The American Journal of Engineering and Technology
ISSN 2689-0984 Volume 08 - 2026

Graph Neural Network (GNN) and XAI for Adaptive, Cross-Institutional Fraud Detection

Ravi Teja Bonam
Independent Researcher, usa

Received: 21 Dec 2025 | Received Revised Version: 19 Jan 2026 | Accepted: 17 Feb 2026 | Published: 28 Feb 2026

Volume 08 Issue 02 2026 | DOI 10.37547/tajet/Volume08Issue02-21

Abstract

The rapid digitalization of financial services has significantly increased the scale and complexity of fraudulent activities across financial institutions. Conventional fraud detection systems, primarily based on rule-based mechanisms and tabular machine learning models, often fail to identify sophisticated and interconnected fraud schemes spanning multiple institutions. Furthermore, increasing regulatory requirements regarding transparency and data privacy restrict direct sharing of financial transaction data among organizations. This study proposes a privacy-preserving and adaptive fraud detection framework that integrates Graph Neural Networks (GNNs), Federated Learning (FL), and Explainable Artificial Intelligence (XAI) to enable collaborative fraud detection across institutions without compromising sensitive customer information.

The proposed framework models financial ecosystems as distributed heterogeneous graphs consisting of accounts, merchants, devices, and transactional relationships. A federated graph learning mechanism enables participating institutions to collaboratively train fraud detection models while preserving local data ownership. To address the interpretability limitations of deep learning systems in highly regulated financial environments, XAI techniques are integrated to provide transparent explanations for fraud predictions at transaction, account, and network levels. Additionally, adaptive learning strategies are incorporated to address evolving fraud patterns and concept drift over time.

The study presents the conceptual architecture, methodological framework, privacy-preserving mechanisms, explainability integration, and evaluation strategy for the proposed system. The framework aims to improve fraud detection accuracy, enhance regulatory compliance, and facilitate collaborative intelligence sharing among financial institutions while maintaining strict privacy guarantees.

Keywords: Graph Neural Networks (GNN), Explainable Artificial Intelligence (XAI), Federated Learning, Financial Fraud Detection, Cross-Institutional Learning, Privacy-Preserving Machine Learning.

Cite This Article: Bonam, R. T. (2026). Graph Neural Network (GNN) and XAI for Adaptive, Cross-Institutional Fraud Detection. The American Journal of Engineering and Technology, 8(2), 245–259. https://doi.org/10.37547/tajet/Volume08Issue02-21

I. INTRODUCTION

The global financial ecosystem has undergone rapid transformation due to the increasing adoption of digital banking, online transactions, mobile payment systems, and open banking infrastructures. Although these developments have improved accessibility and operational efficiency, they have simultaneously

expanded the attack surface for financial fraud. Fraudulent activities such as payment fraud, account takeover, money laundering, synthetic identity fraud, and coordinated financial crimes have become increasingly sophisticated and interconnected across institutions.

Traditional fraud detection systems predominantly rely on rule-based engines and machine learning approaches using tabular transaction records. While these methods are effective in detecting previously observed fraud patterns, they frequently struggle to identify complex relational fraud networks involving multiple entities, devices, and institutions. Modern fraud schemes often involve hidden associations among customer accounts, merchants, devices, IP addresses, and intermediaries, making fraud inherently relational in nature. As a result, conventional record-based detection techniques are often insufficient for identifying collusive fraud behavior and emerging attack patterns.

Graph Neural Networks (GNNs) have emerged as a promising class of deep learning techniques capable of learning from relational and graph-structured data. Unlike traditional machine learning models, GNNs can capture dependencies among interconnected entities through message-passing mechanisms, enabling the detection of hidden fraud rings, suspicious transaction pathways, and multi-hop dependencies across financial networks. Recent studies have demonstrated that graph-based approaches significantly outperform conventional fraud detection models in identifying large-scale and coordinated fraudulent behavior.

Despite the effectiveness of GNN-based approaches, financial institutions face significant challenges in deploying collaborative fraud detection systems due to privacy regulations and competitive concerns. Regulatory frameworks such as the General Data Protection Regulation (GDPR), banking secrecy laws, and regional data residency policies restrict institutions from directly sharing sensitive customer information. Consequently, the inability to exchange raw transactional data limits the effectiveness of cross-institutional fraud detection systems.

