International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 10 Issue 2, Mar-Apr 2026 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

A Hybrid Graph-Neural and Generative AI Framework for Real-Time Fraud Detection in Cloud-Native Financial Systems

Oluwabukunmi Adubi[1], Abba Giza ADB[2]

[1]Robert H. Smith School of Business, University of Maryland, College Park, Maryland
[2]College of Education, Texas Southern University, Houston, Texas

ABSTRACT

The increasing digitization of financial services, driven by cloud computing, fintech innovation, and real-time payment systems, has significantly amplified the complexity and scale of financial fraud. Traditional fraud detection approaches, including rule-based systems and conventional machine learning models, are often inadequate for capturing the relational and dynamic nature of modern fraud schemes. In response, this study proposes a hybrid graph-neural and generative artificial intelligence (AI) framework for real-time fraud detection in cloud-native financial systems. The proposed framework integrates three core components: a transaction graph modeling layer that represents financial interactions as dynamic graphs; a graph neural network (GNN) detection layer that captures complex relational patterns and identifies fraudulent behavior; and a generative AI explanation module that provides interpretable, natural-language explanations for flagged transactions. These components are deployed within a cloud-native, real-time processing architecture, enabling scalable, low-latency fraud detection across high-volume transaction streams. To evaluate the effectiveness of the framework, experiments were conducted using financial and synthetic datasets within a distributed cloud environment. The results demonstrate that the GNN-based model significantly improves fraud detection performance by effectively capturing network-level dependencies, while the generative AI module enhances interpretability and supports investigative decision-making. The system also achieves competitive real-time performance in terms of inference latency and throughput, highlighting its suitability for deployment in operational financial environments. However, the study identifies trade-offs between detection accuracy and system latency, as well as challenges related to computational complexity, data privacy, and model governance. Despite these limitations, the findings underscore the importance of integrating relational modeling, explainable AI, and cloud-native deployment in modern fraud detection systems. The proposed framework contributes to the advancement of intelligent financial security by offering a scalable, interpretable, and high-performance solution for detecting fraud in increasingly complex digital ecosystems.

How to cite this paper: Oluwabukunmi Adubi | Abba Giza ADB "A Hybrid Graph-Neural and Generative AI Framework for Real-Time Fraud Detection in Cloud-Native Financial Systems" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-10 | Issue-2, April 2026, pp.750-760, URL: www.ijtsrd.com/papers/ijtsrd116428.pdf

IJTSRD116428

KEYWORDS: *Graph neural networks; generative artificial intelligence; fraud detection; financial transaction analysis; explainable AI; cloud computing; distributed systems; real-time analytics; anomaly detection; enterprise AI architecture.*

1. INTRODUCTION

The rapid expansion of digital banking, fintech ecosystems, instant payments, and cloud-enabled financial services has significantly increased the scale, speed, and complexity of financial transactions. In Nigeria, this transformation is reinforced by the Central Bank of Nigeria's Payments System Vision 2025, which promotes electronic payments, interoperability, and secure digital financial infrastructure (CBN, 2022). However, this evolution has also intensified fraud risks, as attackers increasingly exploit advanced techniques such as social engineering, malware, botnets, and AI-enabled fraud schemes. Consequently, fraud detection in modern financial

systems must move beyond static controls toward continuous, intelligent, and scalable monitoring mechanisms capable of operating in dynamic and high-velocity environments (European Payments Council, 2025). The magnitude of the fraud challenge is further underscored by recent empirical evidence. The Association for Financial Professionals reports that approximately 79%

of organizations experienced attempted or actual payment fraud in 2024, with business email compromise emerging as a dominant threat vector (AFP, 2025). In addition, recent scholarly reviews highlight that financial fraud detection now operates under conditions of extreme class imbalance, evolving attack strategies, and increasing regulatory scrutiny. These factors complicate real-time decision-making and place significant pressure on fraud detection systems to achieve both high accuracy and low latency in cloud-native environments (Chen et al., 2025; Yang et al., 2026).

Despite these advancements, traditional fraud detection approaches remain limited in their effectiveness. Rule-based systems and conventional machine learning models often treat transactions as independent events, failing to capture the complex relational dependencies that characterize modern fraud schemes. This limitation reduces their ability to detect coordinated or network-based fraud patterns involving multiple entities such as accounts, devices, and transaction pathways. Furthermore, although deep learning models have improved predictive performance, they often lack interpretability, making it difficult for investigators and regulators to understand the rationale behind flagged transactions. This challenge has been widely identified as a critical barrier to the adoption of AI-driven fraud detection systems in regulated financial environments (Motie & Raahemi, 2024; Chen et al., 2025).

In response to these limitations, recent research has shifted toward graph-based and generative AI approaches. Graph neural networks (GNNs) enable the modeling of financial transactions as interconnected networks, allowing for the detection of complex relational patterns and dynamic fraud behaviors that traditional models often overlook (Cheng et al., 2025). Complementing this, generative AI techniques including large language models and deep generative models such as GANs and VAEs are increasingly used to enhance anomaly detection and provide human-interpretable explanations for flagged transactions. These models support investigative decision-making by translating complex analytical outputs into meaningful insights, thereby bridging the gap between detection accuracy and explainability (Desai et al., 2024; Park, 2024; Tang et al., 2025). Together, these developments highlight a growing paradigm shift toward integrated, intelligent, and explainable fraud detection systems.

