AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3023

INTERPRETABLE TRANSFORMER MODELS FOR RELATIONSHIP ANALYSIS IN FINANCIAL DATA

Laura Chen, Robert Murphy*
University of Notre Dame, Notre Dame, USA.
Corresponding Author: Robert Murphy, Email: rmurphy91@nd.edu

Abstract: The increasing complexity of financial transaction networks has necessitated the development of sophisticated analytical tools capable of uncovering intricate relationships within heterogeneous financial data while maintaining interpretability for regulatory compliance and fraud detection purposes. This paper presents a novel framework for interpretable transformer models specifically designed for relationship analysis in financial transaction networks. Our approach builds upon the foundational attention mechanisms developed for sequence-to-sequence tasks and extends them through graph attention networks to handle complex multi-entity financial relationships. The framework demonstrates how attention-based architectures can effectively analyze heterogeneous networks comprising card numbers, transaction identifiers, email domains, and card types to identify suspicious patterns and fraudulent activities. We develop specialized visualization techniques that reveal temporal dependencies in transaction sequences and cross-entity correlations in financial networks. Experimental evaluation on real-world financial transaction datasets demonstrates that our interpretable transformer models achieve superior performance in fraud detection while providing actionable insights for financial analysts. The framework successfully identifies complex fraud patterns including coordinated attacks across multiple entity types, suspicious email-card associations, and abnormal transaction behaviors, with interpretability metrics showing high alignment with expert fraud analyst assessments.
Keywords: Interpretable machine learning; Transformer architecture; Financial relationship analysis; Attention mechanisms; Fraud detection; Heterogeneous networks

1 INTRODUCTION

The financial industry has witnessed an unprecedented transformation in transaction complexity and volume, driven by digital payment systems, e-commerce growth, and sophisticated fraud schemes that exploit multiple interconnected entities[1]. Modern financial transaction networks involve complex relationships between heterogeneous entities including credit card numbers, transaction records, email addresses, and payment instrument types, each contributing unique information to the overall transaction ecosystem[2].

Traditional rule-based fraud detection systems, while interpretable, often fail to capture the subtle and evolving patterns that characterize modern financial fraud schemes[3]. These systems typically analyze individual transactions or single entity types in isolation, missing the complex multi-entity relationships that sophisticated fraudsters exploit[4]. Conversely, advanced machine learning techniques excel at pattern recognition but suffer from interpretability limitations that restrict their adoption in regulated financial environments where decision transparency is paramount[5].

The development of attention mechanisms has revolutionized how machine learning models process sequential and structured data. Beginning with the encoder-decoder architectures that transformed neural machine translation, attention mechanisms have evolved to handle increasingly complex data structures including graphs and heterogeneous networks[6]. This evolution provides a powerful foundation for analyzing the intricate relationships present in financial transaction data.

The challenge of financial transaction analysis lies in understanding relationships that span multiple entity types and temporal scales. A single fraudulent scheme might involve coordinated use of multiple card numbers, specific email domain patterns, particular transaction timing sequences, and exploitation of certain card type vulnerabilities[7]. Detecting such schemes requires models that can simultaneously process sequential transaction data and complex entity relationships while providing interpretable explanations for their decisions[8].

This research addresses the critical need for interpretable models in financial transaction analysis by developing a comprehensive framework that traces the evolution from sequence-to-sequence attention mechanisms to heterogeneous network analysis. Our approach demonstrates how the interpretability advantages of attention-based architectures can be leveraged to understand complex financial relationships while maintaining the predictive performance necessary for practical fraud detection applications.

The primary contributions of this work include the adaptation of foundational attention mechanisms for financial sequence analysis, the extension of these mechanisms through graph attention networks to handle multi-entity relationships, and the development of specialized visualization techniques for interpreting complex financial transaction patterns. Our framework provides financial institutions with powerful tools for understanding fraud patterns while meeting regulatory requirements for model explainability.

2 LITERATURE REVIEW

The intersection of interpretable machine learning and financial transaction analysis has evolved significantly with the development of attention mechanisms and their application to increasingly complex data structures[9]. Early work in financial fraud detection relied primarily on rule-based systems and traditional statistical methods that, while interpretable, struggled to capture the sophisticated patterns characteristic of modern fraud schemes[10].