Federated Learning (FL) provides a viable solution to this challenge by enabling multiple institutions to collaboratively train machine learning models without exposing proprietary or customer-sensitive data. In federated settings, institutions retain local ownership of data while sharing encrypted model updates with a coordinating server for aggregation. Extending federated learning to graph-structured financial data enables collaborative fraud intelligence while preserving confidentiality and regulatory compliance.

Another critical challenge in deploying deep learning systems in financial environments is the lack of interpretability. Financial fraud detection decisions often require justification for auditors, regulators, investigators, and compliance teams. Black-box predictions generated by deep learning models may undermine trust and complicate regulatory approval. Explainable Artificial Intelligence (XAI) techniques address this limitation by generating human-understandable explanations for model decisions. In graph-based systems, XAI methods can identify influential nodes, edges, subgraphs, and features contributing to suspicious activity predictions.

Motivated by these challenges, this study proposes a privacy-preserving framework integrating Graph Neural Networks, Federated Learning, and Explainable Artificial Intelligence for adaptive cross-institutional financial fraud detection. The proposed framework enables collaborative fraud intelligence sharing while maintaining privacy guarantees and providing transparent explanations to stakeholders. Furthermore, the framework incorporates adaptive learning mechanisms to continuously respond to evolving fraud strategies and concept drift in financial environments.

The major contributions of this study are summarized as follows:

1. A distributed graph-based representation of financial transaction ecosystems for modeling cross-institutional fraud relationships.
2. A federated GNN framework that enables collaborative fraud detection without exposing raw institutional data.
3. An integrated explainability mechanism to generate interpretable transaction-level and network-level fraud explanations.
4. An adaptive learning strategy for detecting emerging fraud patterns and mitigating concept drift.
5. A comprehensive evaluation framework to assess detection performance, privacy

preservation, communication overhead, and explainability effectiveness.

II. RELATED WORK

A. Financial Fraud Detection: Current Challenges

Financial fraud detection has become increasingly challenging due to the growing complexity of digital transaction ecosystems. Traditional fraud detection approaches largely rely on rule-based systems and supervised machine learning algorithms operating on structured tabular datasets. Although these methods remain effective for identifying previously known fraud signatures, they often suffer from high false-positive rates and limited adaptability to emerging fraud behaviors.

Contemporary financial fraud schemes are highly interconnected and frequently involve coordinated interactions among multiple entities, including customer accounts, merchants, payment devices, IP addresses, and intermediaries. Such fraud activities often span multiple institutions and evolve dynamically over time, making static and institution-specific detection methods insufficient. Moreover, severe class imbalance, scarcity of labeled fraud data, and the requirement for near real-time detection further complicate effective fraud prevention.

Recent advances in deep learning have improved fraud detection capabilities by enabling the extraction of complex behavioral patterns from large-scale transactional datasets. However, most conventional deep learning techniques continue to rely on tabular representations and often fail to capture hidden relationships among entities participating in fraudulent activities.

B. Graph Neural Networks for Financial Fraud Detection

Graph Neural Networks (GNNs) have emerged as a powerful approach for learning from graph-structured data and have demonstrated considerable promise in financial fraud detection. Unlike traditional machine learning models, GNNs explicitly capture relational dependencies among interconnected entities through message-passing and neighborhood aggregation mechanisms.

In financial systems, transactional ecosystems naturally form graph structures in which nodes represent customer accounts, merchants, devices, cards, terminals, or IP addresses, while edges represent transactional or behavioral relationships. By modeling such ecosystems as graphs, GNNs can uncover latent patterns of suspicious interactions that may remain undetected in conventional record-based systems.

Several GNN architectures have shown effectiveness in fraud detection applications:

1) Graph Convolutional Networks (GCN)

Graph Convolutional Networks aggregate neighborhood information to generate node embeddings that capture structural relationships among entities. GCN-based approaches have demonstrated strong performance in fraud classification tasks involving dense financial interaction networks.

2) GraphSAGE

GraphSAGE introduces inductive representation learning by sampling and aggregating neighboring node information. This architecture is particularly suitable for large-scale financial systems because it supports dynamic graphs and generalization to previously unseen nodes.

3) Graph Attention Networks (GAT)

Graph Attention Networks employ attention mechanisms to assign varying importance to neighboring entities. In fraud detection scenarios, GATs improve interpretability by highlighting influential relationships contributing to suspicious behavior predictions.

