International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 6, Issue 4, PP 319-335, 2025
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V6I4P138

Original Article

Federated Learning for Privacy-Preserving Fraud Detection across Financial Institutions

Sakthi Sankara Balaji Sathyamurthy
Truist Financial Corporation, USA.

Received On: 11/11/2025 ***Revised On: 13/12/2025*** ***Accepted On: 25/12/2025*** ***Published On: 30/12/2025***

Abstract: *Financial fraud has become increasingly sophisticated as digital banking, online payments, fintech platforms, and cross-border financial services continue to expand. Conventional fraud detection systems are often limited by institutional data silos, where each bank or financial organisation trains models only on its own transaction records. This restricts the ability to detect coordinated fraud patterns that move across multiple institutions. However, direct sharing of customer transaction data raises serious privacy, regulatory, commercial, and security concerns. Federated learning offers a promising solution by allowing financial institutions to collaboratively train fraud detection models without transferring raw data outside their local environments. This article examines the use of federated learning for privacy-preserving fraud detection across financial institutions. It outlines a framework in which participating institutions train local models on private transaction data and share protected model updates for secure aggregation into a global fraud detection model. The study further considers privacy-enhancing mechanisms such as secure aggregation, differential privacy, encrypted model updates, and model auditing. It also discusses key challenges, including non-identically distributed data, communication overhead, model poisoning, inference attacks, regulatory compliance, and explainability. The article argues that federated learning can improve collaborative fraud intelligence while maintaining stronger data protection, provided that technical safeguards, governance structures, and human oversight are properly implemented. The study contributes to the growing discussion on privacy-preserving artificial intelligence in financial crime detection and digital banking security.*

Keywords: *Federated Learning, Fraud Detection, Financial Institutions, Privacy-Preserving Machine Learning, Secure Aggregation, Differential Privacy, Financial Crime Analytics, Digital Banking Security, Collaborative Intelligence, Financial Technology.*

1. Introduction

Financial fraud remains one of the most persistent risks facing modern financial institutions. The expansion of digital banking, mobile payments, card-not-present transactions, fintech platforms, and real-time payment systems has created new opportunities for fraudsters to exploit weaknesses in transaction monitoring systems. Fraudulent activities are no longer limited to isolated accounts or single institutions. They often involve coordinated behaviours across banks, payment processors, merchants, devices, and digital identities. As a result, fraud detection has become a data-intensive task that requires timely access to diverse behavioural and transactional patterns.

Traditional fraud detection systems rely on rule-based controls, statistical models, anomaly detection, and machine learning classifiers. These systems can be useful in identifying suspicious activities within a single institution, but they are often limited by data silos. Each financial institution usually trains and maintains its own fraud detection system using internal data only. This creates a narrow view of fraud behaviour, especially when criminals move across institutions or use multiple accounts and platforms to avoid detection. Vanini et al. (2023) noted that online payment fraud requires approaches that move beyond isolated anomaly detection and connect fraud detection with broader risk management practices. Similarly, Cherif et al. (2023) observed that fraud detection systems must adapt to changing technologies and increasingly complex fraud strategies.

A major challenge is that direct data sharing between financial institutions is difficult. Transaction data contains sensitive information about customers, accounts, merchants, payment behaviour, location patterns, and financial history. Sharing such data may expose institutions to privacy breaches, regulatory violations, reputational damage, and competitive risks. Regulations on data protection and financial privacy also limit the extent to which institutions can pool customer-level datasets for collaborative model training. Although centralised fraud detection models may benefit from richer datasets, they require sensitive data to be transferred into a shared repository, which increases the risks of unauthorised access, misuse, and data leakage.

Federated learning offers a practical alternative to centralised data sharing. It allows multiple institutions to collaboratively train a shared machine learning model while keeping raw data within each local environment. In a federated setting, each financial institution trains a local model using its own private transaction data. Instead of

transferring customer records, the institution sends model updates to an aggregation server, where updates from participating institutions are combined to improve a global model. This structure supports collaborative intelligence while reducing the need to expose raw financial data. McMahan et al. (2017) introduced federated learning as a communication-efficient method for training models from decentralised data, while Kairouz et al. (2021) identified federated learning as a major direction for privacy-preserving machine learning across distributed systems.

The relevance of federated learning to fraud detection is supported by recent studies. Abdul Salam et al. (2024) applied federated learning to credit card fraud detection and showed its potential when combined with data balancing techniques. Reddy et al. (2024) also examined deep learning-based credit card fraud detection in a federated learning setting, highlighting the value of decentralised learning for sensitive financial data. More recent work has extended this discussion by combining federated learning with explainable artificial intelligence, graph-based modelling, blockchain, and privacy-enhancing mechanisms for financial fraud detection (Aljunaid et al., 2025; Awosika et al., 2024; Baabdullah et al., 2024; Xia et al., 2025).

Despite its promise, federated learning is not free from risks. Model updates may still leak information if they are not properly protected. Attackers may attempt membership inference, model inversion, gradient leakage, poisoning attacks, or malicious update manipulation. Zhao et al. (2024) noted that federated learning systems require careful attention to privacy attacks, defences, applications, and policy concerns. Privacy-preserving mechanisms such as secure aggregation, differential privacy, secure multiparty computation, and encrypted model updates are therefore important for financial use cases. Liu et al. (2022) also emphasised that privacy-preserving aggregation is central to protecting federated learning systems from exposure during model update exchange.

This article examines federated learning as a privacy-preserving approach to fraud detection across financial institutions. It focuses on how financial organisations can collaborate to improve fraud detection without directly sharing raw customer transaction data. The article discusses the limitations of centralised and institution-specific fraud detection, reviews the role of federated learning in financial fraud analytics, and considers technical safeguards needed to address privacy and security concerns. It also highlights key implementation issues, including data heterogeneity, model explainability, communication cost, governance, regulatory compliance, and human oversight.

The main contribution of this article is to provide a structured discussion of federated learning for cross-institution fraud detection. It brings together technical, privacy, security, and governance perspectives to show how financial institutions can benefit from collaborative model training while maintaining control over sensitive data. The article argues that federated learning can support more effective fraud intelligence when combined with appropriate privacy mechanisms, transparent model monitoring, and institutional accountability.

2. Literature Review

2.1. Financial Fraud Detection in Digital Banking

Fraud detection has become a central concern in digital finance due to the growing volume, speed, and complexity of financial transactions. Banks and payment providers process large numbers of transactions across digital channels, including mobile banking, online purchases, card payments, peer-to-peer transfers, and fintech platforms. These transactions generate behavioural and contextual data that can be used to identify suspicious patterns. However, fraud detection is difficult because fraudulent transactions often represent a very small proportion of total transactions, and fraud patterns change over time.

Early fraud detection systems relied heavily on fixed rules, manual investigation, and expert-defined thresholds. Such methods remain useful for known fraud patterns, but they often perform poorly when fraudsters change their tactics. Machine learning methods have therefore become important in financial fraud detection because they can learn complex relationships from transaction data. Common approaches include logistic regression, decision trees, random forests, support vector machines, gradient boosting, neural networks, and anomaly detection models. Seera et al. (2024) presented an intelligent payment card fraud detection system and demonstrated the continuing relevance of machine learning methods in payment fraud analytics.

Recent studies have also explored graph-based and deep learning methods. Fraud often involves relationships between accounts, merchants, devices, addresses, and transaction networks. Graph neural networks are useful because they can model these relationships rather than treating each transaction as an isolated event. Motie and Raahemi (2024) reviewed the use of graph neural networks in financial fraud detection and showed that graph-based methods are increasingly important for identifying hidden fraud networks. Chen et al. (2024) also examined intelligent sampling and self-supervised learning for credit card fraud detection, reflecting growing interest in methods that address class imbalance and limited fraud labels.

2.2. Limitations of Traditional Fraud Detection Models

Although machine learning has improved fraud detection, traditional institutional models remain limited in several ways. First, financial institutions often operate in separate data environments. A bank may detect suspicious activity within its own customer base but may not see similar activity occurring across another bank, payment platform, or merchant network. This weakens detection when fraudsters distribute activities across multiple institutions.

Second, fraud datasets are usually highly imbalanced. Legitimate transactions greatly outnumber fraudulent ones, which can cause models to favour the majority class and miss rare fraud cases. Abdul Salam et al. (2024) addressed this challenge by combining federated learning with data

balancing techniques for credit card fraud detection. Their work shows that model performance in fraud detection depends not only on the learning architecture but also on how imbalance is handled.

Third, centralised fraud detection creates privacy and security concerns. A centralised model may perform better because it can train on a larger dataset, but it requires institutions to transfer sensitive transaction records into a shared environment. This approach is difficult to justify in highly regulated financial settings. It may also increase the impact of a breach because a central repository can become an attractive target for attackers.