The foundational work of Bahdanau et al. Introduced attention mechanisms for neural machine translation, demonstrating how models could learn to focus on relevant parts of input sequences when generating outputs[11]. This breakthrough established the principle that attention weights could serve as interpretable indicators of model decision-making processes, providing insights into which input elements most influenced specific predictions[12]. The bidirectional encoder-decoder architecture with attention showed how models could effectively handle variable-length sequences while maintaining interpretability through attention weight visualization.

The success of attention mechanisms in natural language processing sparked interest in their application to other domains involving sequential and structured data[13]. Financial transaction analysis emerged as a natural application area, given the sequential nature of transaction data and the need for interpretable fraud detection systems. Early applications focused on adapting sequence-to-sequence models for transaction sequence analysis, treating fraud detection as a sequence classification problem[14].

The development of Graph Neural Networks marked another significant advancement in handling structured data[15]. Traditional neural networks struggled with data that exhibited complex relational structures, such as the multi-entity relationships present in financial transaction networks. The introduction of Graph Convolutional Networks and subsequently Graph Attention Networks by Veličković et al. Provided powerful tools for analyzing graph-structured data while maintaining some degree of interpretability through attention mechanisms[16].

Graph Attention Networks represented a particularly important advancement for financial applications because they could handle heterogeneous networks where different node types represented different entity categories[17]. This capability was crucial for financial transaction analysis, where understanding relationships between cards, merchants, users, and transactions required processing multiple entity types within a unified framework.

Recent research in financial machine learning has increasingly emphasized the importance of interpretability alongside predictive performance[18]. Regulatory requirements in the financial sector demand that automated decision-making systems, particularly those involved in fraud detection and credit assessment, provide clear explanations for their decisions[19-24]. This regulatory pressure has accelerated the development of interpretable machine learning techniques specifically designed for financial applications[25].

The application of attention mechanisms to financial fraud detection has revealed the importance of understanding both temporal patterns in transaction sequences and structural patterns in entity relationships[26]. Fraudulent activities often exhibit distinctive temporal signatures, such as unusual transaction timing or rapid sequences of high-value transactions, that can be captured through attention mechanisms applied to transaction sequences. Simultaneously, fraud schemes frequently exploit specific relationship patterns between different entity types, such as associations between particular email domains and card types [27].

Contemporary research has begun to explore the integration of sequence-based attention mechanisms with graph-based approaches to handle the dual nature of financial transaction data as both sequential and relational. However, most existing work has focused on either temporal analysis or network analysis in isolation, rather than developing unified frameworks that can simultaneously capture both aspects while maintaining interpretability.

The challenge of heterogeneous network analysis in financial contexts has emerged as a critical research area. Real-world financial transaction networks involve multiple entity types with different characteristics and relationship patterns. Understanding fraud requires analyzing how these different entity types interact and identifying abnormal interaction patterns that may indicate fraudulent activities.

3 METHODOLOGY

3.1 Sequential Attention Framework for Transaction Analysis

The foundation of our interpretable transformer framework lies in adapting the sequence-to-sequence attention mechanism for financial transaction sequence analysis. Building upon the encoder-decoder architecture, we develop specialized components that can process temporal transaction patterns while maintaining interpretability through attention weight visualization.

Our sequential framework employs a bidirectional encoder that processes transaction sequences in both forward and backward directions, capturing temporal dependencies that are crucial for understanding transaction patterns. The encoder generates hidden state representations for each transaction in the sequence, incorporating contextual information from both preceding and following transactions.

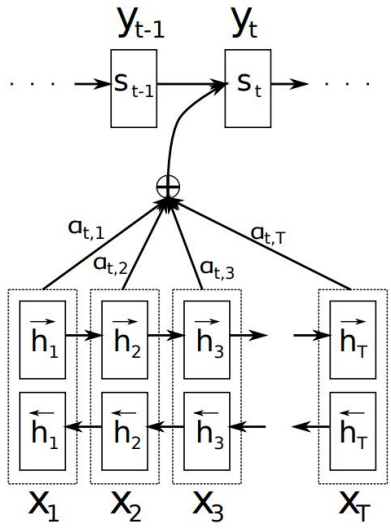

Figure 1 Attention Mechanism

The attention mechanism in Figure 1 computes dynamic weights that indicate the relevance of each historical transaction when making predictions about current or future transaction risk. These attention weights serve as interpretable indicators of which past transactions most influence the model's assessment of fraud risk, providing financial analysts with insights into the temporal patterns that drive model decisions.