Empirical studies indicate that graph-based fraud detection systems significantly outperform conventional machine learning approaches by improving detection sensitivity, reducing false positives, and identifying collusive fraud structures such as fraud rings and mule account networks. Nevertheless, most existing GNN implementations are limited to single-institution environments, restricting their ability to identify coordinated fraudulent behavior across financial ecosystems.

C. Federated Learning for Privacy-Preserving Fraud Detection

Data privacy regulations and competitive concerns present major obstacles to collaborative fraud detection among financial institutions. Regulations such as the General Data Protection Regulation (GDPR), banking secrecy requirements, and regional data governance laws prohibit unrestricted sharing of customer-sensitive transactional information.

Federated Learning (FL) has emerged as an effective paradigm for enabling collaborative machine learning without exposing raw data. In FL settings, institutions train local models independently and share model parameters or gradients rather than sensitive datasets. A central aggregation server combines institutional updates to generate an improved global model.

The application of federated learning to financial fraud detection provides several advantages:

- Preservation of institutional data confidentiality.
- Compliance with privacy and regulatory frameworks.
- Enhanced detection capabilities through collective intelligence.
- Reduction of data-sharing risks among competing organizations.

Recent studies have demonstrated the feasibility of applying federated learning in banking systems for fraud detection, particularly when reinforced with privacy-enhancing technologies such as secure aggregation and differential privacy.

However, extending federated learning to graph-structured financial systems introduces unique challenges. Unlike conventional tabular data, graph-based learning requires preserving structural dependencies among distributed entities. Recent advances in Federated Graph Learning (FGL) have begun addressing these challenges by enabling collaborative graph representation learning without centralized graph construction.

Despite these developments, concerns remain regarding model leakage attacks, including gradient inversion and membership inference attacks. To mitigate such vulnerabilities, recent studies recommend integrating secure multiparty computation, homomorphic encryption, and differential privacy mechanisms into federated graph learning environments.

D. Explainable Artificial Intelligence for Graph Neural Networks

Although deep learning methods have improved predictive performance, their black-box nature creates significant barriers to deployment in highly regulated domains such as banking and financial risk management. Fraud detection decisions frequently require transparent justification for regulators, auditors, compliance officers, and fraud analysts.

Explainable Artificial Intelligence (XAI) aims to improve transparency by providing interpretable explanations of model decisions. In graph neural networks, explainability extends beyond feature importance and includes the identification of influential nodes, edges, neighborhoods, and subgraphs contributing to predictions.

Several graph explainability methods have gained prominence:

1) GNNExplainer

GNNExplainer identifies critical graph structures and feature subsets responsible for predictions. It provides instance-level explanations by highlighting influential neighborhood connections associated with suspicious behavior.

2) PGExplainer

PGExplainer generates probabilistic explanations for graph predictions and supports scalable explanation generation across large networks.

3) GraphLIME and GraphSHAP

These approaches extend explainability by quantifying feature importance and local graph behavior through interpretable approximations.

In financial fraud detection, explainable AI provides several operational advantages:

- Transparent reasoning behind fraud alerts.
- Improved trust among analysts and compliance teams.

- Regulatory accountability and auditability.
- Enhanced model debugging and drift monitoring.

However, existing literature suggests that no single explainability method universally outperforms others. The effectiveness of explanations often depends on stakeholder requirements, operational context, and graph complexity.

E. Cross-Institutional and Adaptive Fraud Detection

Fraudulent activities increasingly exploit fragmentation among financial institutions, allowing fraudsters to distribute malicious activities across multiple banks, payment systems, and digital platforms. Consequently, isolated fraud detection systems frequently fail to identify broader criminal networks.

Cross-institutional fraud detection seeks to overcome this limitation through collaborative intelligence sharing while preserving institutional privacy. Recent developments in federated graph learning have created opportunities for identifying distributed fraud patterns without requiring centralized data repositories.

Another major challenge in fraud detection is concept drift. Fraud patterns continuously evolve as attackers modify their behavior to evade detection mechanisms. Static detection systems often degrade in performance over time due to shifting behavioral distributions.

Adaptive fraud detection strategies proposed in literature include:

- Temporal GNNs for evolving transaction sequences.
- Continuous retraining mechanisms using streaming data.
- Drift detection systems for triggering model updates.
- Self-supervised and contrastive learning approaches to improve representation quality under limited labels.

Despite significant progress, existing studies largely focus on isolated aspects such as graph learning, privacy preservation, or explainability. Few frameworks comprehensively integrate these capabilities into a unified system suitable for cross-institutional fraud detection.