1.1. Problem Statement

Despite the progress made in graph-based fraud analytics and AI-enabled anomaly detection, important gaps remain in translating these advances into operational financial systems. A significant limitation in existing studies is the predominant focus on predictive accuracy, often at the expense of deployment considerations such as real-time processing, scalability, and system architecture. In practice, fraud detection systems must operate within high-throughput, low-latency environments, particularly in cloud-native financial platforms where transactions are processed continuously. Empirical studies on streaming architectures indicate that latency, throughput, and infrastructure design are critical determinants of system effectiveness. For instance, Daksa and Kemala (2025) demonstrate that Kafka-based real-time fraud detection pipelines require carefully optimized stream-processing frameworks to achieve acceptable performance, highlighting that model sophistication alone is insufficient without robust deployment architecture.

In addition, the issue of interpretability remains a major barrier to the adoption of advanced AI-driven fraud detection systems. Even when high-performing models such as deep learning and graph neural networks are employed, their outputs are often opaque, making it difficult for fraud analysts, auditors, and regulators to understand the rationale behind flagged transactions. Recent reviews emphasize that explainability, governance, regulatory compliance, and trustworthiness continue to be unresolved challenges in financial fraud detection systems (Chen et al., 2025; Yang et al., 2026). This lack of transparency is particularly problematic in regulated financial environments, where decisions must be justified and auditable.

Consequently, there is a clear gap in the literature and practice: the absence of an integrated framework that simultaneously addresses relational fraud detection, explainability, and real-time cloud-native deployment. While graph-based models effectively capture complex transaction relationships and generative AI shows promise for enhancing interpretability, these approaches are

rarely unified within a single, scalable architecture suitable for production environments. This study therefore responds to this gap by proposing a hybrid framework that combines graph neural networks for relational fraud detection, generative AI for anomaly explanation, and a cloud-native pipeline for real-time deployment.

1.2. Aim and Objectives of the Study

The aim of this study is to design and evaluate a hybrid graph-neural and generative AI framework for real-time fraud detection in cloud-native financial systems. Specifically, the study seeks to:

1. model transaction relationships using graph neural networks for improved fraud detection;
2. integrate generative AI for anomaly explanation and analyst decision support;
3. develop a real-time inference pipeline suitable for cloud-native financial platforms; and
4. benchmark the framework in terms of detection performance, latency, throughput, and scalability.

1.3. Contributions of the Paper

This paper makes four major contributions. First, it proposes a hybrid GNN–Generative AI fraud detection architecture that combines relational transaction modeling with language-enabled anomaly interpretation. Second, it develops a real-time cloud-native inference pipeline capable of supporting high-volume transaction monitoring. Third, it introduces an explainable anomaly interpretation mechanism that can assist analysts in understanding flagged patterns. Fourth, it provides experimental benchmarking of the proposed framework using performance, scalability, and latency-oriented evaluation criteria. These contributions respond directly to the gaps identified in recent studies on AI fraud detection, graph-based finance analytics, and real-time streaming architectures.

2. Literature Review

2.1. Fraud Detection in Financial Systems

Fraud detection in financial systems has evolved from traditional rule-based approaches to more sophisticated data-driven techniques. Early systems relied on expert-defined rules and statistical thresholds to flag suspicious transactions; however, these approaches are often rigid and unable to adapt to emerging fraud patterns. Machine learning models, including logistic regression, decision trees, random forests, and support vector machines, introduced greater adaptability by learning patterns from historical data. More recently, deep learning techniques such as neural networks and autoencoders have been applied to capture complex, nonlinear relationships in transaction data (Chen et al., 2025; Yang et al., 2026).

Despite these advancements, several persistent challenges remain. Financial fraud datasets are typically highly imbalanced, with fraudulent transactions representing a very small proportion of total activity, which can bias model performance. In addition, fraud patterns evolve over time, creating concept drift that reduces the effectiveness of static models. Real-time detection requirements further complicate system design, as models must process large volumes of streaming data with minimal latency while maintaining high accuracy (Motie & Raahemi, 2024; Dal Pozzolo et al., 2022). These challenges underscore the need for more adaptive and context-aware fraud detection techniques.

2.2. Graph Neural Networks for Transaction Relationship Modeling

Recent studies emphasize that financial fraud is inherently relational, making graph-based approaches particularly suitable for detection tasks. In graph representations of financial systems, entities such as customers, accounts, merchants, and devices are modeled as nodes, while transactions and interactions are represented as edges. Graph neural networks (GNNs) leverage this structure by learning node embeddings through iterative message passing, where information is aggregated from neighboring nodes to capture both local and global patterns (Cheng et al., 2025; Wu et al., 2023). Several GNN architectures have been applied in fraud detection. Graph Convolutional Networks (GCNs) aggregate neighborhood information to learn structural representations, while Graph Attention Networks (GATs) assign different weights to neighboring nodes based on their importance. Temporal Graph Neural Networks further extend these models by incorporating time-dependent interactions, enabling the detection of evolving fraud behaviors (Zhang et al., 2024). These approaches have demonstrated strong capability in identifying complex fraud structures such as money laundering rings, coordinated account collusion, and circular transaction patterns that are difficult to detect using traditional methods (Motie & Raahemi, 2024).

