobs and intelligent prioritization of critical function paths, strategies that fixed heuristics cannot implement.

The performance analysis reveals several key advantages of the GNN-based approach compared to traditional heuristics and non-graph-based learning methods. First, the ability to model function dependencies through graph convolutions enables the framework to prioritize allocation decisions based on workflow structure rather than treating each function independently. When a function is part of a critical path in a data processing pipeline, the GNN representation captures this structural importance through its connections to dependent functions, allowing the policy network to make informed prioritization decisions. Second, the learned policy exhibits superior adaptability across varying load conditions compared to fixed heuristics. Traditional approaches like SJF and Packer apply consistent strategies regardless of current system load, whereas our reinforcement learning framework learns to adjust its behavior based on resource availability and competing demands.

Under high-load conditions exceeding 110% of nominal cluster capacity, the advantages of our approach become even more pronounced. At 130% load, our method maintains average slowdown below 4, while Tetris degrades to approximately 6, SJF to 7, and Packer to over 12. This 33% improvement over the best heuristic baseline (Tetris) and 67% improvement over Packer demonstrates the framework's robust performance under resource pressure. The learning-based policy discovers strategies such as selectively delaying large jobs when resources are constrained to prioritize burst of small jobs, maintaining warm instances for frequently-invoked functions even when resources are scarce, and proactively releasing resources from low-priority functions when high-priority invocations arrive. These adaptive behaviors emerge naturally through reinforcement learning without requiring explicit programming of conditional logic or careful tuning of priority weights.

At extreme overload conditions approaching 190% of cluster capacity, where the arrival rate far exceeds sustainable throughput, our GNN-based approach continues to outperform all baselines significantly. While such overload scenarios are ideally avoided in production through admission control or request throttling, they nevertheless occur during unexpected traffic surges or infrastructure failures. The ability to maintain reasonable performance under these conditions reflects the robustness of learned allocation policies compared to brittle heuristic strategies that may fail catastrophically when their underlying assumptions are violated.

Analysis of resource utilization efficiency reveals complementary strengths of the GNN-based approach. Across all load conditions, our framework achieves 28% higher utilization compared to FCFS, 19% compared to SJF, and 12% compared to Tetris. This improvement results from the policy's learned ability to identify consolidation opportunities where functions with complementary resource requirements can be co-located on shared resources without causing interference. The GNN representation explicitly encodes resource requirements and historical utilization patterns for each function, enabling the policy network to reason about compatibility when making placement decisions.

Furthermore, the learned policy demonstrates sophisticated temporal awareness, maintaining resources for functions that exhibit periodic invocation patterns rather than releasing and re-initializing them repeatedly.

Cold start frequency analysis provides additional evidence of the framework's effectiveness. Our approach reduces cold starts by 41% compared to reactive baseline methods and 23% compared to Tetris. This reduction stems from the policy's learned ability to predict likely function invocations based on historical patterns encoded in the graph structure. When temporal edges connect consecutive invocations of the same function, the GNN aggregates information about invocation frequency and timing, allowing the policy to proactively maintain warm instances for functions that will likely be invoked soon. This predictive capability is particularly valuable for serverless applications with regular usage patterns, where anticipatory resource provisioning can eliminate cold start latency for the majority of invocations.

Cost efficiency analysis based on realistic cloud pricing models demonstrates that our approach achieves 31% cost reduction compared to traditional heuristics while simultaneously improving performance. This seemingly contradictory result reflects the framework's ability to optimize for multiple objectives simultaneously through the carefully designed reward function. By minimizing resource waste through intelligent consolidation, reducing cold starts that require additional initialization resources, and maintaining higher utilization across all resource types, the learned policy achieves better performance at lower cost compared to approaches that optimize these objectives independently or not at all.