F. Research Gap

Although substantial advances have been made in Graph Neural Networks, Federated Learning, and Explainable Artificial Intelligence for fraud detection, a comprehensive framework that simultaneously addresses collaborative learning, privacy preservation, interpretability, and adaptivity remains largely underexplored.

Specifically, current research lacks an integrated architecture capable of:

1. Modeling distributed financial ecosystems as interconnected graph structures.
2. Enabling collaborative GNN training across institutions without exposing sensitive transactional information.
3. Providing transparent and regulator-friendly explanations for fraud predictions.
4. Continuously adapting to evolving fraud strategies and concept drift.

To address these limitations, this study proposes a unified privacy-preserving federated GNN framework integrated with XAI mechanisms for adaptive cross-institutional financial fraud detection.

III. PROPOSED FRAMEWORK AND METHODOLOGY

A. System Architecture Overview

This study proposes a privacy-preserving and adaptive fraud detection framework that integrates Graph Neural Networks (GNNs), Federated Learning (FL), and Explainable Artificial Intelligence (XAI) to facilitate collaborative fraud detection across multiple financial institutions. The proposed architecture is designed to address three major challenges in modern financial fraud detection: fragmented institutional intelligence, strict privacy requirements, and the lack of transparency in deep learning models.

The framework consists of five primary architectural layers:

1) Local Institutional Layer

Each participating financial institution, such as banks, payment service providers (PSPs), or digital financial platforms, maintains its own local transaction environment and customer records. Instead of sharing raw transactional information, institutions independently construct graph-based representations of their financial ecosystems and locally train fraud detection models.

This decentralized architecture ensures that institutions preserve data ownership and remain compliant with privacy regulations while contributing to collaborative fraud intelligence.

2) Federated Learning Coordination Layer

A federation orchestrator coordinates model updates among participating institutions. Rather than centralizing sensitive financial information, the orchestrator aggregates encrypted model updates received from institutional participants and generates an improved global model.

The federated coordination layer is responsible for:

- Synchronizing federated training rounds.
- Aggregating institutional model parameters.
- Maintaining global model consistency.
- Managing communication among participants.
- Supporting secure model update mechanisms.

The orchestrator does not access raw institutional data, thereby minimizing privacy risks and regulatory violations.

3) Graph Neural Network Layer

At the core of the framework lies a Graph Neural Network-based fraud detection model. The GNN processes graph-structured financial relationships and captures hidden interactions among accounts, merchants, devices, IP addresses, and transactional pathways.

Unlike traditional tabular machine learning approaches, GNNs enable learning over interconnected structures and support the identification of coordinated fraud patterns that may otherwise remain undetected.

The GNN backbone is shared across institutions while model training occurs locally under the federated learning paradigm.

4) Explainability Layer

To ensure transparency and regulatory compliance, an Explainable Artificial Intelligence (XAI) layer is integrated into the fraud detection pipeline.

This layer generates interpretable explanations for fraud predictions by identifying:

- Influential neighboring entities.
- Suspicious transaction pathways.
- Significant graph substructures.
- Critical behavioral features contributing to model decisions.

The explainability layer supports fraud investigators, compliance officers, and auditors by improving decision interpretability and operational trust.

5) Privacy-Preserving Security Layer

A dedicated privacy-preserving layer strengthens institutional confidentiality through advanced security mechanisms.

The framework integrates:

- **Secure Aggregation** to protect institutional model updates.
- **Differential Privacy (DP)** to prevent information leakage.
- **Homomorphic Encryption (HE)** for secure encrypted computation.
- **Secure Multi-Party Computation (SMPC)** for collaborative model optimization.

These techniques collectively reduce vulnerability to model inversion, gradient leakage, and membership inference attacks.

B. Financial Graph Construction and Data Representation

Financial ecosystems exhibit inherently relational characteristics involving multiple interacting entities. Consequently, this study models transactional environments as heterogeneous graphs.

In the proposed graph representation:

1) Nodes

Nodes represent important financial entities, including:

- Customer accounts
- Merchants
- Payment cards
- Mobile devices
- IP addresses
- Automated Teller Machines (ATMs)
- Transaction terminals

Each node contains descriptive features reflecting behavioral and contextual information.

For example:

Customer account features:

- Transaction frequency
- Historical fraud risk
- Customer activity patterns
- KYC (Know Your Customer) summaries

Merchant features:

- Merchant category
- Transaction distribution
- Fraud exposure history

2) Edges

Edges represent relationships among financial entities.