Fourth, fraud detection systems must balance detection accuracy with customer experience. A model with high false positives may block legitimate transactions, delay payments, frustrate customers, and increase investigation costs. A model with high false negatives may allow fraudulent transactions to proceed. For this reason, fraud detection performance must be assessed using more than general accuracy. Precision, recall, F1-score, ROC-AUC, PR-AUC, false positive rate, and false negative rate are more useful indicators for fraud detection tasks.

2.3. Federated Learning: Concept and Relevance

Federated learning is a decentralised machine learning approach that enables multiple parties to train a shared model without transferring raw data to a central location. In a typical federated process, a global model is first distributed to participating clients. Each client trains the model locally using its own data and then sends model updates to a server. The server aggregates these updates and sends the improved model back to the clients. This process continues until the model reaches an acceptable level of performance.

McMahan et al. (2017) introduced federated averaging as a communication-efficient method for training deep networks from decentralised data. Kairouz et al. (2021) later provided a broad review of federated learning, identifying major challenges such as communication efficiency, statistical heterogeneity, privacy, system reliability, and security. These issues are highly relevant to financial institutions because banks often have different data distributions, customer bases, risk profiles, fraud rates, and technology infrastructures.

In financial fraud detection, federated learning is useful because it enables institutions to benefit from shared model training while keeping sensitive transaction records local. Instead of building isolated models, institutions can contribute to a global fraud detection model through protected updates. This can improve the model's ability to detect fraud patterns that appear across multiple institutions. At the same time, federated learning reduces the privacy risks associated with direct data pooling.

Several recent studies have applied federated learning to financial fraud detection. Abdul Salam et al. (2024) developed a federated learning model for credit card fraud detection with data balancing techniques. Reddy et al. (2024) examined deep learning-based credit card fraud detection in a federated learning environment. Li et al. (2024) proposed a model combining federated learning, graph attention networks, and dilated convolutional neural networks for credit card fraud detection. These studies suggest that federated learning can support fraud detection while responding to the privacy limitations of centralised financial data sharing.

2.4. Types of Federated Learning for Financial Institutions

The literature commonly identifies three major forms of federated learning: horizontal federated learning, vertical federated learning, and federated transfer learning. Horizontal federated learning is suitable when institutions have similar feature spaces but different users or transactions. For example, several banks may collect similar transaction features, such as transaction amount, time, merchant category, channel type, and fraud label, but each bank holds data for different customers. This form is highly relevant to cross-bank fraud detection.

Vertical federated learning is used when institutions share overlapping users but hold different features. For instance, a bank, credit bureau, and fintech provider may have information about some of the same customers but from different perspectives. One may hold payment history, another may hold credit behaviour, and another may hold device or platform activity. In this case, vertical federated learning can help combine predictive signals without requiring direct exchange of all raw features.

Federated transfer learning is useful when both the users and feature spaces differ, but knowledge can still be transferred across participants. This may be relevant in financial settings where institutions operate in different markets or transaction environments but still face related fraud behaviours. However, federated transfer learning can be more complex because it requires techniques for aligning knowledge across different data structures.

The choice of federated learning type depends on the institutional context. For a network of banks using similar transaction monitoring systems, horizontal federated learning may be the most practical option. For collaborations between banks, fintech firms, insurers, credit bureaus, and payment providers, vertical or transfer-based approaches may be more appropriate.

2.5. Privacy-Enhancing Mechanisms in Federated Fraud Detection

Although federated learning keeps raw data local, it does not automatically guarantee privacy. Model updates may reveal sensitive information if attackers can reconstruct training data or infer membership from gradients or model parameters. For this reason, privacy-enhancing mechanisms are necessary in federated fraud detection systems.

Secure aggregation is one of the most important mechanisms. It allows the server to aggregate model updates without seeing each institution's individual update. Liu et al. (2022) reviewed privacy-preserving aggregation techniques

and showed that secure aggregation is central to protecting client updates in federated learning. This is particularly important in financial settings because model updates may reflect patterns in customer behaviour, fraud labels, or institution-specific risk exposure.

Differential privacy is another important safeguard. It adds controlled noise to data, model updates, or outputs so that the contribution of any individual record becomes difficult to identify. Fu et al. (2024) reviewed differentially private federated learning and showed that it can strengthen privacy protection, although it may also reduce model performance if the noise level is too high. In fraud detection, this creates a privacy-utility trade-off. Stronger privacy can protect customer information, but excessive noise may reduce the model's ability to detect rare fraudulent transactions.

Other mechanisms include secure multiparty computation, homomorphic encryption, access controls, audit logging, and encrypted communication channels. Hu et al. (2024) noted that federated learning requires both privacy and security measures because the system involves multiple participants, distributed computation, and repeated model update exchange. In financial institutions, these safeguards must be supported by legal agreements, governance rules, and compliance monitoring.

2.6. Security Risks in Federated Learning

Federated learning introduces new security risks because participating institutions exchange model updates over multiple training rounds. Malicious or compromised participants may attempt to manipulate the global model by sending poisoned updates. Poisoning attacks can reduce model accuracy, introduce hidden vulnerabilities, or cause the model to misclassify specific fraud patterns. Xia et al. (2024) reviewed privacy-preserving federated learning against poisoning attacks and highlighted the need for defence mechanisms that can identify abnormal or harmful client updates.

Inference attacks are also a major concern. Membership inference attacks attempt to determine whether a specific data record was used during training. Model inversion attacks attempt to reconstruct sensitive input features from model behaviour or updates. Bai et al. (2024) discussed membership inference attacks and defences in federated learning, showing that privacy leakage remains possible even when raw data is not directly shared. Zhao et al. (2024) also emphasised that federated learning must be evaluated against both privacy attacks and policy requirements.

In a financial setting, these risks are serious because transaction data may reveal personal spending behaviour, account activity, business relationships, merchant patterns, and fraud investigation signals. A poorly secured federated learning system may therefore create privacy risks even without centralised data sharing. For this reason, federated fraud detection requires technical defences such as anomaly detection for client updates, robust aggregation, differential privacy, secure aggregation, encryption, and continuous monitoring.

2.7. Explainability in Federated Fraud Detection

Fraud detection models must be accurate, but they must also be interpretable enough to support investigation and regulatory review. Financial institutions need to understand why a transaction has been flagged, especially when model outputs lead to transaction blocking, account restriction, customer notification, or further investigation. Black-box models may be difficult to justify in high-stakes financial decisions. Explainable artificial intelligence can help fraud analysts interpret model outputs through risk scores, feature importance, local explanations, and transaction-level indicators. Awosika et al. (2024) examined the relationship between transparency, privacy, explainable AI, and federated learning in financial fraud detection. Their study supports the view that privacy-preserving fraud detection should not focus only on predictive performance. It must also provide explanations that are useful to analysts, auditors, and compliance teams.

Aljunaid et al. (2025) also studied explainable AI-driven federated learning for financial fraud detection, showing the growing interest in combining privacy-preserving learning with transparent decision support. This is important because fraud detection is not usually a fully automated process. In most financial institutions, model alerts are reviewed by analysts, and final decisions may involve compliance, risk, and customer service teams.

2.8. Research Gap

The reviewed literature shows that federated learning has strong potential for privacy-preserving fraud detection across financial institutions. Existing studies have demonstrated its use in credit card fraud detection, deep learning-based fraud analytics, graph-based fraud modelling, explainable fraud detection, and privacy-preserving model aggregation (Abdul Salam et al., 2024; Aljunaid et al., 2025; Li et al., 2024; Reddy et al., 2024; Xia et al., 2025). However, several gaps remain. First, many studies focus mainly on model performance, while fewer provide a complete discussion of governance, privacy risks, regulatory concerns, and practical deployment requirements. Second, there is still a need for stronger frameworks that combine federated learning with secure aggregation, differential privacy, explainability, and human oversight. Third, financial institutions often operate under non-identical data conditions, but the effect of data heterogeneity on federated fraud detection remains a continuing challenge. Fourth, there is limited discussion of how federated fraud detection systems should be audited, monitored, and protected against malicious institutional participants.

This article addresses these gaps by presenting federated learning as both a technical and governance-based approach to privacy-preserving fraud detection. It focuses not only on model training but also on the privacy, security, explainability, and institutional requirements needed for cross-institution deployment.