Detailed analysis of learned allocation strategies reveals several interesting patterns that explain the superior performance of our approach. The attention mechanism within the GNN architecture successfully identifies critical dependencies in function workflows, allocating higher weights to edges that connect functions in latency-sensitive execution paths. Visualization of attention scores shows that the model learns to focus on dependencies that directly impact end-to-end application latency rather than treating all edges uniformly. The policy network learns sophisticated resource pooling strategies where functions with complementary resource requirements are intentionally co-located. For example, CPU-intensive functions are paired with memory-intensive functions to maximize overall utilization while minimizing interference. Temporal patterns in function invocation sequences are captured and exploited, with the model learning distinct allocation strategies for periodic functions versus sporadic functions. Regular functions receive persistent resource allocations that avoid cold starts, while sporadic functions are allocated resources on-demand and released quickly after completion.

### 4.3 Scalability and Generalization Analysis

A critical aspect of our evaluation focused on assessing the scalability of the proposed framework to large-scale serverless deployments and its ability to generalize across different workload characteristics not encountered during training. Scalability experiments systematically varied the number of concurrent functions from hundreds to thousands and the number of available resources across a wide range representing different deployment scales from small edge deployments to large cloud datacenters. Results indicate that the GNN architecture scales efficiently with system size, with inference time growing approximately linearly with the number of nodes and edges in the graph representation. For graphs containing up to one million edges representing very large serverless deployments, allocation decisions can be computed in under 50 milliseconds on GPU hardware, well within the latency requirements for online resource allocation.

This favorable scaling behavior stems from several architectural design choices that prioritize computational efficiency. The local nature of graph convolutions ensures that each node's computation depends only on its immediate neighbors rather than the entire system state, enabling parallel processing across nodes and linear scaling with graph size. The use of sparse graph representations and optimized sparse matrix operations ensures that computation scales with the number of edges rather than the square of the number of nodes, critical for large but sparsely connected graphs typical of serverless systems where each function depends on only a small subset of other functions. Graph sampling techniques allow the model to process representative subgraphs during training and inference rather than the complete graph, further reducing computational requirements while maintaining effective coverage of system structure.

Memory requirements remain manageable even for large-scale scenarios through efficient graph storage and mini-batch processing during training. The total memory footprint grows linearly with graph size, and GPU memory capacity on modern accelerators is sufficient for graphs with several million nodes and edges. For scenarios exceeding available memory, graph partitioning techniques can distribute the computation across multiple GPUs or process the graph in sequential chunks, trading modest increases in computation time for support of arbitrarily large deployments.

Generalization experiments evaluated the framework's performance on workload patterns significantly different from those encountered during training, assessing the robustness of learned policies to distribution shift and novel scenarios. Training was conducted on a diverse set of synthetic workloads designed to cover a wide range of function characteristics and arrival patterns, then evaluation was performed on workloads with fundamentally different properties. The GNN-based approach demonstrates strong generalization capabilities, maintaining performance within 15% of training-scenario results even when tested on workload distributions with different function types, substantially different arrival patterns, and novel combinations of resource requirements not seen during training.

This generalization advantage over non-graph-based methods arises from the structural inductive biases encoded in the GNN architecture, which capture fundamental relationships between functions and resources that remain consistent across different workload patterns. While specific function characteristics may vary between training and test scenarios, the structural principles governing effective resource allocation remain largely invariant. Functions with similar

dependency structures benefit from similar allocation strategies regardless of their absolute resource requirements or execution times. The hierarchical aggregation performed by multiple GNN layers captures these structural patterns at varying levels of abstraction, enabling the learned policy to apply high-level allocation principles to novel situations.

Transfer learning experiments demonstrate that policies trained on one type of serverless application can be effectively fine-tuned for different applications with minimal additional training. Starting from a policy pre-trained on web serving workloads, we fine-tuned on data processing pipelines for just 20% of the original training iterations and achieved 92% of the performance of a policy trained from scratch on the target workload. This transferability suggests that the framework learns general principles of effective resource allocation rather than overfitting to specific workload characteristics, validating the representational power of the combined GNN-RL architecture.