Examples include:

- Account-to-merchant transactions
- Shared device relationships
- Shared IP address usage
- Account-to-account transfers
- Behavioral similarity links

Each edge contains temporal and transactional attributes such as:

- Transaction amount
- Currency
- Transaction timestamp
- Payment channel
- Device information
- Risk indicators

Time-stamped graph structures allow the framework to model evolving fraud behavior over time.

By learning from graph topology, the proposed framework captures hidden relational patterns including fraud rings, mule account networks, suspicious merchant clusters, and coordinated attack pathways.

C. Graph Neural Network Model Design

The proposed fraud detection system employs a Graph Neural Network architecture designed specifically for relational financial learning.

The model architecture consists of four key components:

1) Input Embedding Layer

The input layer transforms node and edge features into dense numerical representations suitable for deep learning.

Input features include:

- Behavioral attributes
- Temporal transaction information
- Device metadata

- Historical fraud indicators
- Institutional context

Embedding mechanisms improve feature representation and facilitate graph-level learning.

2) Message Passing and Neighborhood Aggregation

The core learning process occurs through message-passing operations where information propagates between neighboring entities.

This study considers two major GNN architectures:

GraphSAGE

GraphSAGE is selected due to its scalability and ability to generalize to previously unseen nodes. It efficiently handles large financial transaction graphs where new customers and accounts continuously emerge.

Graph Attention Networks (GAT)

Graph Attention Networks introduce attention mechanisms that assign importance weights to neighboring entities. This enables the model to focus on highly influential relationships, improving fraud explainability and predictive accuracy.

The message-passing mechanism enables detection of:

- Hidden fraud rings
- Multi-hop suspicious relationships
- Shared fraudulent infrastructure
- Coordinated money laundering pathways

3) Readout and Classification Layer

The readout layer transforms graph representations into fraud prediction outputs.

The framework supports multiple prediction objectives:

Node-level classification

- Account fraud risk prediction

Edge-level classification

- Fraudulent transaction detection

Subgraph-level scoring

- Detection of suspicious relational structures

The final prediction layer generates fraud probability scores using sigmoid or softmax activation functions.

Threshold-based ranking mechanisms classify suspicious activities requiring investigation.

4) Imbalanced Fraud Learning

Financial fraud datasets typically suffer from severe class imbalance because fraudulent transactions represent only a small fraction of total activity.

To mitigate this issue, the framework incorporates:

- Weighted binary cross-entropy loss
- Focal loss optimization
- Minority oversampling strategies
- Cost-sensitive learning mechanisms

These techniques improve fraud sensitivity and reduce false negatives.

D. Federated Learning Protocol

The proposed framework adopts a federated learning strategy to enable collaborative model development without exposing raw institutional data.

The training procedure consists of the following stages:

Step 1: Global Model Initialization

A central federation server initializes the global GNN model and distributes it to participating institutions.

Step 2: Local Institutional Training

Each institution independently trains the model using locally available transaction data and graph structures.

Training occurs exclusively within institutional boundaries to preserve privacy.

Step 3: Model Update Sharing

Instead of transferring financial data, institutions share encrypted model updates, including gradients or trained parameters.

Step 4: Secure Aggregation

The federation orchestrator aggregates institutional updates using weighted averaging methods such as Federated Averaging (FedAvg).

The updated parameters are used to construct a stronger global model.

Step 5: Global Redistribution

The improved global model is redistributed to institutions for subsequent training rounds.

This iterative process enables collaborative fraud intelligence while preserving institutional confidentiality.

E. Privacy-Preserving Security Mechanisms

Protecting customer confidentiality is a fundamental design objective of the proposed system.

Several privacy-preserving mechanisms are incorporated:

1) Secure Aggregation

Secure aggregation prevents the federation server from observing individual institutional updates.

Only aggregated model information becomes accessible during training.

2) Differential Privacy

Differential privacy introduces carefully controlled statistical noise to gradients and model parameters.

This limits the probability of reconstructing sensitive customer information.

3) Homomorphic Encryption

Homomorphic encryption enables encrypted mathematical computation without exposing plaintext model information.

This mechanism becomes particularly valuable in highly regulated banking environments.

4) Secure Multi-Party Computation

Secure multi-party computation enables institutions to collaboratively optimize models without revealing proprietary internal information.

These mechanisms collectively ensure compliance with global privacy standards and financial regulations.