Table 1: Summary of Key Literature Related to Federated Fraud Detection

Author(s)	Focus of Study	Relevance to This Article
McMahan et al. (2017)	Federated averaging for decentralised model training	Provides the foundational training approach for federated learning
Kairouz et al. (2021)	Challenges and open problems in federated learning	Explains core issues such as privacy, communication, and data heterogeneity
Abdul Salam et al. (2024)	Federated learning for credit card fraud detection	Directly supports the use of FL in fraud detection
Reddy et al. (2024)	Deep learning-based fraud detection in federated settings	Shows how deep learning can be applied to federated fraud detection
Li et al. (2024)	Federated fraud detection using graph attention and convolutional models	Supports advanced modelling for financial fraud patterns
Liu et al. (2022)	Privacy-preserving aggregation in federated learning	Supports the need for secure aggregation in financial applications
Fu et al. (2024)	Differential privacy in federated learning	Supports discussion of privacy-utility trade-offs
Zhao et al. (2024)	Federated learning privacy attacks, defences, and policies	Supports discussion of attacks, safeguards, and policy concerns
Awosika et al. (2024)	Explainable AI and federated learning in financial fraud detection	Supports the need for transparent fraud alerts
Xia et al. (2024)	Poisoning attacks in privacy-preserving federated learning	Supports the security risk analysis of malicious updates

3. Conceptual Framework

The conceptual framework of this article is built on the idea that financial institutions can improve fraud detection through collaborative model training without transferring raw transaction data into a central repository. In conventional fraud detection, each institution develops its own model using its internal records. This approach protects institutional control over data, but it limits the model's ability to learn fraud patterns that occur across multiple institutions. Federated learning addresses this limitation by allowing distributed institutions to train a shared model while keeping customer and transaction data within their local systems.

The framework assumes a multi-institution financial environment involving banks, fintech companies, card issuers, payment processors, and other regulated financial service providers. Each institution holds sensitive transaction data, including account history, transaction amount, merchant category, payment channel, device information, customer behaviour, and fraud labels. Instead of sharing these records directly, each institution trains a local fraud detection model on its own dataset. The local model updates are then protected and transmitted to a federated aggregation server, where they are combined to update a global fraud detection model. The improved global model is sent back to the participating institutions for further local training and fraud monitoring.

This structure follows the general principle of federated learning introduced by McMahan et al. (2017), where learning is performed across decentralised data sources through iterative local training and model aggregation. In the financial fraud detection context, this approach is useful because fraud signals may be distributed across institutions, while the data required to identify these signals remains private and regulated. Kairouz et al. (2021) also emphasised that federated learning is suitable for settings where data is distributed across multiple clients and cannot easily be centralised due to privacy, system, or institutional constraints.

The proposed conceptual framework consists of five major layers. The first layer is the institutional data layer, where each financial institution retains its own transaction records and fraud labels. The second layer is the local model training layer, where each institution trains a fraud detection model using its private data. The third layer is the privacy protection layer, where mechanisms such as secure aggregation, differential privacy, encrypted model updates, and access controls are applied. The fourth layer is the federated aggregation layer, where protected model updates are combined to improve the global model. The fifth layer is the fraud intelligence and governance layer, where the global model is used for fraud scoring, alert generation, analyst review, compliance monitoring, and model auditing.

The central value of this framework lies in its ability to support cross-institution fraud intelligence without exposing raw data. For example, one institution may observe suspicious transaction behaviours linked to account takeover, while another may observe similar device or merchant-level patterns. Through federated learning, these institutions can contribute to a stronger shared model without directly revealing customer-level records. This improves the ability to detect coordinated fraud schemes while reducing the privacy risks associated with centralised data sharing.

However, the framework also recognises that federated learning does not remove all privacy and security risks. Model updates can still reveal sensitive information if they are not properly protected. Zhao et al. (2024) noted that federated learning systems remain vulnerable to privacy attacks such as inference attacks, gradient leakage, and model inversion. For this reason, the framework includes a privacy protection layer as a core component rather than an optional addition. Secure aggregation helps prevent the server from viewing individual

institutional updates, while differential privacy can reduce the risk that model updates reveal information about specific customers or transactions (Liu et al., 2022; Fu et al., 2024).

The framework also includes governance and explainability because fraud detection is a high-stakes financial activity. A fraud detection model may influence transaction blocking, account review, customer investigation, or regulatory reporting. Therefore, model outputs should be interpretable enough for fraud analysts and compliance teams to review. Awosika et al. (2024) argued that transparency is important in financial fraud detection systems that combine federated learning and privacy-preserving techniques. In this article, explainability is positioned as part of the decision-support layer, where model outputs are translated into risk scores, fraud alerts, and feature-level explanations that support human judgement.

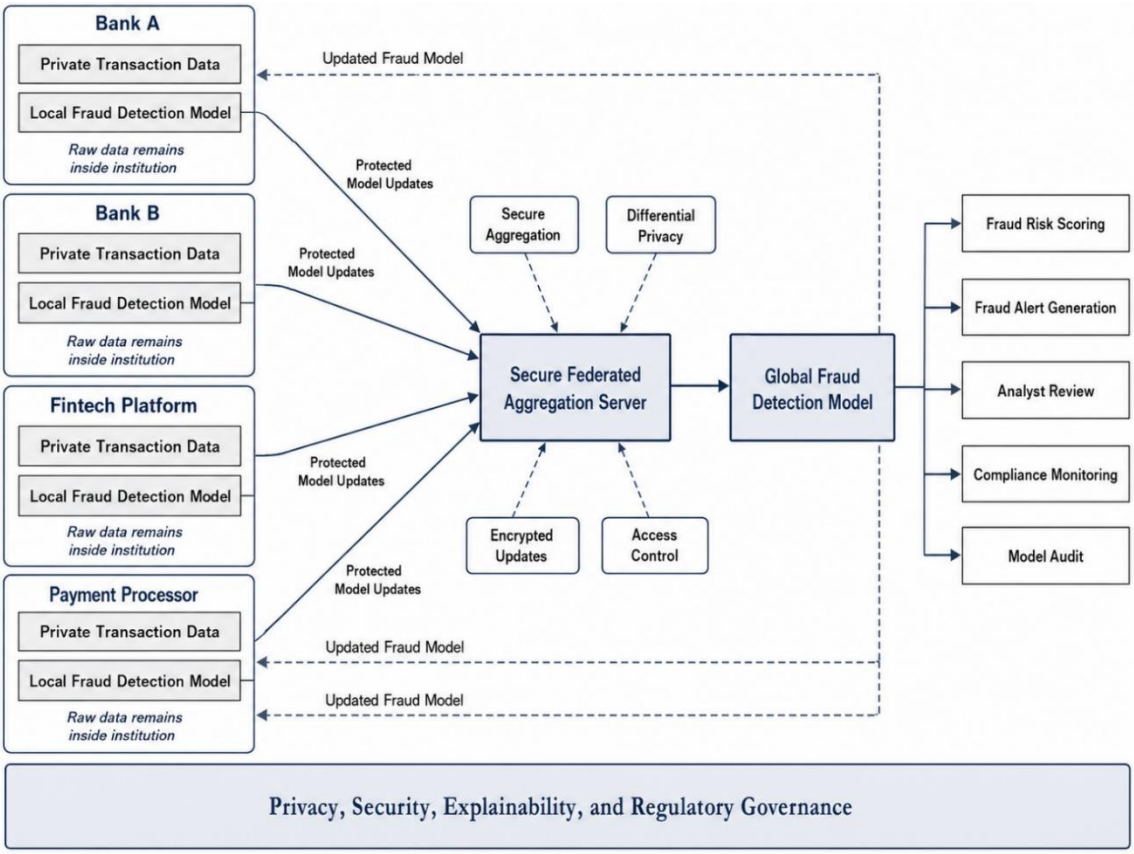

Fig 1: Federated Learning Architecture for Privacy-Preserving Fraud Detection across Financial Institutions

4. Proposed Federated Fraud Detection Framework

The proposed framework is designed to support privacy-preserving fraud detection across financial institutions through coordinated but decentralised model training. It responds to two major needs in financial fraud analytics. The first is the need for wider fraud intelligence across institutions, since fraudsters often exploit gaps between banks, fintech platforms, payment processors, and card networks. The second is the need to protect sensitive financial data, since customer transaction records cannot be freely transferred or pooled without creating privacy, legal, and security risks.

The framework is structured around a federated training cycle. At the beginning of the process, a baseline fraud detection model is initialised and distributed to participating institutions. Each institution trains the model locally using its own private transaction data. After local training, the institution does not send raw transaction records to the central server. Instead, it sends protected model updates, such as weights, gradients, or parameter changes. These updates are then aggregated to produce an improved global fraud detection model. The global model is redistributed to the institutions, and the cycle continues over multiple training rounds.

This approach reflects the federated learning structure used in decentralised model training, where learning takes place locally and only model updates are exchanged (McMahan et al., 2017). In fraud detection, this design is particularly useful because it enables institutions to learn from broader fraud patterns while maintaining control over customer-level data. Abdul Salam et al. (2024) showed that federated learning can be applied to credit card fraud detection, especially when combined with techniques that address class imbalance. Reddy et al. (2024) also demonstrated the relevance of deep learning-based fraud detection in a federated learning setting.