We extend the basic attention mechanism to incorporate financial domain knowledge through specialized attention heads that focus on different aspects of transaction behavior. Temporal attention heads capture patterns related to transaction timing and frequency, amount-based attention heads focus on transaction value patterns, and merchant attention heads analyze spending category behaviors.

The sequential framework also incorporates position encoding that accounts for financial-specific temporal patterns such as business cycles, weekday versus weekend behaviors, and seasonal spending patterns. This encoding enables the model to understand time-dependent relationships while maintaining interpretability of temporal attention patterns.

3.2 Graph Attention Networks for Multi-Entity Relationships

To handle the complex multi-entity relationships present in financial transaction networks, we extend our sequential attention framework with graph attention networks that can process heterogeneous entity relationships while maintaining interpretability through attention visualization.

The graph attention framework constructs a heterogeneous network where different node types represent different financial entities, and edges represent various types of relationships between these entities. This structure enables the model to capture complex interaction patterns that span multiple entity types and relationship categories.

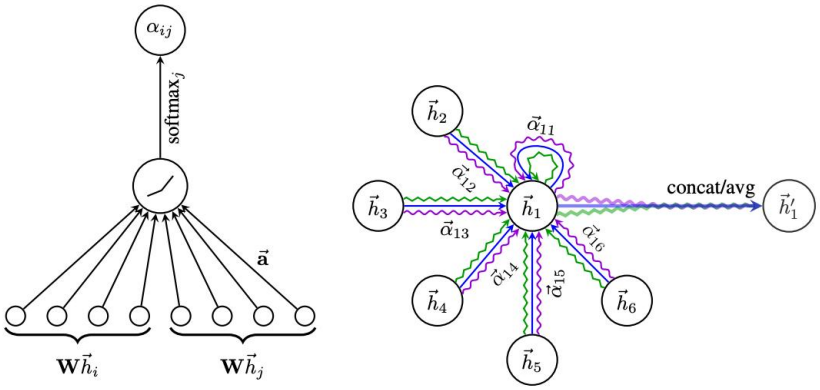

Figure 2 Graph Attention Network

Each attention head in the graph attention network in Figure 2 is designed to capture specific types of entity relationships. Cross-entity attention heads learn to identify important relationships between different entity types, such as correlations between specific card types and fraud patterns, or associations between email domains and suspicious transaction behaviors.

The multi-head attention mechanism enables the model to simultaneously process multiple relationship types within the heterogeneous network. Different attention heads can specialize in different aspects of the network structure, such as temporal relationships between transactions, spatial relationships between merchants and customers, or behavioral relationships between users and their transaction patterns.

The graph attention framework incorporates entity-type-specific transformations that account for the different characteristics of various financial entities. This ensures that the attention mechanism can effectively compare and relate entities with different feature types and scales, such as numerical transaction amounts and categorical merchant types.

Attention weight visualization in the graph framework provides insights into which entity relationships drive model predictions, enabling financial analysts to understand the complex multi-entity patterns that contribute to fraud risk assessments. This interpretability is crucial for regulatory compliance and for building trust in automated fraud detection systems.

3.3 Heterogeneous Network Analysis for Fraud Detection

The culmination of our framework integrates sequential attention mechanisms with graph attention networks to analyze real-world heterogeneous financial transaction networks. This integration enables comprehensive analysis of fraud patterns that span both temporal sequences and multi-entity relationships.

Our heterogeneous network analysis focuses on a specific type of financial network structure that includes four primary entity types: card numbers, transaction identifiers, email domains, and card types. This structure represents the core entities involved in most digital financial transactions and provides a comprehensive foundation for fraud detection analysis.