F. Explainability Integration

Interpretability is essential in financial fraud detection because investigators and regulators require transparent explanations for automated decisions.

The proposed framework integrates XAI as a first-class component.

Local-Level Explanations

For each suspicious transaction or flagged account, explanation mechanisms identify:

- Critical neighboring entities
- Suspicious transaction pathways
- Important behavioral attributes
- Fraud-related subgraph structures

Methods such as **GNNExplainer**, **PGExplainer**, and **GraphSHAP** are incorporated to generate localized explanations.

Global-Level Explanations

Aggregated explanation statistics across institutions provide:

- Common fraud motifs
- Frequently recurring suspicious patterns
- Shared attack signatures

Importantly, explanation sharing remains anonymized to preserve institutional confidentiality.

Analyst Support Interface

Fraud analysts receive explanations through intuitive dashboards including:

- Graph visualizations

- Risk contribution scores
- Textual justification summaries
- Transaction pathway explanations

These interfaces improve trust, operational efficiency, and auditability.

G. Adaptivity and Concept Drift Handling

Fraud behavior continuously evolves, making adaptability essential.

The proposed framework incorporates several adaptive learning mechanisms.

Online Incremental Learning

Periodic federated updates enable continuous model improvement using recent transaction streams.

Drift Detection

Performance degradation and distributional changes are monitored to identify concept drift.

Detected shifts trigger:

- Model retraining
- Threshold recalibration
- Feature adaptation

Self-Supervised Learning

Self-supervised graph tasks such as:

- Link prediction
- Contrastive graph learning
- Neighbor reconstruction

enhance representation quality when labeled fraud data are limited.

Few-Shot Adaptation

Few-shot learning mechanisms improve detection of newly emerging fraud strategies using limited training samples.

These adaptive capabilities improve long-term robustness and maintain fraud detection effectiveness in evolving financial ecosystems.

IV. EVALUATION STRATEGY

A. Experimental Design

The proposed framework will be evaluated using a combination of simulated and real-world inspired financial transaction datasets to assess its effectiveness in cross-institutional fraud detection. Since direct access to proprietary banking datasets is often restricted due to confidentiality and regulatory concerns, this study adopts a realistic experimental setup that reflects operational financial environments.

The evaluation process will involve multi-institutional transaction settings in which participating institutions maintain independent datasets while collaboratively training the fraud detection model through federated learning mechanisms.

Two experimental scenarios are considered:

1) Simulated Cross-Institutional Financial Environment

Synthetic financial transaction datasets will be generated to emulate realistic fraud behavior across multiple institutions. These datasets will include:

- Customer account networks
- Merchant interactions
- Shared device relationships
- Suspicious behavioral patterns
- Fraud rings and mule account structures

The simulated environment enables controlled experimentation and allows benchmarking under varying fraud complexities.

2) Distributed Institutional Data Scenario

A realistic federated setting will be emulated by partitioning financial datasets into institution-specific subsets. Each institution independently retains local customer records while participating in collaborative model training.

This approach enables evaluation of:

- Cross-institutional intelligence sharing
- Privacy preservation effectiveness
- Communication overhead
- Generalization capability

To mimic real-world deployment environments, the dataset will be divided temporally into:

- **Training dataset**
- **Validation dataset**
- **Testing dataset**

Time-aware evaluation ensures realistic assessment of fraud evolution and deployment robustness.

B. Performance Metrics

The proposed system will be evaluated using multiple performance indicators to comprehensively measure predictive effectiveness, adaptability, explainability, and privacy preservation.

1) Fraud Detection Performance

The following classification metrics will be used:

Accuracy

Measures overall prediction correctness:

[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
]

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Precision

Evaluates the proportion of correctly identified fraudulent transactions among predicted fraud cases.

Recall (Sensitivity)

Measures the model's ability to correctly identify actual fraudulent behavior.

F1-Score

The harmonic mean of precision and recall, particularly useful for imbalanced fraud datasets.

AUC-ROC (Area Under Receiver Operating Characteristic Curve)

Assesses the discrimination ability of the model between fraudulent and non-fraudulent behavior.

AUC-PR (Area Under Precision-Recall Curve)

Provides a more informative evaluation for highly imbalanced fraud datasets.

2) Federated Learning Metrics

To assess federated efficiency, the following indicators will be evaluated:

- Communication overhead per federated round
- Model convergence time
- Training latency
- Institutional scalability
- Resource utilization

These measures help determine the operational feasibility of deployment across large financial networks.