4.1. System Architecture

The proposed system architecture contains six main components: participating institutions, local transaction databases, local fraud detection models, a privacy protection layer, a federated aggregation server, and a fraud decision-support layer.

Participating institutions are the clients in the federated learning system. These may include commercial banks, digital banks, fintech companies, credit card issuers, mobile money operators, payment gateways, and other financial service providers. Each participant stores its transaction data locally and performs model training within its own computing environment. This helps reduce the risks associated with transferring sensitive data into a shared repository.

The local transaction database contains the data used for fraud detection. This may include transaction amount, transaction time, merchant category, payment channel, device identifier, login behaviour, transaction frequency, geographic pattern, account age, previous chargebacks, and fraud labels. The actual features may vary across institutions depending on data availability, regulatory requirements, and system design.

The local fraud detection model is trained within each institution. The model may be a logistic regression model, gradient boosting model, neural network, graph-based model, or hybrid classifier. The choice of model depends on the size of the dataset, the need for interpretability, the available computing infrastructure, and the complexity of fraud patterns. For large transaction networks, graph-based models may be useful because they can capture relationships between accounts, merchants, devices, and transaction flows. Motie and Raahemi (2024) showed that graph neural networks are increasingly relevant for financial fraud detection because fraud often involves hidden relationships rather than isolated transaction events.

The privacy protection layer protects the model updates before they are shared. This layer may include secure aggregation, differential privacy, encrypted communication, and update validation. Secure aggregation prevents the aggregation server from viewing each institution's individual model update. Differential privacy adds controlled noise to reduce the risk that model updates reveal information about specific customers or transactions. Liu et al. (2022) described privacy-preserving aggregation as a major requirement for federated learning, while Fu et al. (2024) showed that differentially private federated learning can strengthen privacy protection when properly designed.

The federated aggregation server combines the protected model updates received from participating institutions. Its role is not to access raw data, but to coordinate the training process and update the global model. Depending on the system design, the aggregation server may be operated by a trusted financial consortium, a regulator-approved technology provider, a payment network, or a neutral governance body. In more advanced settings, decentralised aggregation may be used to reduce dependence on a single central server.

The fraud decision-support layer applies the trained model to detect suspicious transactions. It generates fraud risk scores, transaction alerts, feature-level explanations, and investigation priorities. Since fraud detection can affect customers and financial operations, the output should support human review rather than operate as a fully opaque automated decision system. Awosika et al. (2024) highlighted the importance of explainability in privacy-preserving financial fraud detection, especially where model outputs must be interpreted by analysts and compliance teams.

4.2. Data Structure and Local Training

Each institution trains its model using local transaction data. The dataset may contain structured transaction features, behavioural variables, customer account features, merchant-level indicators, device-level attributes, and fraud labels. Since financial fraud is usually rare compared with legitimate transactions, class imbalance is expected. This means that a model may achieve high accuracy while still failing to detect actual fraud cases. For this reason, the training process should include methods such as balanced sampling, class weighting, cost-sensitive learning, or synthetic minority oversampling where appropriate.

Local training should also reflect the differences between participating institutions. One bank may serve retail customers, another may focus on corporate accounts, while a fintech platform may process small but frequent mobile payments. These differences can create non-identically distributed data across institutions. Kairouz et al. (2021) identified statistical heterogeneity as one of the central challenges in federated learning. In fraud detection, this challenge is serious because fraud rates, customer behaviour, transaction volumes, and merchant categories may differ widely across institutions.

To address this, the proposed framework allows institutions to train locally while contributing to a shared global model. The global model learns broader fraud patterns, while each institution may also fine-tune the model to its own environment. This hybrid approach helps balance general fraud intelligence with institution-specific risk patterns.

4.3. Federated Training Process

The federated training process follows a structured sequence. First, the federation coordinator initialises a baseline fraud detection model and distributes it to participating institutions. Second, each institution trains the model locally using its private transaction dataset. Third, each institution applies privacy protection to its model updates. Fourth, protected updates are sent to the aggregation server. Fifth, the server aggregates the updates to generate an improved global model. Sixth, the updated global model is sent back to all participating institutions. Seventh, institutions apply the model to local fraud detection tasks and continue training in later rounds.

This iterative process allows the model to improve over time as it learns from distributed fraud patterns. It also reduces the need for centralised data storage. However, the quality of

the global model depends on the reliability of participating institutions, the representativeness of local data, and the security of the update aggregation process. If some institutions submit poor-quality, biased, or malicious updates, the global model may be weakened. For this reason, the framework includes update validation and monitoring procedures.

4.4. Privacy Protection Layer

The privacy protection layer is one of the most important parts of the proposed framework. Although federated learning prevents raw data transfer, model updates may still contain sensitive information. Attackers may attempt to infer whether a customer was included in a local dataset, reconstruct transaction features, or identify institution-specific fraud patterns. Zhao et al. (2024) showed that federated learning systems face several privacy risks, including inference attacks and model leakage.

To reduce these risks, the framework includes four privacy safeguards. The first is secure aggregation, which ensures that the server sees only the combined update rather than the individual update from each institution. The second is differential privacy, which adds controlled noise to limit the exposure of individual records. The third is encrypted communication, which protects model updates during transmission. The fourth is access control and authentication, which ensures that only approved institutions participate in the federation.

These safeguards must be carefully balanced. Excessive privacy noise may reduce the model's ability to detect rare fraud events, while weak privacy controls may expose sensitive information. Therefore, the framework treats privacy as a measurable design factor rather than a general claim. The system should monitor privacy budget, model performance, and communication cost across training rounds.

4.5. Fraud Detection Model Options

The proposed framework can support different fraud detection models depending on the research design and institutional requirements. Simpler models such as logistic regression and decision trees may be useful when interpretability is a priority. Ensemble methods such as random forests and gradient boosting may improve predictive performance while still offering some level of explainability. Deep learning models may be useful for large-scale transaction data, especially when fraud patterns are complex and nonlinear.

Graph-based models may also be appropriate where fraud involves relationships among accounts, merchants, devices, and transactions. Li et al. (2024) proposed a federated fraud detection model combining graph attention networks and convolutional methods, showing the growing role of graph-based learning in federated financial fraud detection. Xia et al. (2025) also explored graph-based federated learning for privacy-preserving credit card fraud detection, which supports the view that relationship-based modelling can strengthen fraud intelligence.

For practical deployment, the model choice should consider accuracy, recall, false positive rate, interpretability, computational cost, and communication efficiency. In fraud detection, recall is especially important because missing fraudulent transactions can lead to financial loss. However, precision is also important because excessive false alerts increase operational costs and may harm customer trust.

4.6. Explainability and Analyst Decision Support

The proposed framework includes explainability as a core part of fraud decision support. Fraud alerts should not only produce a risk score but also provide useful reasons for the alert. These may include unusual transaction amount, abnormal login location, unfamiliar device, high-risk merchant category, rapid transaction frequency, or deviation from the customer's normal behaviour.

Explainability supports fraud analysts by helping them understand which features contributed to a model decision. It also supports compliance review because financial institutions may need to justify why a transaction was blocked or why an account was flagged for investigation. Aljunaid et al. (2025) showed that explainable AI can be integrated with federated learning for financial fraud detection, especially where transparency and trust are important.

In the proposed framework, explainability is applied at the local institutional level. This means each institution can interpret fraud alerts using its own customer and transaction context. The global model provides shared fraud intelligence, but the final review remains with local analysts who understand institutional policies, customer history, and regulatory obligations.

4.7. Security and Governance Controls

A federated fraud detection framework must include security controls against malicious or unreliable participants. A participating institution may submit corrupted updates due to system compromise, poor data quality, or intentional manipulation. Poisoning attacks are especially concerning because they can weaken the global model or cause it to misclassify specific fraud patterns. Xia et al. (2024) noted that poisoning attacks remain a major challenge for privacy-preserving federated learning.

To address this risk, the framework includes update validation, anomaly detection for model updates, robust aggregation, participant authentication, audit trails, and periodic security review. Institutions should also agree on participation rules, minimum data quality standards, model evaluation procedures, incident reporting obligations, and liability arrangements.

Governance is equally important. A federated fraud detection network requires clear rules on who can participate, how model updates are protected, how performance is evaluated, how privacy budgets are managed, and how disputes are resolved. Regulatory compliance should be built into the framework from the beginning, especially where

institutions operate across jurisdictions or process data subject to financial privacy laws.

5. Methodology

5.1. Research Design

This study adopts a framework-based and experimental research design to examine how federated learning can be used for privacy-preserving fraud detection across financial institutions. The framework-based component explains the structure, processes, privacy safeguards, and governance requirements of a federated fraud detection system. The experimental component provides a basis for evaluating whether federated learning can achieve competitive fraud detection performance when compared with institution-specific and centralised modelling approaches.