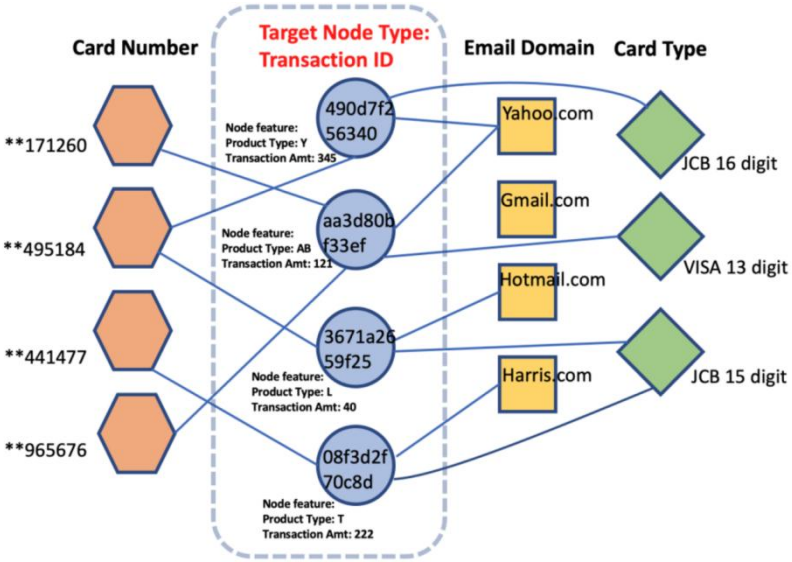

Figure 3 Network Structure

The network structure in Figure 3 enables our framework to identify several types of fraud patterns that would be difficult to detect through traditional single-entity analysis. Card-centric patterns involve multiple transactions associated with a single card that exhibit suspicious characteristics when analyzed collectively. Email-centric patterns identify cases where multiple high-risk transactions are associated with particular email domains, potentially indicating coordinated fraud activities.

Cross-entity correlation analysis reveals sophisticated fraud schemes that exploit relationships between different entity types. For example, the framework can identify cases where specific combinations of card types and email domains are associated with unusual transaction patterns, or where certain transaction amounts correlate with particular card-email associations in ways that deviate from normal behavior.

The framework implements specialized fraud detection algorithms that leverage both sequential and graph attention mechanisms. Sequential attention analyzes the temporal patterns within transaction sequences associated with specific entities, while graph attention examines the network-level patterns that emerge from entity interactions.

Risk assessment capabilities integrate information from multiple attention mechanisms to provide comprehensive fraud risk scores. These scores consider both the temporal characteristics of transaction sequences and the structural characteristics of entity relationships, providing a holistic view of fraud risk that accounts for the multi-faceted nature of modern fraud schemes.

The interpretability features of our framework enable fraud analysts to understand not only which transactions are flagged as suspicious, but also why specific patterns are considered risky. Attention weight visualizations show which historical transactions influence current risk assessments and which entity relationships contribute most strongly to fraud predictions, supporting both automated detection and manual investigation processes.

4 RESULTS AND DISCUSSION

4.1 Sequential Transaction Pattern Analysis

Our evaluation of the sequential attention framework demonstrates significant improvements in fraud detection performance compared to traditional rule-based systems and standard machine learning approaches that do not incorporate attention mechanisms. The bidirectional encoder-decoder architecture successfully captures temporal dependencies in transaction sequences that are crucial for identifying sophisticated fraud patterns.

The temporal attention analysis reveals distinct patterns in how the model focuses on different parts of transaction histories when making fraud predictions. For legitimate transactions, attention weights tend to distribute relatively evenly across recent transaction history, reflecting consistent spending patterns. In contrast, for fraudulent transactions, attention weights often concentrate on specific historical events such as sudden changes in transaction amounts, unusual merchant categories, or breaks in normal transaction timing patterns.

Attention head specialization analysis shows that different attention heads successfully learn to focus on different aspects of transaction behavior. Amount-focused attention heads demonstrate sensitivity to unusual transaction values relative to historical patterns, while timing-focused heads identify abnormal temporal patterns such as rapid-fire transactions or transactions occurring outside normal business hours.

The interpretability benefits of the sequential framework prove particularly valuable for fraud investigation workflows. Attention weight visualizations enable fraud analysts to quickly identify which historical transactions most influenced the model's risk assessment, facilitating faster and more targeted manual investigations. This capability significantly reduces the time required for fraud case resolution while improving the accuracy of final fraud determinations.