3) Privacy Preservation Metrics

Privacy robustness will be evaluated through simulated attack scenarios including:

- Membership inference attacks
- Gradient inversion attacks
- Model reconstruction attacks

Performance comparisons will be conducted:

- With differential privacy
- Without differential privacy

- With secure aggregation
- Without secure aggregation

This analysis will assess the trade-off between privacy guarantees and predictive performance.

4) Adaptivity Metrics

The framework's ability to adapt to evolving fraud behavior will be evaluated using:

- Detection accuracy under concept drift
- Recovery performance after retraining
- Detection of previously unseen fraud patterns
- Time-to-adaptation for emerging fraud strategies

C. Explainability Evaluation

The explainability component of the proposed framework will be evaluated through both quantitative and qualitative methods.

1) Quantitative Evaluation

The following explainability indicators will be measured:

Faithfulness

Evaluates whether identified explanations genuinely influence prediction outcomes.

Stability

Measures consistency of explanations across similar fraud scenarios.

Sparsity

Assesses the simplicity and interpretability of generated explanations.

2) Qualitative Expert Assessment

Fraud analysts and domain experts will evaluate explanation usefulness based on:

- Interpretability
- Actionability
- Trustworthiness
- Regulatory suitability

Feedback will be collected through structured evaluation frameworks and expert scoring.

3) Case Study Analysis

Representative fraud cases will be analyzed to assess whether the framework successfully identifies:

- Fraud rings
- Money laundering pathways
- Suspicious account clusters
- Coordinated fraudulent behavior

Explanation outputs will be examined to determine whether generated reasoning supports operational fraud investigation.

V. DISCUSSION

The proposed framework addresses several limitations associated with traditional financial fraud detection systems. By integrating Graph Neural Networks with Federated Learning and Explainable Artificial Intelligence, the system enables collaborative fraud detection while maintaining privacy and transparency.

One of the most significant contributions of the framework lies in its ability to model financial systems as interconnected graph structures. Traditional machine learning approaches operating on tabular records often fail to identify hidden dependencies among entities participating in fraudulent behavior. The graph-based approach enables richer relational learning and facilitates detection of coordinated fraud networks.

Furthermore, the federated learning mechanism addresses privacy concerns that frequently prevent institutions from sharing customer-sensitive information. By exchanging model updates rather than raw data, institutions can collaboratively strengthen fraud intelligence while remaining compliant with financial regulations and privacy laws.

The integration of explainable AI further improves practical applicability. In regulated financial environments, fraud alerts must be interpretable and auditable. The proposed explainability layer enables

fraud analysts to understand why suspicious activities are flagged, thereby improving operational trust and reducing resistance to AI adoption.

Despite these advantages, the framework also presents several challenges. Federated graph learning introduces communication overhead and computational complexity, particularly in large-scale financial ecosystems. Moreover, implementing advanced privacy-preserving mechanisms such as homomorphic encryption may increase computational requirements.

Another challenge concerns data heterogeneity across institutions. Differences in institutional policies, customer demographics, fraud patterns, and data quality may affect model consistency and performance.

Nevertheless, the proposed framework provides a promising foundation for collaborative, transparent, and privacy-preserving fraud detection systems in increasingly interconnected financial environments.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

This study proposed a privacy-preserving and adaptive framework for cross-institutional financial fraud detection through the integration of Graph Neural Networks, Federated Learning, and Explainable Artificial Intelligence.

The framework addresses critical challenges in contemporary fraud detection, including fragmented institutional intelligence, privacy constraints, lack of interpretability, and continuously evolving fraud behavior.

By representing financial ecosystems as graph structures, the proposed framework captures hidden relationships among customers, merchants, devices, and transactional pathways. The integration of Graph Neural Networks enables the identification of sophisticated fraud rings and relational attack patterns that are difficult to detect through traditional methods.

Federated Learning facilitates collaborative intelligence sharing while preserving institutional confidentiality and regulatory compliance. Simultaneously, Explainable Artificial Intelligence enhances transparency by providing human-interpretable reasoning behind fraud predictions, thereby improving trust, accountability, and auditability.

Additionally, adaptive learning mechanisms improve robustness against concept drift and emerging fraud strategies, ensuring sustained performance over time.

Overall, the proposed framework presents a scalable and practical direction toward collaborative fraud detection in modern financial ecosystems.

B. Limitations

Although the proposed framework demonstrates strong conceptual potential, several limitations remain.