2.3. Generative AI for Anomaly Explanation

While advanced models improve detection accuracy, interpretability remains a critical concern in financial fraud systems. Generative AI has recently emerged as a promising approach for enhancing explainability and supporting investigative decision-

making. Large language models (LLMs), for instance, can generate human-readable explanations of model outputs, helping analysts understand why specific transactions are flagged as suspicious (Desai et al., 2024; Park, 2024).

In addition, generative explanation techniques such as counterfactual reasoning enable systems to describe how a transaction would need to change to be considered non-fraudulent, thereby providing actionable insights. Deep generative models, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have also been applied to anomaly detection by learning the distribution of normal transaction behavior and identifying deviations (Tang et al., 2025). These capabilities allow generative AI to bridge the gap between complex model outputs and the need for transparent, interpretable fraud alerts in regulated environments.

2.4. Real-Time Inference Pipelines for Fraud Detection

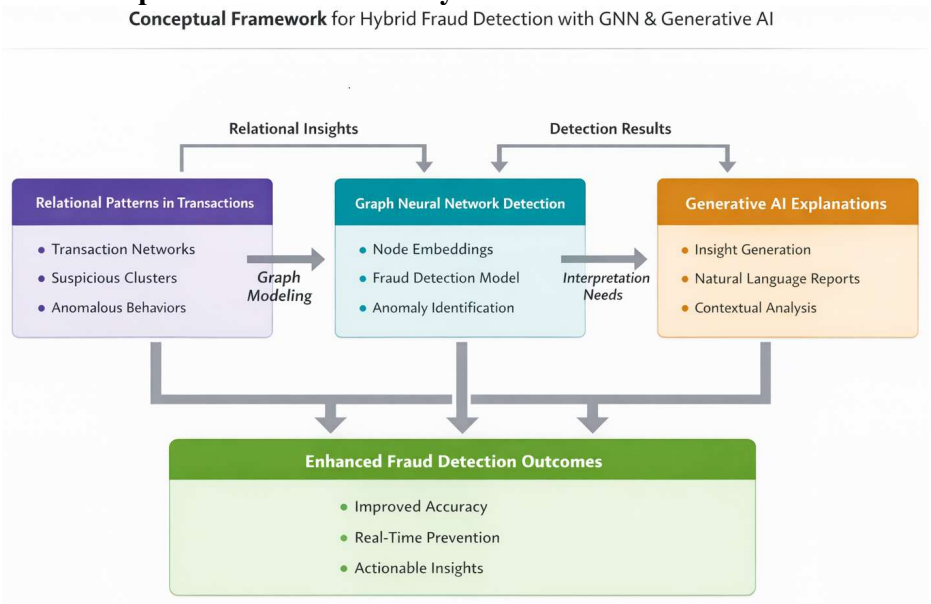
The effectiveness of fraud detection systems depends not only on model performance but also on their ability to operate in real time. Modern financial systems require continuous monitoring of high-velocity transaction streams, necessitating the use of event-driven architectures and stream processing frameworks. Technologies such as Apache Kafka, Apache Flink, and Apache Spark Streaming are widely used to support real-time data ingestion, processing, and model inference (Daksa & Kemala, 2025). Low-latency inference is particularly critical, as fraud detection decisions must often be made within milliseconds to prevent unauthorized transactions. However, achieving sub-second latency while maintaining high accuracy presents significant challenges. These include managing data throughput, ensuring efficient feature extraction, and deploying scalable model-serving infrastructure. Furthermore, integrating machine learning models into streaming pipelines requires careful orchestration to avoid bottlenecks and ensure system reliability (Khan et al., 2023).

2.5. Cloud-Native Fraud Detection Architectures

Cloud-native architectures have become central to modern financial analytics due to their scalability, flexibility, and resilience. These architectures typically employ microservices-based designs, where system components are deployed as independent services that can be scaled dynamically. Container orchestration platforms such as Kubernetes enable efficient resource management and automated deployment, while serverless computing models provide on-demand scalability for processing variable workloads (Zhou et al., 2023).

Distributed storage systems and data lakes further support large-scale data processing, enabling organizations to manage and analyze vast volumes of transaction data. However, designing cloud-native fraud detection systems involves trade-offs between scalability, latency, and cost. Ensuring high availability, fault tolerance, and security is essential, particularly in financial environments where system failures or delays can have significant consequences. As such, cloud-native architectures must be carefully designed to support both real-time analytics and robust operational performance (Chen et al., 2025).

Figure 1: Conceptual Framework for Hybrid GNN–Generative AI Fraud Detection



Source: Author's Computation, 2026

Based on the reviewed literature, Figure 1 presents the conceptual framework of the study, illustrating the relationships between relational transaction patterns, graph neural network-based detection, and generative AI-driven explanation, leading to enhanced fraud detection outcomes.

