International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 08 Augustl-2026 | Impact Factor: 3.5

SecureFedShield: An Adaptive Privacy-Preserving Federated Defense Framework Against Adversarial Attacks in Financial Fraud Detection

Kriti Mishra

[1] University of Mumbai, Mumbai, India

How to Cite this Article:

Mishra, K. (2026). SecureFedShield: An Adaptive Privacy-Preserving Federated Defense Framework Against Adversarial Attacks in Financial Fraud Detection. International Journal of Creative and Open Research in Engineering and Management, 02(8), 1-9. https://doi.org/10.55041/ijcope.v2i8.059

https://doi.org/10.55041/ijcope.v2i8.059

Abstract

The rapid digital transformation of financial services has significantly increased the volume and complexity of electronic transactions, making automated fraud detection an essential component of modern banking systems. Machine learning (ML) techniques have demonstrated remarkable success in identifying fraudulent activities by learning complex transaction patterns from historical financial data. However, conventional centralized machine learning approaches require organizations to consolidate sensitive customer information into centralized repositories, increasing the risk of privacy breaches, unauthorized access, and regulatory non-compliance with frameworks such as the General Data Protection Regulation (GDPR) and other financial data protection standards [2, 4].

Federated Learning (FL) has emerged as a promising distributed learning paradigm that enables multiple organizations to collaboratively train machine learning models without exchanging raw data [1,4]. Although FL significantly improves data privacy, recent studies have demonstrated that federated learning remains vulnerable to adversarial attacks, including model poisoning, data poisoning, backdoor attacks, membership inference, and gradient inversion attacks, all of which can compromise model integrity and reveal confidential information [7,13].

This paper proposes SecureFedShield, a privacy-preserving federated learning framework designed for secure financial fraud detection in adversarial environments. The proposed framework integrates adaptive privacy protection, trust-aware client evaluation, adversarial update detection, and robust model aggregation into a unified architecture. By combining these complementary mechanisms, SecureFedShield aims to improve resilience against malicious participants while preserving high fraud detection accuracy. The framework is intended to be evaluated using publicly available financial fraud datasets and compared with state-of-the-art federated learning aggregation methods. The proposed architecture provides a practical foundation for deploying secure collaborative machine learning in privacy-sensitive financial institutions.

Keywords— Federated Learning, Financial Fraud Detection, Differential Privacy, Adversarial Machine Learning, Privacy Preservation, Trust Management, Robust Aggregation

1. Introduction

The financial sector has experienced rapid digitalization over the past decade, with online banking, mobile payments, digital wallets, and electronic commerce becoming integral components of the global economy. While these technologies have improved accessibility and operational efficiency, they have also created new opportunities for cybercriminals to conduct increasingly sophisticated financial fraud. Consequently, financial institutions rely heavily on Artificial Intelligence (AI) and Machine Learning (ML) techniques to automatically identify fraudulent transactions, assess credit

risk, detect money laundering, and support regulatory compliance [21,22].

Traditional machine learning models require centralized access to large-scale transaction datasets collected from multiple organizations. Although centralized learning often achieves high predictive performance, it introduces significant privacy and security concerns because sensitive customer information must be transferred and stored in centralized databases. Such repositories become attractive targets for cyberattacks and increase the likelihood of data leakage and regulatory violations [2], [3]. Furthermore, financial organizations are frequently prohibited from sharing customer information because of legal and regulatory requirements, limiting opportunities for collaborative fraud detection.

Federated Learning (FL) has emerged as an alternative distributed learning paradigm that enables multiple organizations to collaboratively train a shared machine learning model while retaining customer data within their local infrastructure [1], [4]. Instead of exchanging raw financial records, participating institutions periodically transmit model updates to a central server, which aggregates these updates to construct an improved global model. This decentralized learning process substantially reduces the need for direct data sharing while enabling knowledge transfer across multiple organizations.