The proposed design is suitable because fraud detection in financial institutions involves both technical and regulatory concerns. A purely technical model evaluation would not fully address privacy, institutional trust, data governance, and explainability. Similarly, a purely conceptual discussion would not be sufficient to show whether federated learning can produce useful fraud detection results. For this reason, the methodology combines model comparison, privacy assessment, communication analysis, and practical deployment considerations.

Federated learning has already been applied to credit card fraud detection in recent studies. Abdul Salam et al. (2024) demonstrated the relevance of federated learning for credit card fraud detection, especially when combined with data balancing techniques. Reddy et al. (2024) also examined deep learning-based credit card fraud detection under a federated learning setting. These studies support the use of an experimental design that compares federated learning with conventional fraud detection approaches.

5.2. Data Source

The study may use either a publicly available financial fraud dataset, a synthetic multi-institution transaction dataset, or anonymised institutional data where access is available. If real banking data is unavailable, a public credit card fraud dataset can be partitioned to simulate multiple financial institutions. This approach allows the study to test federated learning under controlled conditions while avoiding the ethical and legal issues associated with exposing sensitive customer data.

The dataset should contain transaction-level records with variables that are relevant to fraud detection. These may include transaction amount, transaction time, merchant category, payment channel, account behaviour, device indicators, location-related patterns, and fraud labels. Where a public dataset has anonymised features, the study should clearly state that the variables have been transformed or masked for confidentiality.

Since fraud cases are usually rare, the dataset is expected to show class imbalance. This imbalance should be handled carefully because a model may report high overall accuracy while failing to detect fraudulent transactions. Abdul Salam et al. (2024) addressed this problem by combining federated learning with data balancing methods. In this study, class imbalance may be handled through class weighting, balanced sampling, cost-sensitive learning, or suitable oversampling methods, depending on the dataset structure.

5.3. Data Partitioning Strategy

To simulate a multi-institution environment, the dataset should be divided across several participating institutions. For example, the data may be split into five simulated financial institutions, each representing a bank, fintech platform, payment processor, card issuer, or digital wallet provider. Each institution should retain its own local data and train a model without sharing raw records.

Two data partitioning scenarios should be considered. The first is an independent and identically distributed setting, where each institution receives a relatively similar distribution of transaction patterns and fraud labels. The second is a non-independent and non-identically distributed setting, where institutions receive different fraud rates, transaction behaviours, customer profiles, or merchant categories. The second setting is closer to real financial environments, where institutions differ in size, customer base, transaction volume, risk exposure, and fraud typology.

Kairouz et al. (2021) identified statistical heterogeneity as one of the major challenges in federated learning. This is important for fraud detection because data differences between institutions may affect model convergence, performance, and fairness. For this reason, the methodology should evaluate the federated model under both balanced and heterogeneous institutional data conditions.

5.4. Model Development

The study should compare three main modelling approaches. The first is the local-only model, where each institution trains and evaluates a fraud detection model using only its own data. This represents the conventional institutional approach to fraud detection. The second is the centralised model, where all data is pooled into one dataset and used to train a single model. This represents the ideal performance benchmark, although it may not be practical or legally acceptable in real banking environments. The third is the federated learning model, where each institution trains locally and shares only protected model updates for aggregation.

The local-only model is important because it shows how well an institution can detect fraud without collaboration. The centralised model is important because it provides a reference point for maximum performance under direct data pooling. The federated model is the main focus because it tests whether collaborative learning can approach centralised model performance while preserving local data control.

Possible models include logistic regression, random forest, gradient boosting, multilayer perceptron, or graph-based models. The final model choice should depend on the

dataset, computational resources, interpretability requirements, and research scope. For a balanced research design, the study may use one conventional machine learning model and one deep learning model. Where transaction relationships are available, a graph-based model may also be tested, since graph neural networks have become relevant in financial fraud detection research (Motie & Raahemi, 2024).

5.5. Federated Learning Procedure

The federated learning process should follow a repeated training cycle. First, a global fraud detection model is initialised. Second, the global model is distributed to all participating institutions. Third, each institution trains the model locally using its private transaction data. Fourth, the institution generates model updates after local training. Fifth, privacy protection mechanisms are applied to the updates before transmission. Sixth, the protected updates are sent to the aggregation server. Seventh, the aggregation server combines the updates to generate a new global model. Finally, the updated global model is redistributed to the institutions for the next training round.

Federated averaging may be used as the aggregation method, since it is one of the foundational methods for communication-efficient distributed learning (McMahan et al., 2017). However, the aggregation method should be supported by security checks to reduce the influence of unreliable or malicious updates. Where privacy protection is included, secure aggregation and differential privacy should be considered. Liu et al. (2022) noted that privacy-preserving aggregation is central to secure federated learning, while Fu et al. (2024) showed that differential privacy can reduce information leakage in federated systems.

5.6. Privacy and Security Configuration

The experimental design should include privacy and security settings that reflect the risks of financial data sharing. Three privacy configurations may be tested. The first is federated learning without additional privacy protection, where local updates are transmitted normally. The second is federated learning with secure aggregation, where the aggregation server receives only combined model updates. The third is federated learning with differential privacy, where controlled noise is added to reduce the risk of exposing information about individual transactions.

The study should also consider update validation. Since federated learning involves multiple participants, there is a risk that one or more institutions may submit corrupted, poor-quality, or malicious updates. Poisoning attacks are a known challenge in federated learning systems, and defences are needed to protect the global model from harmful updates (Xia et al., 2024). The methodology should therefore include a basic update monitoring process to identify abnormal update patterns.

5.7. Evaluation Metrics

Fraud detection should not be evaluated using accuracy alone because fraud datasets are usually imbalanced. A model may classify most legitimate transactions correctly and still fail to detect fraud. Therefore, the evaluation should include precision, recall, F1-score, ROC-AUC, PR-AUC, false positive rate, and false negative rate.

Recall is important because it measures the proportion of fraudulent transactions correctly detected. Precision is important because it shows how many flagged transactions are actually fraudulent. F1-score provides a balance between precision and recall. PR-AUC is particularly useful in imbalanced fraud detection tasks because it focuses on the relationship between precision and recall. ROC-AUC can also be reported, but it should be interpreted carefully where fraud cases are rare.

In addition to predictive performance, the study should assess communication cost, training time, convergence behaviour, and privacy-utility trade-off. Communication cost may be measured by the number of training rounds, size of model updates, and total transmitted data. Privacy-utility trade-off may be assessed by comparing model performance under different levels of differential privacy.

5.8. Experimental Scenarios

The study should evaluate the proposed model under the following experimental scenarios:

Table 2: Comparative Analysis of Fraud Detection Training Scenarios in Federated Learning Environments

Scenario	Description	Purpose
Local-only model	Each institution trains a separate fraud detection model using only local data	Measures the limitation of isolated institutional learning
Centralised model	All data is pooled and used to train one model	Provides a benchmark for maximum possible model performance
Federated learning without privacy layer	Institutions train locally and share model updates	Tests collaborative learning without added privacy protection
Federated learning with secure aggregation	Individual updates are protected during aggregation	Tests privacy-preserving update aggregation
Federated learning with differential privacy	Noise is added to model updates before aggregation	Tests the privacy-utility trade-off
Federated learning under non-IID data	Institutions have different transaction and fraud distributions	Tests performance under realistic institutional heterogeneity

5.9. Ethical and Regulatory Considerations

The methodology should ensure that raw transaction data remains within each participating institution. If public or synthetic data is used, the study should clearly state the source and explain how the multi-institution setting was simulated. If real institutional data is used, the study should describe anonymisation, access control, ethical approval, data minimisation, and compliance procedures.

The methodology should also recognise that fraud detection can affect customers directly. False positives may lead to blocked transactions, account review, or customer inconvenience. False negatives may allow financial losses and expose institutions to risk. For this reason, the proposed system should include human review, explainability, and audit logging. Awosika et al. (2024) emphasised that transparency is important when federated learning and explainable artificial intelligence are applied to financial fraud detection.

6. Results and Analysis

6.1. Overview of Expected Analysis

The results and analysis section should present the performance of the proposed federated fraud detection framework across the experimental scenarios described in the methodology. Since this section depends on empirical testing, numerical values should only be inserted after model training and evaluation have been completed. The analysis should compare local-only, centralised, and federated models using predictive performance, privacy protection, communication cost, and practical suitability for financial institutions.

The main purpose of the analysis is to determine whether federated learning can provide a practical balance between fraud detection accuracy and data privacy. A centralised model may produce strong predictive performance because it has access to all transaction records, but such an approach is often difficult to implement in regulated financial settings. A local-only model protects institutional data but may fail to learn wider fraud patterns. The federated model should be assessed based on whether it can approach centralised performance while preserving local data control.