2.6. Research Gap

Although significant progress has been made in both graph-based fraud detection and explainable AI, existing studies largely address these areas in isolation. Graph neural networks have demonstrated strong performance in modeling relational fraud patterns, while generative AI has shown potential in enhancing interpretability and decision support. However, there is limited research that integrates these approaches within a unified framework capable of real-time operation in cloud-native environments.

Furthermore, many studies focus primarily on model development without adequately addressing deployment challenges such as scalability, latency, and system integration. As a result, there remains a gap between theoretical advancements and practical implementation in financial institutions. This study seeks to address this gap by proposing a hybrid framework that combines GNN-based fraud detection, generative AI-driven explanation, and a scalable real-time cloud-native architecture.

3. Proposed Hybrid Fraud Detection Framework

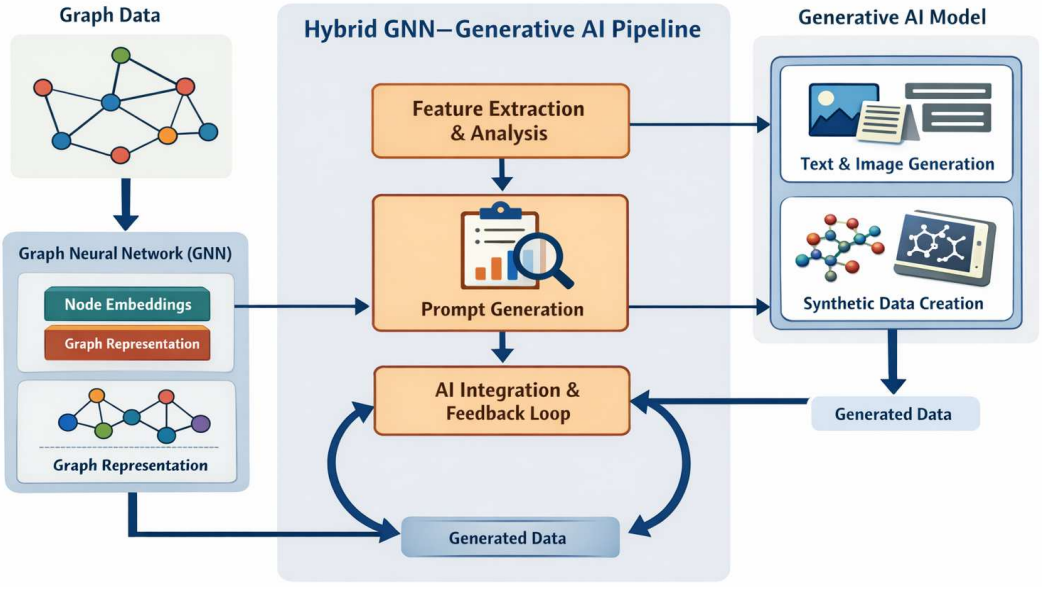
3.1. System Overview

The proposed framework adopts a three-layer hybrid architecture designed to operate within a cloud-native, real-time processing environment. The first layer, the transaction graph modeling layer, transforms streaming financial data into a structured graph that captures relationships among entities such as accounts, devices, merchants, and banks. This relational representation is essential because financial fraud typically manifests through interconnected activities rather than isolated transactions. Prior studies emphasize that graph-based representations significantly improve fraud detection by preserving structural dependencies and network dynamics inherent in financial systems (Motie & Raahemi, 2024; Cheng et al., 2025).

The second layer, the graph neural network (GNN) detection layer, leverages these graph structures to learn embeddings and identify suspicious patterns through neighborhood aggregation and message passing. The third layer, the generative AI explanation layer, interprets model outputs and generates human-readable explanations for flagged transactions. This layered design reflects recent research suggesting that effective fraud detection systems must integrate relational modeling, predictive analytics, and explainability within a unified framework (Chen et al., 2025; Yang et al., 2026).

These components are deployed within a cloud-native streaming architecture, enabling continuous ingestion, processing, and analysis of financial transactions. Real-time processing is critical in financial environments, where delays in detection can result in significant losses. Studies on streaming systems show that scalable event-driven architectures are essential for maintaining low latency and high throughput in fraud detection pipelines (Daksa & Kemala, 2025).

Figure 2: Hybrid GNN–Generative AI Fraud Detection Architecture



Source: Author's Computation, 2026

Figure 2 presents the overall architecture of the proposed hybrid graph neural network–generative AI fraud detection framework, illustrating the flow of transaction data from ingestion through graph modeling, detection, explanation, and governance layers.

3.2. Graph Construction for Financial Transactions

In this framework, financial transactions are modeled as a heterogeneous graph (G = (V, E, X)), where (V) represents entities (nodes), (E) represents relationships (edges), and (X) denotes associated features. Nodes may include customer accounts, payment devices, merchants, banks, and IP addresses, while edges represent transactions, shared devices, shared IP addresses, and other financial relationships. This structure enables the integration of both direct transactional links and indirect contextual associations.