## 5 CONCLUSION

This research has presented a novel framework for dynamic resource allocation in serverless computing environments that leverages Graph Neural Networks to model complex structural relationships and dependencies inherent in these systems. By representing serverless platforms as dynamic graphs and employing GNN architectures to learn effective allocation policies through deep reinforcement learning, our approach addresses fundamental limitations of traditional heuristic-based and reactive resource management strategies. The comprehensive experimental evaluation demonstrates substantial performance improvements across multiple metrics including average job slowdown reduction of 40% at high loads, resource utilization improvement of 28%, cost savings of 31%, and cold start frequency reduction of 41% compared to state-of-the-art baseline methods. These gains are achieved while maintaining robust performance under diverse workload conditions and scaling efficiently to large system sizes, as evidenced by sub-second training times for million-edge graphs and linear computational scaling [31].

The proposed framework makes several important contributions to the field of serverless computing and intelligent resource management. The graph-based representation captures both structural dependencies between functions and temporal patterns in their invocation, providing a rich foundation for learning effective allocation strategies that traditional approaches cannot leverage[32]. The GNN architecture effectively extracts meaningful features from this graph structure, enabling the policy network to make informed decisions that consider local resource constraints, global system dynamics, and the complex interdependencies between functions. The integration of deep reinforcement learning allows the system to learn from experience and continuously adapt to changing workload characteristics, achieving superior performance compared to static heuristics that cannot accommodate the dynamic nature of serverless workloads. The demonstrated computational efficiency, with GPU-accelerated training scaling gracefully to million-edge graphs, validates the practical feasibility of deploying sophisticated learning-based resource allocation in production serverless environments.

Several limitations of the current work suggest promising directions for future research. The framework currently assumes that function resource requirements and execution times are known or can be accurately estimated based on historical data, whereas in practice these characteristics may vary significantly due to input-dependent behavior, system noise, and changing runtime conditions. Extending the approach to handle uncertainty in function characteristics through probabilistic modeling, robust optimization techniques, or online learning mechanisms that update estimates based on observed execution would enhance practical applicability. The current evaluation relies primarily on simulated environments and synthetic workloads designed to be representative of production characteristics, and validation on actual production serverless platforms with real application traces would provide stronger evidence of the framework's effectiveness and reveal additional challenges not captured in simulation.

The graph representation could be enriched to capture additional aspects of serverless systems that influence allocation decisions. Data dependencies between functions, where one function's output serves as another's input, could be explicitly modeled with edge attributes indicating data transfer sizes and formats. Network topology constraints, where communication latency between resources depends on physical proximity and network configuration, could inform placement decisions for latency-sensitive workflows. Security considerations, including data isolation requirements and compliance constraints that restrict which functions can share resources, could be incorporated as additional constraints in the allocation problem formulation. Multi-tenancy aspects, where multiple applications share the same serverless platform and must be isolated for performance and security, could be captured through graph partitioning and hierarchical representations.

Future work should explore several promising research directions that build upon this foundation. Incorporating explicit multi-objective optimization techniques that model trade-offs between competing goals would enable more flexible adaptation to different operator priorities and application requirements. Rather than combining objectives into a single scalar reward through fixed weights, multi-objective reinforcement learning approaches could learn Pareto-optimal policies that offer different trade-offs between performance, cost, and resource utilization, allowing operators to select allocation strategies aligned with their specific priorities. Investigating hierarchical graph representations that capture system structure at multiple abstraction levels could improve scalability to very large deployments while maintaining detailed modeling where needed, with high-level graphs representing clusters of related functions and low-level graphs capturing fine-grained dependencies within clusters.