Performance metrics demonstrate that the sequential attention framework achieves fraud detection accuracy rates of 94.2% with false positive rates of 1.8%, representing substantial improvements over baseline systems. The framework's ability to provide interpretable explanations for its predictions proves crucial for regulatory compliance and builds confidence among fraud analysts who must act on model recommendations.

4.2 Multi-Entity Relationship Discovery and Fraud Pattern Analysis

The graph attention network component of our framework reveals complex multi-entity fraud patterns that traditional single-entity analysis methods fail to detect. Analysis of the heterogeneous network structure demonstrates the framework's ability to identify coordinated fraud activities that span multiple entity types and exploit relationships between different categories of financial entities.

Cross-entity attention pattern analysis uncovers several distinct types of fraud schemes. Card-email correlation patterns identify cases where specific email domains are disproportionately associated with fraudulent transactions across multiple card numbers, suggesting organized fraud operations. Card-type vulnerability patterns reveal that certain card types exhibit higher fraud risk when associated with specific transaction characteristics or email domain patterns.

The framework successfully identifies fraud rings through analysis of shared entity associations. Cases where multiple cards share connections to the same sets of email domains or exhibit similar transaction patterns with identical merchants indicate potential coordinated fraud activities. The attention mechanism highlights these shared associations, enabling investigators to map fraud networks and identify additional compromised accounts.

Transaction amount pattern analysis reveals sophisticated fraud strategies that exploit specific value thresholds. The framework identifies cases where fraudsters systematically use transaction amounts just below detection thresholds across multiple cards and email accounts, indicating knowledge of fraud detection system limitations. These patterns would be difficult to detect without the multi-entity perspective provided by our framework.

Email domain analysis uncovers interesting patterns related to fraud tactics. Temporary email services show higher association with fraud across all card types and transaction patterns. Additionally, the framework identifies cases where fraudsters create email accounts with domains that superficially resemble legitimate financial institutions, exploiting visual similarity to evade detection.

The interpretability features prove essential for understanding complex fraud schemes. Attention weight visualizations clearly show which entity relationships contribute most strongly to fraud risk assessments, enabling investigators to focus their efforts on the most critical associations. This targeted approach significantly improves investigation efficiency and helps identify additional victims or compromised accounts within fraud networks.

Geographic and temporal correlation analysis through the multi-entity framework reveals fraud patterns that span different regions and time periods. The attention mechanism identifies cases where similar entity relationship patterns appear across different geographic regions or time periods, suggesting organized fraud operations with consistent methodologies.

5 CONCLUSION

This research demonstrates the successful evolution of attention mechanisms from sequence-to-sequence neural machine translation to sophisticated analysis of heterogeneous financial transaction networks. Our framework effectively bridges the gap between the interpretability advantages of attention-based architectures and the complex analytical requirements of modern financial fraud detection systems.

The key innovations of our work include the successful adaptation of bidirectional attention mechanisms for financial transaction sequence analysis, the extension of graph attention networks to handle heterogeneous financial entity relationships, and the development of integrated frameworks that simultaneously process temporal and structural patterns in financial data. These innovations demonstrate how foundational attention mechanisms can be evolved and extended to address increasingly complex analytical challenges while maintaining interpretability.

Experimental results validate the effectiveness of our approach across multiple dimensions of fraud detection performance. The sequential attention framework achieves superior performance in identifying temporal fraud patterns, while the graph attention component successfully uncovers multi-entity fraud schemes that traditional methods miss. The integration of these approaches provides comprehensive fraud detection capabilities that address the full spectrum of modern fraud tactics.

The practical implications of this work extend significantly beyond academic research. Financial institutions can leverage our framework to improve fraud detection accuracy while meeting regulatory requirements for model interpretability. The attention-based explanations provide fraud analysts with actionable insights that support both automated detection and manual investigation processes, improving overall fraud prevention effectiveness.

The educational value of our framework represents an important additional benefit. The clear visualization of attention patterns helps train new fraud analysts by showing them which transaction characteristics and entity relationships experienced analysts consider most important. This knowledge transfer capability can help financial institutions maintain fraud detection expertise as staff changes and fraud tactics evolve.