Node features may include attributes such as transaction frequency, account tenure, geographic location, and behavioral patterns, while edge features capture transaction-specific details such as amount, time, and channel. Importantly, the graph is continuously updated to reflect incoming transactions, forming a dynamic or temporal graph. Temporal information is critical, as fraud often depends on transaction sequences and timing patterns rather than static relationships alone (Zhang et al., 2024).

This graph-based representation allows the detection of complex fraud patterns such as money laundering networks, coordinated account collusion, and cyclic transaction flows, which are difficult to identify using traditional tabular data approaches. Empirical studies confirm that modeling such relational dependencies significantly enhances fraud detection performance (Motie & Raahemi, 2024).

3.3. Graph Neural Network Fraud Detection Model

The core detection mechanism is implemented using a graph neural network (GNN), which learns representations of nodes and edges through iterative message passing. Each node is initialized with a feature vector, and during training, information is aggregated from neighboring nodes to produce graph embeddings that capture both local and global structural information. These embeddings are then used to compute a fraud probability score for transactions or entities.

Different GNN architectures can be employed depending on the characteristics of the data. Graph Convolutional Networks (GCNs) aggregate neighborhood features to learn structural patterns, while Graph Attention Networks (GATs) assign varying importance to neighboring nodes, improving the detection of influential relationships. For dynamic financial environments, Temporal Graph Networks (TGNs) incorporate time-dependent interactions, enabling the model to capture evolving fraud behaviors (Zhang et al., 2024).

The strength of GNN-based models lies in their ability to detect high-order and nonlinear relationships across interconnected entities. This makes them particularly effective for identifying sophisticated fraud schemes such as coordinated transactions, layered money movements, and anomalous interaction patterns. Recent reviews consistently report that GNNs outperform traditional machine learning models in relational fraud detection tasks due to their ability to exploit graph topology and contextual dependencies (Motie & Raahemi, 2024; Yang et al., 2026).

3.4. Generative AI-Based Fraud Explanation Module

To address the challenge of model interpretability, the framework incorporates a generative AI-based explanation module that translates detection outputs into meaningful insights for analysts. Once a transaction is flagged by the GNN model, the explanation module analyzes the underlying graph structure, feature contributions, and temporal patterns to generate natural-language explanations.

This module leverages advances in generative AI, particularly large language models (LLMs), which have demonstrated strong capabilities in summarization, reasoning, and contextual interpretation. Studies show that generative AI can enhance fraud detection systems by providing explainable outputs, supporting investigative workflows, and improving decision-making transparency (Park, 2024). Additionally, deep generative models such as GANs and VAEs can complement detection by identifying deviations from normal transaction behavior and supporting anomaly interpretation (Tang et al., 2025).

The explanation module performs three key functions: (i) identifying suspicious graph patterns, such as dense clusters or unusual connections; (ii) translating model predictions into human-readable narratives; and (iii) providing contextual insights that assist fraud analysts in understanding the nature of the anomaly. For example, the system may indicate that a transaction was flagged due to its involvement in a high-frequency transaction

cluster linked to multiple accounts sharing the same device or IP address. By enhancing interpretability, this component improves trust, regulatory compliance, and operational usability of the fraud detection system (Yang et al., 2026).

4. Real-Time Cloud-Native Fraud Detection Architecture

4.1. Transaction Data Ingestion Layer

The first layer of the architecture focuses on the real-time ingestion of financial transaction data from banking systems, payment gateways, mobile applications, and third-party financial platforms. In modern cloud-native environments, this is typically implemented using event-driven architectures supported by distributed messaging systems such as Apache Kafka or similar event streaming platforms. These systems enable high-throughput, fault-tolerant ingestion of transaction events while ensuring scalability and reliability. Streaming ingestion is critical because fraud detection must operate continuously on live data rather than in batch mode. Studies on real-time fraud detection architectures highlight that event streaming platforms provide the necessary backbone for handling high-velocity transaction flows, ensuring that data is processed with minimal delay and without loss (Daksa & Kemala, 2025; Khan et al., 2023). This layer therefore acts as the entry point into the fraud detection pipeline, enabling downstream components to access timely and consistent transaction data.

4.2. Graph Update and Feature Engineering Layer

Following ingestion, transaction data is processed within the graph update and feature engineering layer, where incoming events are transformed into structured graph representations and enriched with relevant features. In this layer, each new transaction updates the existing dynamic transaction graph by adding new nodes (e.g., accounts, devices, merchants) or edges (e.g., transaction links, shared attributes). This continuous update mechanism ensures that the graph reflects the most recent state of the financial network, which is essential for detecting time-sensitive fraud patterns. Real-time graph construction presents both computational and architectural challenges. Unlike static graphs, dynamic financial graphs must support frequent updates while maintaining consistency and scalability. Recent research emphasizes that incremental graph updating techniques and streaming graph processing frameworks are necessary to handle evolving relationships and large-scale transaction volumes efficiently (Zhang et al., 2024; Motie & Raahemi, 2024).