Despite its advantages, federated learning is not inherently secure. Recent research has demonstrated that malicious participants can manipulate local model updates through model poisoning and data poisoning attacks, causing significant degradation of the global model [11]–[13]. Moreover, privacy inference techniques, including membership inference and gradient inversion attacks, have shown that sensitive information can sometimes be reconstructed from exchanged model gradients despite the absence of direct data sharing [7]–[10]. These findings indicate that federated learning alone cannot fully guarantee privacy or robustness in adversarial environments.

Several defense mechanisms have been proposed to address these challenges. Differential Privacy introduces controlled perturbations into model updates to reduce information leakage [2], [3], while robust aggregation algorithms such as Krum, Multi-Krum, Trimmed Mean, and Bulyan improve resilience against malicious updates
[11]–[13]. However, most existing approaches address privacy preservation and adversarial robustness independently. Few studies integrate adaptive privacy protection, client trust evaluation, adversarial detection, and robust aggregation into a single framework specifically designed for financial fraud detection.

To address this research gap, this paper proposes SecureFedShield, an adaptive federated learning framework that combines privacy preservation with trust-aware collaborative learning. The framework introduces adaptive differential privacy, dynamic client trust assessment, gradient similarity-based adversarial detection, and hybrid robust aggregation to strengthen the security of collaborative financial machine learning while maintaining high predictive performance.

The major contributions of this work are summarized as follows:

1. An adaptive privacy-preserving framework that dynamically adjusts privacy protection according to client reliability.
2. A trust-based client evaluation mechanism that continuously monitors participant behavior throughout federated training.
3. A gradient similarity-based adversarial detection module capable of identifying suspicious model updates before aggregation.
4. A hybrid robust aggregation strategy that combines statistical robustness with trust-aware weighting to reduce the influence of malicious participants.

5. A unified security architecture that integrates privacy preservation, adversarial robustness, and collaborative financial fraud detection within a single federated learning framework.

The remainder of this paper is organized as follows. Section 2 reviews existing literature on privacy-preserving machine learning, federated learning, and adversarial attacks. Section 3 presents the proposed SecureFedShield framework. Section 4 describes the experimental methodology, while Section 5 discusses the experimental results. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

2.1 Privacy-Preserving Machine Learning

Protecting sensitive financial information has become one of the primary challenges in modern machine learning systems. Conventional machine learning algorithms typically require centralized storage of training data, increasing the risk of unauthorized access, insider attacks, and large-scale data breaches. To mitigate these risks, researchers have proposed several privacy-preserving learning techniques, including Differential Privacy (DP), Secure Multi-Party Computation (SMPC), Homomorphic Encryption (HE), and Federated Learning (FL) [2,5].

Differential Privacy provides mathematically provable privacy guarantees by introducing carefully calibrated noise into model training or query responses, thereby limiting the influence of individual training records on the final model [2], [3]. Although differential privacy effectively reduces information leakage, excessive noise may reduce model accuracy, particularly in highly imbalanced financial fraud datasets. Consequently, recent studies have explored adaptive privacy mechanisms that seek to balance privacy protection with predictive performance [4].

Federated Learning complements differential privacy by enabling collaborative model training without transferring raw customer information between organizations [1,4]. Nevertheless, research has shown that exchanging gradients alone does not completely eliminate privacy risks because gradient inversion and membership inference attacks may still recover sensitive information from communicated model updates [7,10]. These findings highlight the need for integrated frameworks that simultaneously address privacy preservation and adversarial robustness.

2.2 Federated Learning for Financial Fraud Detection

Federated Learning (FL) has emerged as one of the most promising distributed machine learning paradigms for developing collaborative artificial intelligence systems while preserving data privacy. Unlike conventional centralized machine learning approaches, federated learning enables multiple organizations to jointly train a global model without transferring their local datasets to a centralized repository. Instead, participating institutions perform model training independently using their own private data and share only model parameters or gradients with a central aggregation server [1,4,26].

The decentralized nature of federated learning makes it particularly attractive for the financial sector, where customer transaction records contain highly sensitive personal and financial information. Banks, insurance companies, payment gateways, and credit card providers often possess valuable fraud-related datasets; however, legal restrictions, privacy regulations, and competitive concerns prevent direct sharing of customer information. Federated learning addresses this challenge by allowing organizations to collaboratively improve fraud detection models while ensuring that raw transaction data remain within local institutional boundaries [4, 26, 27].