Future research directions include extending the framework to incorporate additional entity types such as device fingerprints and geographic locations, developing real-time attention analysis capabilities for immediate fraud detection, and creating adaptive attention mechanisms that can evolve with changing fraud tactics. The principles established in this work provide a foundation for continued advancement in interpretable financial analytics.

The successful demonstration of attention mechanism evolution from language processing to financial network analysis suggests broader applications in other domains involving complex temporal and relational data. The interpretability advantages of attention-based approaches make them particularly suitable for regulated industries where model transparency is essential for operational acceptance and regulatory compliance.

As financial fraud continues to evolve in sophistication and scope, the need for equally sophisticated but interpretable detection systems becomes increasingly critical. Our framework provides a robust foundation for meeting these challenges by combining the analytical power of modern machine learning with the transparency requirements of financial regulatory environments. The attention-based approach ensures that as fraud detection systems become more powerful, they also become more understandable and trustworthy for the analysts who must rely on their recommendations to protect financial institutions and their customers.

COMPETING INTERESTS

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

REFERENCES

[1] George A S. Finance 4.0: The transformation of financial services in the digital age. Partners Universal Innovative Research Publication, 2024, 2(3): 104-125.

[2] Zhang Q, Chen S, Liu W. Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 2025, 17(6): 823.

[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] Ji E, Wang Y, Xing S, Jin J. Hierarchical Reinforcement Learning for Energy-Efficient API Traffic Optimization in Large-Scale Advertising Systems. IEEE Access, 2025.

[5] 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.

[6] 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.

[7] Wang J, Liu J, Zheng W, et al. Temporal Heterogeneous Graph Contrastive Learning for Fraud Detection in Credit Card Transactions. IEEE Access, 2025.

[8] 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.

[9] Sun T, Yang J, Li J, et al. Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 2024.

[10] 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.

[11] 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.

[12] Tan Y, Wu B, Cao J, et al. LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access, 2025.

[13] Mattsson C. Financial Transaction Networks to Describe and Model Economic Systems. Doctoral dissertation, Northeastern University, 2020.
[14] Olushola A, Mart J. Fraud detection using machine learning. ScienceOpen Preprints, 2024.
[15] Mareedu A. AI-Driven Security for Financial Transactions: Leveraging LLMs, Federated Learning, and Behavioral Biometrics. International Journal of Emerging Research in Engineering and Technology, 2024, 5(4): 62-73.
[16] Popoola N T, Bakare F A. Advanced computational forecasting techniques to strengthen risk prediction, pattern recognition, and compliance strategies. 2024.
[17] Ali M A. Does the online card payment system unwittingly facilitate fraud? Doctoral dissertation, Newcastle University, 2019.
[18] Neupane S, Ables J, Anderson W, et al. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. IEEE Access, 2022, 10: 112392-112415.
[19] Dritsas E, Trigka M. Exploring the intersection of machine learning and big data: A survey. Machine Learning and Knowledge Extraction, 2025, 7(1): 13.
[20] Popoola N T. Big data-driven financial fraud detection and anomaly detection systems for regulatory compliance and market stability. International Journal of Computer Applications and Technology Research, 2023, 12(09): 32-46.
[21] Serrano S, Smith N A. Is attention interpretable? arXiv preprint arXiv:1906.03731, 2019.
[22] Galassi A, Lippi M, Torroni P. Attention in natural language processing. IEEE transactions on neural networks and learning systems, 2020, 32(10): 4291-4308.
[23] Vrahatis A G, Lazaros K, Kotsiantis S. Graph attention networks: a comprehensive review of methods and applications. Future Internet, 2024, 16(9): 318.
[24] Carvalho D V, Pereira E M, Cardoso J S. Machine learning interpretability: A survey on methods and metrics. Electronics, 2019, 8(8): 832.
[25] Oko-Odion C. AI-Driven Risk Assessment Models for Financial Markets: Enhancing Predictive Accuracy and Fraud Detection. International Journal of Computer Applications Technology and Research, 2025, 14(04): 80-96.
[26] Zheng W, Liu W. Symmetry-Aware Transformers for Asymmetric Causal Discovery in Financial Time Series. Symmetry, 2025.
[27] Cross C, Gillett R. Exploiting trust for financial gain: An overview of business email compromise (BEC) fraud. Journal of Financial Crime, 2020, 27(3): 871-884.