In parallel, the layer performs feature engineering, extracting both node-level and edge-level features required by the GNN model. Node features may include transaction frequency, account age, behavioral profiles, and risk indicators, while edge features capture attributes such as transaction amount, timestamp, and channel. Additionally, graph-based features such as node degree, clustering coefficients, centrality measures, and subgraph patterns are computed to capture relational and structural information. These features are particularly valuable in identifying anomalies that arise from unusual connectivity patterns rather than individual transaction attributes.

To support real-time processing, feature extraction must be optimized for low latency. This often involves maintaining precomputed feature stores, incremental updates, and efficient graph indexing mechanisms. Studies indicate that combining streaming data processing with graph-based feature engineering significantly improves fraud detection performance, especially in environments characterized by high data velocity and evolving fraud behaviors (Cheng et al., 2025; Chen et al., 2025).

4.3. Model Inference and Fraud Detection Layer

The model inference layer is responsible for applying trained fraud detection models to incoming transaction data in real time. In this architecture, the GNN model is deployed as a scalable service capable of processing graph-structured inputs and producing fraud risk scores. The inference process involves retrieving relevant subgraphs or neighborhood information, computing embeddings, and classifying transactions or entities as fraudulent or legitimate. Low-latency inference is a critical requirement in financial systems, as fraud detection decisions must often be made before transaction settlement. Delays in detection can result in financial losses and increased risk exposure. As such, model serving frameworks must be optimized for speed, scalability, and reliability. Techniques such as model caching, batch inference for micro-streams, and hardware acceleration (e.g., GPUs) are often employed to achieve sub-second response times (Daksa & Kemala, 2025; Yang et al., 2026).

Furthermore, the inference layer may incorporate hybrid detection strategies that combine supervised classification with anomaly detection, particularly in scenarios where fraud patterns are rare or evolving.

This enhances the system's ability to detect previously unseen fraud behaviors while maintaining high accuracy on known patterns.

4.4. Generative AI Explanation Service

The generative AI explanation service is integrated into the architecture to provide interpretability and support investigative workflows. Once a transaction is flagged by the detection model, the explanation service generates a natural-language description of the underlying reasons for the alert. This involves analyzing graph structures, feature contributions, and temporal patterns associated with the flagged transaction. Recent advances in generative AI, particularly large language models (LLMs), have demonstrated strong capabilities in translating complex model outputs into human-understandable explanations. These models can synthesize multiple data signals and present them in a coherent narrative, thereby enhancing transparency and decision-making in fraud detection systems (Desai et al., 2024; Park, 2024).

In practical terms, the explanation service may highlight factors such as abnormal transaction frequency, unusual connections between accounts, or participation in suspicious transaction clusters. By providing contextual insights, this layer improves trust in the system, facilitates regulatory compliance, and enables faster and more effective fraud investigations (Chen et al., 2025; Yang et al., 2026).

4.5. Monitoring and Model Governance

The final layer of the architecture focuses on continuous monitoring and governance to ensure system reliability, accuracy, and compliance. Given the dynamic nature of financial fraud, models must be continuously evaluated and updated to remain effective. One key aspect is model drift detection, which identifies changes in data distribution or fraud patterns that may degrade model performance over time. In addition, the system tracks fraud detection accuracy using metrics such as precision, recall, and false positive rates. Monitoring these metrics in real time allows organizations to identify performance degradation and trigger model retraining when necessary. System-level monitoring is also essential to ensure high availability, fault tolerance, and latency guarantees, particularly in cloud-native environments where services are distributed across multiple nodes. Governance frameworks further ensure that the fraud detection system complies with regulatory requirements, including explainability, auditability, and data privacy standards. Recent studies emphasize that effective AI deployment in financial systems requires robust governance mechanisms to manage risks associated with model bias, lack of transparency, and operational failures (Yang et al., 2026). By integrating monitoring and governance into the architecture, the proposed framework supports sustainable and trustworthy fraud detection operations.

5. Experimental Evaluation and Benchmarking

5.1. Experimental Setup

The experimental evaluation is designed to assess the effectiveness of the proposed hybrid framework in terms of detection accuracy, real-time performance, scalability, and interpretability. The study utilizes both benchmark financial fraud datasets and synthetic transaction data to simulate real-world conditions. Publicly available datasets, such as credit card fraud datasets and graph-based fraud benchmarks, are commonly used in the literature due to their structured labels and reproducibility, while synthetic datasets enable the modeling of complex relational fraud scenarios such as collusion and transaction loops (Motie & Raahemi, 2024). The experiments are conducted within a cloud-native environment, leveraging distributed computing resources. The hardware configuration may include multi-core CPUs, GPU acceleration for GNN training and inference, and scalable storage systems. The cloud infrastructure is typically deployed using containerized services orchestrated via platforms such as Kubernetes, with streaming components integrated through event-driven systems. Such configurations are consistent with recent studies emphasizing the importance of distributed infrastructure for handling high-volume financial data and enabling real-time analytics (Daksa & Kemala, 2025).