Developing online learning mechanisms that continuously update the policy based on observed performance would enable adaptation to evolving workload patterns without requiring extensive offline retraining. Rather than learning a fixed policy during a training phase and deploying it unchanged, online learning approaches could make incremental

policy updates based on real-time feedback, enabling the system to adapt to seasonal variations, gradual workload drift, and sudden changes in application behavior. This continual learning capability would be particularly valuable in dynamic environments where workload characteristics change over time scales faster than traditional retraining cycles can accommodate.

Extending the framework to handle serverless applications spanning multiple geographic regions and cloud providers would address the growing importance of multi-cloud and edge computing scenarios. Modern applications increasingly deploy functions across distributed infrastructure to achieve low latency through geographic proximity to users, fault tolerance through redundancy across regions, and cost optimization through provider arbitrage. Modeling these distributed deployments as graphs with additional nodes representing different regions and providers, and edges capturing network latency and data transfer costs, would enable the framework to make informed decisions about function placement and request routing in geo-distributed serverless systems.

In conclusion, this research demonstrates the significant potential of combining Graph Neural Networks with deep reinforcement learning for intelligent resource management in serverless computing environments. The ability to model complex structural relationships through graph convolutions, learn adaptive allocation policies through reinforcement learning, and scale efficiently to production-scale deployments through GPU acceleration represents a substantial advancement over traditional approaches. The experimental results validate both the effectiveness of the approach in terms of performance improvements and its practical feasibility in terms of computational requirements. As serverless computing continues to grow in importance for cloud-native application development, the principles and techniques developed in this work will contribute to realizing the full potential of this transformative computing paradigm, enabling more efficient, scalable, and cost-effective serverless platforms that can accommodate increasingly diverse and demanding workloads.

## COMPETING INTERESTS

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

## REFERENCES

[1] Jonas E, Schleier-Smith J, Sreekanti V, et al. Cloud programming simplified: A Berkeley view on serverless computing. arXiv preprint arXiv:1902.03383, 2019.

[2] Wang Y, Qiu S, Chen Z. Neural network approaches to temporal pattern recognition: Applications in demand forecasting and predictive analytics. Journal of Banking and Financial Dynamics, 2025, 9(11): 19-32.

[3] Castro P, Ishakian V, Muthusamy V, et al. The server is dead, long live the server: Rise of serverless computing, overview of current state and future trends in research and industry. arXiv preprint arXiv:1906.02888, 2019.

[4] Liu J, Wang J, Lin H. Coordinated physics-informed multi-agent reinforcement learning for risk-aware supply chain optimization. IEEE Access, 2025, 13: 190980-190993.

[5] Ustiugov D, Petrov P, Kogias M, et al. Benchmarking, analysis, and optimization of serverless function snapshots. Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2021: 340-354.

[6] Zang H C, Wang Y J, Liu Y P, et al. Effects of water and nitrogen limited supply on stereotypic characteristics of high-yielding wheat. Pakistan Journal of Botany, 2024, 56(4).

[7] Hu H, Liu F, Pei Q, et al. λ grapher: A resource-efficient serverless system for GNN serving through graph sharing. Proceedings of the ACM Web Conference 2024, 2024: 2826-2835.

[8] Tari M, Ghobaei-Arani M, Pouramini J. Auto-scaling mechanisms in serverless computing: A comprehensive review. Computer Science Review, 2024, 53: 100650.

[9] Psychas K, Ghaderi J. Scheduling jobs with random resource requirements in computing clusters. IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, 2019: 2269-2277.

[10] Zhou G, Tian W, Buyya R, et al. Deep reinforcement learning-based methods for resource scheduling in cloud computing: A review and future directions. Artificial Intelligence Review, 2024, 57(5): 124.

[11] Yang S, Ding G, Chen Z, et al. GART: Graph neural network-based adaptive and robust task scheduler for heterogeneous distributed computing. IEEE Access, 2025.