6.2. Model Performance Comparison

The first stage of analysis should compare fraud detection performance across the local-only, centralised, and federated learning models. The comparison should include precision, recall, F1-score, ROC-AUC, and PR-AUC. In fraud detection, recall and PR-AUC are especially important because fraudulent transactions are usually rare. A high recall value indicates that the model detects a larger proportion of fraud cases, while PR-AUC provides a better view of performance under class imbalance.

The expected outcome is that the local-only model may perform inconsistently across institutions because each institution has access only to its own data. Smaller institutions or those with fewer fraud cases may have weaker models due to limited training examples. The centralised model may produce the strongest performance because it is trained on the combined dataset. However, the federated learning model is expected to perform better than most local-only models because it benefits from collaborative training across institutions.

This expectation is consistent with recent studies showing that federated learning can support credit card fraud detection while limiting raw data sharing (Abdul Salam et al., 2024; Reddy et al., 2024). However, performance may depend on the number of institutions, data quality, class imbalance, aggregation method, and privacy configuration.

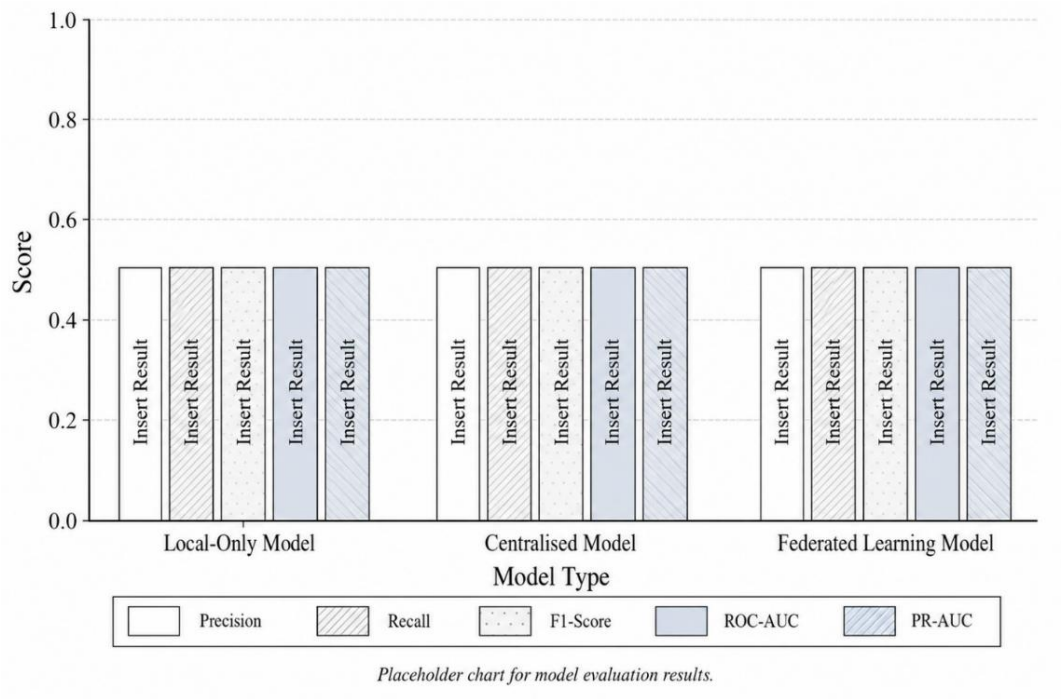

Fig 2: Performance Comparison of Fraud Detection Models

6.3. Analysis of Precision, Recall, and F1-Score

Precision should be analysed to determine how many flagged transactions are truly fraudulent. High precision is important because excessive false alerts increase the workload of fraud analysts and may inconvenience customers. However, precision alone is not enough because a model may flag only the most obvious fraud cases and miss many others.

Recall should be analysed to determine how many actual fraud cases are detected. In fraud detection, recall is often a priority because undetected fraud can lead to financial loss, customer harm, and regulatory concern. However, recall must be balanced against false positives. A model that flags too many legitimate transactions may reduce customer trust and increase investigation costs.

F1-score should be used to assess the balance between precision and recall. If the federated learning model achieves a higher F1-score than the local-only models, this would suggest that cross-institution training improves fraud detection while keeping data decentralised. If the federated model performs close to the centralised model, the result would support the argument that federated learning provides a practical privacy-preserving alternative to data pooling.

6.4. Privacy-Utility Trade-Off

The second stage of analysis should examine the relationship between privacy protection and model performance. Privacy-preserving mechanisms such as differential privacy can reduce the risk of information leakage, but they may also affect model accuracy if the noise level is too high. For this reason, the study should test different privacy settings and compare their effect on recall, F1-score, and PR-AUC.

The expected pattern is that moderate privacy protection may maintain acceptable fraud detection performance, while very strict privacy settings may reduce the model's ability to identify rare fraud patterns. This privacy-utility trade-off has been identified in differentially private federated learning research (Fu et al., 2024). In the context of financial fraud detection, the trade-off is particularly important because the model must protect customer data while still detecting suspicious transactions accurately.

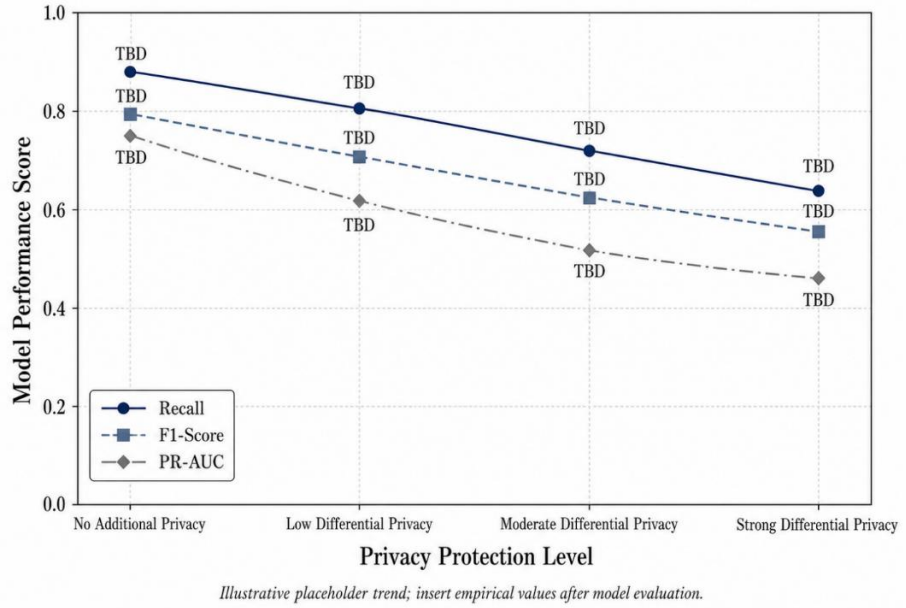

Fig 3: Privacy-Utility Trade-Off in Federated Fraud Detection

6.5. Communication and Scalability Analysis

The third stage of analysis should assess communication cost and scalability. Federated learning requires repeated communication between the participating institutions and the aggregation server. Each training round involves sending the global model to institutions, local training, transmission of model updates, aggregation, and redistribution of the updated model. As the number of institutions increases, communication requirements may also increase.

The analysis should report the number of communication rounds, model update size, total transmitted data, and training time. These indicators are important because financial institutions may have different levels of technical infrastructure. A model that performs well but requires excessive communication may be difficult to deploy in real-time fraud detection environments. Bonawitz et al. (2019) highlighted the importance of system design for federated learning at scale, especially where large numbers of participants and repeated communication rounds are involved.

The analysis should also consider whether all participating institutions contribute equally to model improvement. Large institutions may have more transaction data and therefore greater influence on the global model. Smaller institutions may benefit from the shared model but contribute fewer updates. This raises practical questions about fairness, weighting, and institutional participation in a federated fraud detection network.

6.6. Performance under Non-IID Data

The fourth stage of analysis should examine model performance under non-IID data conditions. In real financial environments, institutions rarely have identical transaction distributions. One institution may process high-value corporate transfers, while another may process low-value retail transactions. A fintech platform may have frequent mobile payments, while a credit card issuer may have merchant-based card transactions. These differences can affect model training and aggregation.

The federated learning model should therefore be tested under heterogeneous institutional data distributions. If performance declines significantly under non-IID conditions, the analysis should discuss possible reasons, such as uneven fraud rates, different feature distributions, or institution-specific transaction behaviour. If performance remains stable, this would support the suitability of federated learning for real-world financial settings.

Kairouz et al. (2021) identified non-IID data as a central challenge in federated learning. In fraud detection, this issue is especially important because fraud patterns may be concentrated in certain institutions, channels, or customer groups. The analysis should therefore report results separately for each institution as well as for the overall global model.