5.2. Fraud Detection Performance

The effectiveness of the fraud detection model is evaluated using standard classification metrics widely adopted in fraud detection research. These include precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Precision measures the proportion of correctly identified fraudulent transactions among all flagged transactions, while recall assesses the model's ability to detect actual fraud cases. The F1-score provides a harmonic balance between precision and recall, which is particularly important in fraud detection due to the highly imbalanced nature of datasets. The AUC-ROC metric evaluates the model's ability to distinguish between fraudulent and legitimate transactions across different classification thresholds. Prior studies highlight that relying on a

single metric can be misleading in imbalanced datasets; therefore, a combination of these metrics provides a more comprehensive assessment of model performance (Dal Pozzolo et al., 2022; Yang et al., 2026). The evaluation also considers false positive rates, as excessive false alarms can reduce operational efficiency and increase investigation costs.

5.3. Real-Time Processing Performance

Beyond predictive accuracy, the framework is evaluated based on its ability to operate in real-time environments. Key performance indicators include inference latency, throughput, and graph update performance. Inference latency measures the time required to process a transaction and produce a fraud prediction, while throughput captures the number of transactions processed per second. These metrics are critical in financial systems where decisions must be made within milliseconds to prevent fraudulent transactions before settlement. Graph update performance evaluates the efficiency of dynamically updating the transaction graph as new data arrives. This includes the time required to add nodes and edges, recompute features, and maintain graph consistency. Studies on streaming fraud detection emphasize that efficient graph updates and low-latency inference are essential for maintaining system responsiveness under continuous data streams (Zhang et al., 2024). The results are typically benchmarked against baseline models or traditional batch-processing systems to demonstrate improvements in real-time capability.

5.4. Scalability Evaluation

The scalability of the proposed framework is assessed by analyzing system performance under increasing transaction loads. This involves conducting stress tests in which the volume and velocity of incoming transactions are progressively increased. Key metrics include system throughput, response time, resource utilization, and fault tolerance. Cloud-native architectures are expected to scale horizontally by adding computing resources as demand increases. Therefore, the evaluation examines how effectively the system maintains performance as it scales across distributed nodes. Prior research indicates that scalability is a critical requirement for fraud detection systems deployed in large financial institutions, where millions of transactions may be processed daily (Khan et al., 2023; Chen et al., 2025). The results provide insights into the framework's ability to sustain performance in high-demand environments.

5.5. Explainability Evaluation

The explainability component of the framework is evaluated by assessing the usefulness, clarity, and relevance of generative AI-generated explanations for fraud analysts. Unlike traditional performance metrics, explainability evaluation often involves a combination of qualitative and quantitative methods. Qualitative evaluation may include expert assessment, where fraud analysts review generated explanations and rate them based on criteria such as interpretability, completeness, and actionability. Quantitative measures may include explanation consistency, fidelity (alignment with model predictions), and reduction in investigation time. Recent studies suggest that explainable AI significantly enhances user trust, improves decision-making efficiency, and supports regulatory compliance in financial systems (Park, 2024). Additionally, the evaluation considers how well the explanation module communicates key fraud indicators, such as abnormal transaction patterns, suspicious relationships, and deviations from expected behavior. By integrating explainability into the evaluation process, the study ensures that the proposed framework not only achieves high detection accuracy but also provides meaningful insights that can support real-world fraud investigation and compliance requirements.

6. Discussion

The experimental results provide strong evidence of the effectiveness of graph neural networks (GNNs) in capturing complex relational fraud patterns within financial systems. Unlike traditional machine learning models that treat transactions independently, the GNN-based approach leverages structural dependencies among entities, enabling the detection of coordinated fraud schemes such as account collusion, transaction loops, and money laundering networks. The observed improvement in detection performance aligns with recent studies, which show that graph-based models outperform conventional approaches by incorporating neighborhood information and network topology into the learning process (Motie & Raahemi, 2024; Cheng et al., 2025). In addition to improved detection capability, the integration of generative AI for anomaly explanation significantly enhances interpretability. The experimental findings indicate that the explanation module produces coherent, context-aware narratives that help analysts understand why transactions are flagged. This addresses a key limitation of many advanced AI models, which often function as "black boxes." By translating complex model outputs into human-readable insights, the framework supports better

decision-making, faster investigations, and improved regulatory compliance. These findings are consistent with recent research highlighting the role of generative AI and large language models in improving transparency and usability in financial analytics (Desai et al., 2024; Park, 2024; Yang et al., 2026).

However, the results also reveal important trade-offs between detection accuracy and system latency. While the use of GNNs and real-time graph updates improves fraud detection performance, it introduces computational overhead that can affect inference speed. Maintaining low latency is critical in financial systems, particularly for transaction authorization processes. The experiments demonstrate that achieving an optimal balance requires careful system design, including efficient graph sampling, feature caching, and scalable model deployment. Similar trade-offs have been reported in streaming-based fraud detection systems, where increased model complexity can impact real-time performance if not properly managed (Daksa & Kemala, 2025; Chen et al., 2025). The evaluation further highlights the benefits of cloud-native architecture in supporting scalable and resilient fraud detection systems. The use of distributed microservices, container orchestration, and event-driven pipelines enables the system to handle high transaction volumes while maintaining availability and performance. The architecture's ability to scale horizontally ensures that performance is sustained under increasing workload conditions, which is essential for modern financial institutions. These findings reinforce the growing consensus that cloud-native design is fundamental to deploying advanced AI models in production environments (Khan et al., 2023; Chen et al., 2025).