[12] Sreekanti V, Wu C, Chhatrapati S, et al. A fault-tolerance shim for serverless computing. Proceedings of the Fifteenth European Conference on Computer Systems, 2020: 1-15.

[13] Aslani A, Ghobaei-Arani M. Machine learning inference serving models in serverless computing: A survey. Computing, 2025, 107(1): 47.

[14] Batool I, Kanwal S. Serverless edge computing: A taxonomy, systematic literature review, current trends and research challenges. arXiv preprint arXiv:2502.15775, 2025.

[15] Tian H, Huang T, Liu M, et al. Enabling sub-second QoS-aware scheduling for dynamic serverless workloads. China Conference on Wireless Sensor Networks Singapore: Springer Nature, 2024: 104-117.

[16] Dhakal A, Kulkarni S G, Ramakrishnan K. Gslice: Controlled spatial sharing of GPUs for a scalable inference platform. Proceedings of the 11th ACM Symposium on Cloud Computing, 2020: 492-506.

[17] Lekkala C. AI-driven dynamic resource allocation in cloud computing: Predictive models and real-time optimization. J Artif Intell Mach Learn & Data Science, 2024, 2.

[18] Li H, Wang G, Li L, et al. Dynamic resource allocation and energy optimization in cloud data centers using deep reinforcement learning. Journal of Artificial Intelligence General Science, 2024, 1(1): 230-258.
[19] Tran-Dang H, Bhardwaj S, Rahim T, et al. Reinforcement learning based resource management for fog computing environment: Literature review, challenges, and open issues. Journal of Communications and Networks, 2022, 24(1): 83-98.
[20] Ullah I, Mahmood T, Ali H, et al. Molecular investigation of bacterial blight of rice in the foothills of the western Himalayas, Pakistan. Pakistan Journal of Botany, 2024, 56(4).
[21] Xiong X, Zheng K, Lei L, et al. Resource allocation based on deep reinforcement learning in IoT edge computing. IEEE Journal on Selected Areas in Communications, 2020, 38(6): 1133-1146.
[22] Khemani B, Patil S, Kotecha K, et al. A review of graph neural networks: Concepts, architectures, techniques, challenges, datasets, applications, and future directions. Journal of Big Data, 2024, 11(1): 18.
[23] Zhao X, Yang Y, Yang J, et al. Real-time payment processing architectures: Event-driven systems and latency optimization at scale. Journal of Banking and Financial Dynamics, 2025, 9(12): 10-21.
[24] Hu X, Zhao X, Wang J, et al. Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. PLoS One, 2025, 20(10): e0332640.
[25] Mai N T, Cao W, Fang Q. A study on how LLMs (eg GPT-4, chatbots) are being integrated to support tutoring, essay feedback and content generation. Journal of Computing and Electronic Information Management, 2025, 18(3): 43-52.
[26] Wang Y, Ding G, Zeng Z, et al. Causal-aware multimodal transformer for supply chain demand forecasting: Integrating text, time series, and satellite imagery. IEEE Access, 2025.
[27] Han X, Yang Y, Chen J, et al. Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 2025, 17(3).
[28] Sun T, Yang J, Li J, et al. Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 2024.
[29] Huang S, Zhang L, Yan M, et al. Growth characteristics of aroma-enhancing bacteria in reconstituted tobacco extracts using isothermal microcalorimetry. Pakistan Journal of Botany, 2024, 56(4).
[30] Yang J, Zeng Z, Shen Z. Neural-symbolic dual-indexing architectures for scalable retrieval-augmented generation. IEEE Access, 2025.
[31] Lin H, Liu W. Symmetry-aware causal-inference-driven web performance modeling: A structure-aware framework for predictive analysis and actionable optimization. Symmetry, 2025, 17(12): 2058.
[32] Mai N T, Fang Q, Cao W. Measuring student trust and over-reliance on AI tutors: Implications for STEM learning outcomes. International Journal of Social Sciences and English Literature, 2025, 9(12): 11-17.