A typical federated learning workflow consists of four sequential stages. Initially, the central server initializes a global machine learning model and distributes it to all participating financial institutions. Each institution independently trains the model using its local transaction dataset and computes updated model parameters. Rather than transmitting customer records, only the locally trained model updates are communicated back to the aggregation server. Finally, the server combines these updates using an aggregation algorithm to generate an improved global model, which is redistributed for the next communication round [1,4].

One of the earliest and most widely adopted aggregation algorithms is Federated Averaging (FedAvg)**,** introduced by McMahan et al. [1]. FedAvg aggregates local model updates by computing their weighted average according to the number of training samples contributed by each participant. Owing to its computational simplicity and communication efficiency, FedAvg has become the foundation for numerous federated learning applications across healthcare, finance, cybersecurity, and intelligent transportation systems [4,26]. However, FedAvg assumes that all participating clients behave honestly and contribute reliable model updates. This assumption rarely holds in practical financial environments, where compromised institutions or malicious insiders may intentionally manipulate local updates to influence the global model.

To improve robustness against malicious participants, several alternative aggregation algorithms have been proposed**.** Krum selects a client update that is closest to the majority of participating clients, thereby reducing the influence of abnormal gradients [11]. Multi-Krum extends this concept by selecting multiple reliable updates instead of a single client, resulting in improved stability under Byzantine failures [12]. Trimmed Mean removes extreme parameter values before aggregation to reduce the effect of outliers, whereas Bulyan combines multiple robust aggregation strategies to further improve resistance against coordinated poisoning attacks [13]. Although these approaches demonstrate improved robustness compared with FedAvg, they primarily rely on statistical characteristics observed during individual communication rounds and do not incorporate long-term client reliability or adaptive trust evaluation.

Recent studies have also investigated integrating differential privacy, secure aggregation, and cryptographic techniques into federated learning systems. Bonawitz et al. [5] proposed a practical secure aggregation protocol that encrypts model updates during communication, preventing the aggregation server from observing individual client parameters. Similarly, Kairouz et al. [4] identified secure aggregation, communication efficiency, client heterogeneity, and adversarial robustness as major research challenges limiting the large-scale deployment of federated learning. Despite these advances, balancing privacy preservation, model utility, communication efficiency, and adversarial resilience remains an open research problem.

Table 2.1 summarizes representative federated learning aggregation methods and their primary characteristics.

Table 2.1 Comparison of Representative Federated Learning Aggregation Methods

Method	**Robust**	**Privacy**	**Complexity**	**Limitation**
FedAvg [1]	No	No	Low	Vulnerable to malicious updates
Krum [11]	Partial	No	Medium	Limited scalability
Multi-Krum [12]	Yes	No	High	Increased computational overhead
Trimmed Mean	Yes	No	Medium	Sensitive to parameter selection

Although these aggregation algorithms contribute significantly to improving collaborative machine learning, none simultaneously provide adaptive

privacy preservation, continuous client trust assessment, adversarial detection, and robust aggregation within a unified framework. Most existing approaches focus on a single aspect of secure federated learning, leaving practical financial systems vulnerable to sophisticated adversarial attacks. This limitation motivates the development of the proposed SecureFedShield framework, which integrates these complementary mechanisms into a single adaptive architecture for secure financial fraud detection.

2.3 Adversarial Attacks in Federated Learning

Despite its privacy advantages, federated learning remains susceptible to several adversarial attacks that can compromise both model performance and customer confidentiality. Since model updates are exchanged repeatedly during collaborative training, attackers can exploit this communication process to manipulate the global model or infer sensitive information from shared parameters [7,13].

One of the most common threats is model poisoning, in which malicious participants deliberately transmit manipulated model updates to influence the global model. Rather than modifying local datasets, attackers alter gradients or model parameters before transmission to reduce fraud detection accuracy or introduce targeted classification errors. Because traditional aggregation algorithms often assume honest participants, even a small number of malicious clients can significantly affect the global model [11], [12].