6.7. Explainability of Fraud Predictions

The analysis should also assess the interpretability of fraud alerts generated by the model. Fraud detection systems should provide analysts with useful reasons for model decisions. These reasons may include unusual transaction value, unfamiliar merchant, abnormal device activity, rapid transaction frequency, high-risk location, or deviation from normal customer behaviour.

Explainability should be reported through feature importance summaries, local explanation examples, or ranked fraud indicators. This is important because financial institutions must often justify fraud-related decisions to customers, auditors, and regulators. Awosika et al. (2024) argued that explainability is an important requirement in privacy-preserving financial fraud detection systems. Aljunaid et al. (2025) also showed that explainable AI can be combined with federated learning to improve transparency in financial fraud detection.

The analysis should avoid presenting the model as a replacement for fraud analysts. Instead, the model should be treated as a decision-support tool that helps analysts prioritise suspicious transactions and review risk indicators.

6.8. Security Analysis

The security analysis should examine how the framework responds to risks such as poisoned updates, unreliable clients, and inference attacks. A basic test may involve introducing abnormal model updates from one simulated institution and observing whether update validation or robust aggregation reduces the effect on the global model. This type of analysis is useful because federated learning systems depend on repeated contributions from multiple participants.

Poisoning attacks are a known risk in federated learning, especially where a malicious or compromised participant attempts to influence the global model (Xia et al., 2024). In financial fraud detection, this risk could be serious because a manipulated model may fail to detect specific fraud patterns. The results should therefore discuss whether the framework includes sufficient safeguards, such as update validation, anomaly detection, secure aggregation, and audit trails.

6.9. Summary of Results

The final results table should summarise the main outcomes across all experimental scenarios. It should include predictive performance, privacy configuration, communication cost, and practical interpretation. The table should not report values until the model has been tested. If the study is conceptual rather than experimental, the table may be presented as an evaluation template.

Table 3: Summary of Model Performance across Experimental Conditions

Experimental Condition	Precision	Recall	F1-Score	ROC-AUC	PR-AUC	Privacy Mechanism	Communication Cost	Interpretation
Local-only model	Insert result	Insert result	Insert result	Insert result	Insert result	None	Low	Shows baseline institutional performance
Centralised model	Insert result	Insert result	Insert result	Insert result	Insert result	None	Not applicable	Provides benchmark performance but requires data pooling
Federated learning	Insert result	Insert result	Insert result	Insert result	Insert result	Local data retention	Moderate	Tests collaborative learning without raw data sharing
Federated learning with secure aggregation	Insert result	Insert result	Insert result	Insert result	Insert result	Secure aggregation	Moderate	Protects individual institutional updates

Federated learning with differential privacy	Insert result	Insert result	Insert result	Insert result	Insert result	Differential privacy	Moderate	Tests privacy-utility trade-off
Federated learning under non-IID data	Insert result	Insert result	Insert result	Insert result	Insert result	Local data retention and update protection	Variable	Tests performance under realistic institutional differences

6.10. Interpretation of Expected Findings

If the federated learning model performs better than local-only models, the findings would support the claim that collaborative learning can improve fraud detection across institutions. If the federated model performs close to the centralised model, the findings would further suggest that federated learning can provide a practical alternative to data pooling. This would be important for financial institutions that need wider fraud intelligence but cannot freely share customer-level data.

If privacy-enhanced federated learning shows a slight reduction in performance, the result should be interpreted as a trade-off rather than a failure. In financial settings, a small loss in model performance may be acceptable if it significantly improves data protection and regulatory acceptability. However, if performance drops sharply under strong privacy settings, the study should recommend careful privacy budget selection and further tuning.

If the model performs poorly under non-IID data, the analysis should explain that real-world financial institutions may require personalised federated learning, institution-level calibration, or weighted aggregation. If performance remains stable, the result would support the use of federated learning in cross-institution fraud detection networks.

Overall, the results should be interpreted in relation to three main criteria: fraud detection performance, privacy preservation, and operational feasibility. A successful federated fraud detection framework should not only improve predictive performance but also protect sensitive data, reduce unnecessary data sharing, support analyst review, and remain practical for financial institutions to deploy.

7. Limitations of the Study

This study has several limitations. First, access to real multi-institution financial fraud datasets may be restricted due to privacy, regulatory, and commercial concerns. As a result, the study may rely on public or synthetic datasets, which may not fully reflect the complexity of fraud behaviour across actual financial institutions.

Second, the proposed framework may not capture all operational differences between banks, fintech platforms, payment processors, and card issuers. In practice, institutions may use different data structures, fraud labels, transaction monitoring systems, and compliance procedures. These differences may affect the performance and deployment of a federated fraud detection model.

Third, federated learning may be affected by non-identically distributed data. Financial institutions often serve different customer groups, process different transaction types, and experience different fraud patterns. Such variation can influence model convergence, accuracy, and fairness across institutions.

Fourth, the privacy-preserving mechanisms discussed in the study may introduce performance trade-offs. Differential privacy, encrypted updates, and secure aggregation can strengthen data protection, but they may increase computational cost or reduce model performance if not carefully configured.

Finally, the study is limited in its treatment of real-time deployment. Fraud detection often requires immediate response, while federated learning involves repeated training, communication, and aggregation rounds. Further testing would be needed to determine how the framework performs in live financial environments.

8. Future Research Directions

Future research should test the proposed framework using real-world multi-institution datasets, subject to proper ethical approval, anonymisation, and regulatory safeguards. This would provide stronger evidence of how federated learning performs under practical banking and fintech conditions.

Further studies should examine personalised federated learning models that allow institutions to benefit from a shared global model while adapting predictions to their own customer profiles, transaction channels, and fraud patterns. This may be useful where data distributions differ significantly across institutions.

Future work should also explore federated graph neural networks for detecting fraud networks involving linked accounts, merchants, devices, payment channels, and digital identities. Graph-based federated learning may improve the detection of coordinated fraud schemes that are difficult to identify from isolated transaction records.

Another important direction is the development of stronger defences against poisoning attacks, inference attacks, and malicious model updates. Future research should evaluate robust aggregation methods, update validation, anomaly detection, and secure audit mechanisms within federated financial systems.

Further research should also examine explainable federated fraud detection. Financial institutions need models that not only detect suspicious transactions but also provide clear reasons that fraud analysts, auditors, and compliance teams can review.

Finally, future studies should investigate regulatory and governance models for cross-institution federated learning. This includes participation rules, liability arrangements, audit requirements, privacy budget management, and standards for secure collaboration between financial institutions.

9. Conclusion

This article examined federated learning as a privacy-preserving approach to fraud detection across financial institutions. The discussion showed that traditional fraud detection models are often limited by institutional data silos, while centralised data sharing creates privacy, security, and regulatory concerns. Federated learning provides a practical alternative by allowing institutions to train a shared fraud detection model without transferring raw customer transaction data.

The proposed framework combines local model training, protected model update sharing, secure aggregation, global model improvement, fraud risk scoring, analyst review, and governance oversight. It also recognises the need for privacy safeguards such as differential privacy, encrypted updates, access control, and update validation. These mechanisms are necessary because federated learning can still be exposed to inference attacks, poisoning attacks, and model leakage if poorly implemented.

The article concludes that federated learning can strengthen collaborative fraud intelligence while preserving institutional control over sensitive data. However, successful deployment depends on more than model accuracy. It requires privacy protection, explainability, secure infrastructure, regulatory compliance, human oversight, and trust among participating institutions. With proper design and governance, federated learning can become an important tool for privacy-preserving fraud detection in modern financial systems.