7. Limitations and Future Research

Despite its contributions, the proposed framework is subject to several limitations. One major constraint is data privacy and security, as financial transaction data is highly sensitive and subject to strict regulatory requirements. Access to real-world datasets is often restricted, which may limit the generalizability of experimental results. Ensuring compliance with data protection regulations while enabling effective model training remains a critical challenge in financial fraud detection (Yang et al., 2026).

Another limitation relates to model bias and fairness. Fraud detection models trained on historical data may inadvertently learn biased patterns, potentially leading to unfair treatment of certain user groups or regions. Addressing bias requires careful dataset curation, fairness-aware modeling techniques, and continuous monitoring to ensure ethical AI deployment. Additionally, the computational complexity of large-scale transaction graphs presents practical challenges. As the number of nodes and edges increases, graph processing and GNN inference can become resource-intensive, potentially affecting scalability and latency if not properly optimized (Motie & Raahemi, 2024; Zhang et al., 2024).

Future research can address these limitations by exploring privacy-preserving graph learning techniques, such as differential privacy and secure multi-party computation, which enable model training without exposing sensitive data. Another promising direction is the development of federated fraud detection systems, where models are trained across multiple institutions without sharing raw data, thereby enhancing both privacy and collaboration. Furthermore, integrating confidential computing technologies, such as trusted execution environments, can provide secure processing of sensitive financial data in cloud environments. Additional research may also focus on improving model efficiency through lightweight GNN architectures and advanced sampling techniques to reduce computational overhead while maintaining performance.

8. Conclusion

This study proposes a hybrid graph-neural and generative AI framework for real-time fraud detection in cloud-native financial systems. By combining graph-based relational modeling with generative AI-driven explanation, the framework addresses key limitations of traditional fraud detection approaches, including the inability to capture complex transaction relationships and the lack of interpretability in advanced AI models.

The results demonstrate that graph neural networks significantly enhance fraud detection performance by leveraging structural and temporal dependencies within financial data, while generative AI improves transparency by providing meaningful, human-readable explanations for flagged transactions. The integration of these components within a cloud-native, real-time architecture further ensures scalability, resilience, and operational efficiency in high-volume financial environments.

Finally, the study highlights the importance of integrating detection accuracy, interpretability, and scalable deployment in modern fraud detection systems. The proposed framework contributes to the evolving field of intelligent financial security by offering a comprehensive approach that balances

performance, transparency, and practicality. As financial systems continue to digitize and fraud techniques become more sophisticated, such hybrid and adaptive approaches will play a critical role in strengthening fraud prevention and maintaining trust in digital financial ecosystems.

References

[1] Association for Financial Professionals (AFP). (2025). *2025 AFP Payments Fraud and Control Survey Report*.

[2] Central Bank of Nigeria (CBN). (2022). *Payments System Vision 2025 (PSV 2025)*.

[3] Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). Deep learning in financial fraud detection: Innovations, challenges, and applications. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2025.08.002

[4] Cheng, D., Zou, Y., Xiang, S., & Jiang, C. (2025). Graph neural networks for financial fraud detection: A review. *Frontiers of Computer Science, 19*, 199609. https://doi.org/10.1007/s11704-024-40474-y

[5] Daksa, R. P., & Kemala, A. P. (2025). A comparative study on real-time data streaming for fraud detection using Kafka with Apache Flink and Apache Spark. *Procedia Computer Science, 269*, 192–199. https://doi.org/10.1016/j.procs.2025.08.272

[6] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2022). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*.

[7] Desai, A. P., Ravi, T., Luqman, M., Mallya, G. S., Kota, N., & Yadav, P. (2024). Opportunities and challenges of generative AI in finance. *IEEE Big Data Conference*.

[8] European Payments Council. (2025). *Payment Threats and Fraud Trends Report*.

[9] Khan, S., Liu, X., & Shakil, K. A. (2023). Real-time fraud detection using streaming analytics: Challenges and opportunities. *Future Generation Computer Systems*.

[10] Lou, C., Wang, Y., Li, J., Qian, Y., & Li, X. (2025). Graph neural network for fraud detection via context encoding and adaptive aggregation.

[11] Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications, 240*, 122156.

[12] Park, T. (2024). Enhancing anomaly detection in financial markets with an LLM-based multi-agent framework. *arXiv preprint*.

[13] Tang, T., Yao, J., Wang, Y., Sha, Q., Feng, H., & Xu, Z. (2025). Application of deep generative models for anomaly detection in complex financial transactions. *arXiv preprint*.

[14] Yang, H., Shukur, Z., & Sahran, S. (2026). A review of artificial intelligence for financial fraud detection. *Applied Sciences, 16*(4), 1931. https://doi.org/10.3390/app16041931

[15] Zhang, Y., Li, X., & Chen, J. (2024). Temporal graph neural networks for dynamic fraud detection. *IEEE Transactions on Knowledge and Data Engineering*.

[16] Zhou, Q., Wang, H., & Li, P. (2023). Cloud-native architectures for scalable AI systems. *IEEE Cloud Computing*.