First, this study primarily presents a conceptual and methodological framework rather than empirical validation on proprietary banking datasets. Practical deployment performance may vary depending on data quality, labeling availability, and institutional participation.

Second, federated learning introduces communication costs and infrastructure complexity, particularly when scaling across multiple institutions.

Third, advanced privacy-preserving methods such as homomorphic encryption and secure multi-party computation may increase computational overhead and latency.

Finally, explanation quality in graph neural networks remains an active research area, and different explainability techniques may produce varying levels of interpretability.

C. Future Work

Future research may extend this work in several directions.

1) Temporal and Dynamic Graph Learning

Future models may integrate temporal GNN architectures to better capture continuously evolving transaction behavior and fraud dynamics.

2) Advanced Privacy Preservation

Hybrid privacy-preserving mechanisms combining differential privacy, secure aggregation, and cryptographic techniques may further improve institutional confidentiality.

3) Large Language Model Integration

Large Language Models (LLMs) may be incorporated to generate natural language fraud explanations, automate investigation summaries, and assist analysts in querying suspicious graph structures.

4) Real-Time Streaming Fraud Detection

Future systems may incorporate streaming architectures for low-latency fraud identification in high-frequency transaction environments.

5) Domain-Specific Explainability Interfaces

Customized visualization dashboards designed for regulators, auditors, and fraud analysts may improve practical usability and decision support.

In conclusion, the proposed framework establishes a strong foundation for transparent, adaptive, and privacy-preserving financial fraud detection in increasingly interconnected digital financial systems.

REFERENCES

1. W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 1024–1034.
2. T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.
3. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.
4. R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating Explanations for Graph Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, Vancouver, Canada, 2019.
5. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.
6. Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, Feb. 2019.
7. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
8. D. Zhou, J. Cui, H. Zhang, Z. Yang, C. Liu, and M. Sun, "Graph Neural Networks: A Review of Methods and Applications," *AI Open*, vol. 1, pp. 57–81, 2020.
9. C. Agarwal, H. Lakkaraju, and M. Zitnik, "Evaluating Explainability for Graph Neural Networks," *Scientific Data*, vol. 10, no. 1, pp. 1–21, 2023.
10. S. Kanamori, H. Watanabe, and T. Uchida, "Privacy-Preserving Federated Learning for Detecting Fraudulent Financial Transactions Across Multiple Banks," *IPSJ Transactions on Information and Systems*, vol. 30, no. 4, pp. 789–800, 2022.
11. D. Cheng, Y. Zhang, and H. Liu, "Graph Neural Networks for Financial Fraud Detection: A Comprehensive Review," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–35, 2024.
12. L. Ge, J. Han, and K. Xu, "A Review of Privacy-Preserving Research on Federated Graph Neural Networks," *Neurocomputing*, vol. 575, Art. no. 128084, 2024.
13. S. Motie, J. Zhang, and X. Sun, "Financial Fraud Detection Using Graph Neural Networks," *Expert Systems with Applications*, vol. 235, Art. no. 121118, 2024.
14. O. Owoade and K. Petersen, "Federated Graph Learning for Privacy-Preserving Financial Anomaly Detection," *Nature Communications*, vol. 15, Art. no. 3456, 2024.
15. Y. Chen, X. Li, and T. Wang, "Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Future Directions," *Discover Artificial Intelligence*, vol. 5, no. 1, Art. no. 12, 2025.
16. S. K. Aljunaid, M. Al-Hadrani, and Y. H. Al-Mamary, "Explainable AI-Driven Federated Learning Model for Financial Fraud Detection," *Journal of Risk and Financial Management*, vol. 18, no. 4, Art. no. 179, 2025.
17. M. Nandan, R. Gupta, and P. Singh, "A Survey of Graph Neural Networks for Explainable Artificial

Intelligence," *Neural Computing and Applications*, vol. 37, no. 9, pp. 14621–14655, 2025.

18. Z. Xia, J. Li, and R. Huang, "FinGraphFL: Financial Graph-Based Federated Learning with Differential Privacy for Credit Card Fraud Detection," *Mathematics*, vol. 13, no. 9, Art. no. 1396, 2025.

19. Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, "Privacy-Preserving Traffic Flow Prediction: A Federated Learning Approach," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, Aug. 2020.

20. J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph Neural Networks: A Review of Methods and Applications," *AI Open*, vol. 1, pp. 57–81, 2020.