References

1. Abdul Salam, M., Fouad, K. M., Elbably, D. L., & Elsayed, S. M. (2024). Federated learning model for credit card fraud detection with data balancing techniques. *Neural Computing and Applications, 36*, 6231–6256. https://doi.org/10.1007/s00521-023-09410-2
2. Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. *Journal of Risk and Financial Management, 18*(4), 179. https://doi.org/10.3390/jrfm18040179
3. Li, M., Zhang, Y., Wang, X., Chen, J., & Liu, H. (2024). FedGAT-DCNN: Advanced credit card fraud detection using federated learning, graph attention networks, and dilated convolutions. *Electronics, 13*(16), 3169. https://doi.org/10.3390/electronics13163169
4. Baabdullah, T., Alzahrani, A., Alharbi, F., & Alshammari, M. (2024). Efficiency of federated learning and blockchain in preserving privacy and enhancing the performance of credit card fraud detection systems. *Future Internet, 16*(6), 196. https://doi.org/10.3390/fi16060196
5. Reddy, V. V. K., Reddy, R. V. K., Munaga, M. S. K., Karnam, B., Maddila, S. K., & Kolli, C. S. (2024). Deep learning-based credit card fraud detection in federated learning. *Expert Systems with Applications, 251*, 124493. https://doi.org/10.1016/j.eswa.2024.124493
6. Awosika, T., Shukla, R. M., & Pranggono, B. (2024). Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. *IEEE Access, 12*, 64551–64560. https://doi.org/10.1109/ACCESS.2024.3394528
7. Xia, Z., Liu, Y., Zhang, H., Chen, L., & Wang, Q. (2025). FinGraphFL: Financial graph-based federated learning for privacy-preserving credit card fraud detection. *Mathematics, 13*(9), 1396. https://doi.org/10.3390/math13091396
8. Nagraj, A. (2024). GraphQL in Wealth Management Platforms: Optimizing Data Access and Performance. British Journal of Multidisciplinary Studies, 2(1), 16-24.
9. Takon, A. (2024). Data-Driven Threat Intelligence for Energy and Critical Asset Management. International Journal of Technology, Management and Humanities, 10(04), 253-266.
10. Kola, J. N. Longitudinal Cohort Intelligence for Self-Insured Employer Groups: A Predictive Framework for Healthcare Cost Trajectory Modeling and Proactive Risk Intervention.
11. Adepoju, S. A., & Adepoju, M. A. (2024). From Portals to Case Graphs: A Reference Architecture and Benchmark for Safety Investigation Operations with Agentic Orchestration.
12. Takon, A. (2024). Data Science Approaches to Asset Integrity Management in Offshore and Onshore Oil and Gas Operations. Multidisciplinary Innovations & Research Analysis, 5(2), 17-31.
13. Kola, J. N. (2011). An Integrated Framework for Data Mining and Distributed Database Optimization in Resource-Constrained Network Environments. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 2(02), 82-86.
14. Ravikumar, V. (2014). Fair and optimal resource allocation in wireless sensor networks.
15. Naidu, K. J. (2014). Secure OLAP Reporting Architectures: Integrating Role-based Access Control and Query Execution Plan Optimization for Enterprise Analytical Environments. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 5(02), 155-159.
16. Marasani, Y. (2025). Explainable AI Frameworks for Patient-Level Claims Data Analytics. J Artif Intell Mach Learn & Data Sci, 8(1), 3382-3390.
17. Zhao, J. C., Bagchi, S., Avestimehr, S., Chan, K. S., Chaterji, S., Dimitriadis, D., Li, J., Li, N., Nourian, A., & Roth, H. R. (2024). Federated learning privacy: Attacks, defenses, applications, and policy landscape: A survey. *ACM Computing Surveys*. https://doi.org/10.1145/3724113

18. Fu, J., Hong, Y., Ling, X., Wang, L., Ran, X., Sun, Z., Wang, W. H., Chen, Z., & Cao, Y. (2024). Differentially private federated learning: A systematic review. *arXiv*. https://doi.org/10.48550/arXiv.2405.08299
19. Liu, Z., Guo, J., Yang, W., Fan, J., Lam, K.-Y., & Zhao, J. (2022). Privacy-preserving aggregation in federated learning: A survey. *IEEE Transactions on Big Data*. https://doi.org/10.1109/TBDATA.2022.3180613
20. Nagraj, A. (2022). Modernizing Legacy Banking Systems: Migration Strategies and Cost Optimization in Financial Enterprises. Frontiers in Computer Science and Artificial Intelligence, 1(1), 43-52.
21. Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2024). Privacy issues, attacks, countermeasures and open problems in federated learning. *Applied Artificial Intelligence, 38*(1), 2410504. https://doi.org/10.1080/08839514.2024.2410504
22. Hu, K., Li, J., Ding, Y., Bai, X., & Yang, F. (2024). An overview of implementing security and privacy in federated learning. *Artificial Intelligence Review, 57*, 1–39. https://doi.org/10.1007/s10462-024-10846-8
23. Xia, F., Yu, X., Zhang, J., & Yang, L. T. (2024). A survey on privacy-preserving federated learning against poisoning attacks. *Cluster Computing*. https://doi.org/10.1007/s10586-024-04629-7
24. ALAMPALLY, J. (2024). Enhancing data quality and trust in AI systems through robust data engineering. Frontiers in Computer Science and Artificial Intelligence, 3(1), 120-130.
25. Bai, L., Hu, H., Ye, Q., Li, H., Wang, L., & Xu, J. (2024). Membership inference attacks and defenses in federated learning: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2412.06157
26. Chen, Y., Qin, X., Wang, J., Yu, C., & Gao, W. (2020). FedHealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems, 35*(4), 83–93. https://doi.org/10.1109/MIS.2020.2988604
27. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749
28. MARASANI, Y. (2024). Enterprise Readiness for Generative AI: The Critical Role of Data Engineering. Frontiers in Computer Science and Artificial Intelligence, 3(2), 59-71.
29. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083
30. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Van Overveldt, T., Petrou, D., Ramage, D., & Roselander, J. (2019). Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems, 1*, 374–388.
31. ALAMPALLY, J. (2024). Real-Time and Near-Real-Time Analytics in Healthcare Data Ecosystems. Journal of Computer Science and Technology Studies, 6(1), 314-324.
32. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR.
33. Mukherjee, C. Ai-Driven Personalization of Power System Learning Modules Using Student Personas based on Behavioral Analysis of Grid Performance.
34. Nadia, N. Y., Rabby, H. R., Arif, M. H., Tanvir, M. I. M., Ahmed, M., & Firdaus, S. (2025, October). Scalable RNN-Based Transfer Learning for Patient Sentiment Monitoring in Telehealth Platforms. In 2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS) (pp. 1-6). IEEE.
35. Takon, A. (2025). Explainable AI for Threat Modelling and Decision Support in Engineering Assets. Journal of Cyber-Physical Security and Robotics, 1(02), 46-52.
36. Mukherjee, C. (2025). Combating digital media piracy with agentic ai: Leveraging video transcription and character recognition for automated enforcement. Authorea Preprints.
37. Anifowose, K. (2025). Development and Validation of AI-Assisted Analytical Methods for Biochemical Compound Detection in Pharmaceutical Chemistry. Journal of Applied Pharmaceutical Sciences and Research, 8(4), 41-52.
38. Mukherjee, C. (2025). Use of Agentic AI with OpenAI and Prompt Engineering and State-of-the Art Machine Learning Algorithm to detect the patterns in IOT Device Network Intrusion Attacks. Authorea Preprints.
39. Ravikumar, V. (2025). Therapeutic Bot: Ethical Concerns in AI therapy for Neurodivergence. J Int Scient Re Rep.
40. Mukherjee, C. (2025). Use of Agentic AI with LLM and Prompt Engineering and State-of-the Art Machine Learning Algorithm to detect the patterns in IOT Device Network Intrusion Attacks. TechRxiv. August, 6.
41. Takon, A. (2025). 3D Object Detection and Localization for Industrial Threat Monitoring. Well Testing Journal, 34(S3), 850-880.
42. Mukherjee, C. (2025). Harnessing large language models and ai agents for child behavior analytics in day care: a proof of concept for next-generation parental insight using simulated data. Machinery and Production Engineering, 174(2870), 26-34.
43. Mukherjee, C. (2025). Combating digital media piracy with agentic ai: Leveraging video transcription and character recognition for automated enforcement. Authorea Preprints.
44. Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., & Imine, A. (2023). Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University* -

Computer and Information Sciences, 35(1), 145–174. https://doi.org/10.1016/j.jksuci.2022.11.008

45. Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. *Expert Systems with Applications, 240*, 122156. https://doi.org/10.1016/j.eswa.2023.122156
46. MARASANI, Y. (2023). Machine Learning Models for Predicting Patient Treatment Switching Using Claims Data. Frontiers in Computer Science and Artificial Intelligence, 2(1), 59-66.
47. Gao, H., Kou, G., Liang, H., Zhang, H., Chao, X., Li, C.-C., & Dong, Y. (2024). Machine learning in business and finance: A literature review and research opportunities. *Financial Innovation, 10*, 35. https://doi.org/10.1186/s40854-023-00550-z
48. Vanini, P., Rossi, S., Zvizdic, E., & Domenig, T. (2023). Online payment fraud: From anomaly detection to risk management. *Financial Innovation, 9*, 66. https://doi.org/10.1186/s40854-023-00470-y
49. Chen, C., Lee, C., Huang, S., & Peng, W. (2024). Credit card fraud detection via intelligent sampling and self-supervised learning. *ACM Transactions on Intelligent Systems and Technology, 15*(2), 1–29. https://doi.org/10.1145/3636515
50. Seera, M., Lim, C. P., Kumar, A., Dhamotharan, L., & Tan, K. H. (2024). An intelligent payment card fraud detection system. *Annals of Operations Research, 334*, 445–467. https://doi.org/10.1007/s10479-021-04149-2