Another important category is data poisoning, where attackers manipulate local training datasets before model training. Financial transaction labels may be intentionally modified, fraudulent transactions may be disguised as legitimate, or fabricated records may be inserted into local datasets. As these corrupted datasets are used during local training, the resulting model updates gradually degrade the effectiveness of the collaboratively learned fraud detection model [13].

Researchers have also demonstrated the effectiveness of backdoor attacks, in which malicious clients train local models to associate specific trigger patterns with incorrect predictions while maintaining high overall classification accuracy. Because the global model continues to perform well on normal transactions, these attacks often remain undetected during conventional evaluation procedures [13].

In addition to integrity attacks, federated learning also faces privacy inference attacks. Membership inference attacks attempt to determine whether a particular transaction was used during model training, whereas gradient inversion attacks reconstruct sensitive input features directly from exchanged gradients [7,10]. These attacks demonstrate that sharing model parameters alone does not completely eliminate privacy risks, emphasizing the importance of integrating additional privacy-preserving mechanisms into federated learning systems.

The increasing sophistication of adversarial attacks indicates that existing federated learning architectures require stronger security mechanisms capable of simultaneously protecting customer privacy, detecting malicious participants, and maintaining model robustness. These challenges motivate the research gap discussed in the following section.

2.4 Research Gap

The literature review indicates that significant progress has been made in privacy-preserving machine learning and federated learning; however, several critical challenges remain unresolved, particularly for financial fraud detection systems. Existing research has primarily focused on addressing individual aspects of secure federated learning, such as privacy preservation, robust aggregation, or adversarial defense, rather than providing a

unified framework capable of simultaneously addressing these challenges [4,26].

Differential Privacy has become one of the most widely adopted techniques for protecting sensitive information during collaborative model training by injecting calibrated noise into model updates [2,3]. Although this approach offers formal privacy guarantees, the use of a fixed privacy budget often degrades model utility, especially in highly imbalanced financial datasets where preserving subtle fraud patterns is essential for accurate classification. Moreover, most differential privacy mechanisms do not distinguish between trustworthy and potentially malicious participants, resulting in uniform privacy protection regardless of client behavior [4].

Similarly, robust aggregation algorithms such as Krum, Multi-Krum, Trimmed Mean, and Bulyan have improved the resilience of federated learning against Byzantine failures and model poisoning attacks [11]–[13]. Nevertheless, these methods primarily rely on statistical analysis of model updates during a single communication round and assume that abnormal gradients always originate from malicious clients. Such assumptions may not hold in practical financial environments where legitimate clients may generate heterogeneous updates because of non-independent and identically distributed (Non-IID) transaction data, varying transaction volumes, or institution-specific fraud patterns [4], [26]. Consequently, relying solely on statistical filtering may reduce the contribution of honest participants and affect overall model convergence.

Recent research has also investigated secure aggregation protocols and cryptographic techniques to prevent disclosure of individual model updates during communication [5]. While these methods effectively protect transmitted parameters from unauthorized access, they do not prevent adversarial participants from submitting manipulated updates that intentionally degrade the global model. Likewise, privacy-preserving mechanisms alone cannot defend against sophisticated attacks such as model poisoning, backdoor insertion, or collusion among multiple malicious clients [7,13].

Another important limitation identified in existing studies is the absence of continuous client reliability assessment. Most federated learning algorithms evaluate model updates independently in each communication round without considering historical participant behavior. In real-world financial networks, however, client behavior may evolve over time because of compromised systems, insider threats, or gradually executed poisoning attacks. The lack of adaptive trust management limits the ability of existing approaches to identify stealthy adversaries whose updates remain statistically similar to those of legitimate participants [4], [13].

Furthermore, relatively few studies have proposed integrated security architectures specifically designed for financial fraud detection. Financial datasets are characterized by severe class imbalance, rapidly evolving fraud patterns, heterogeneous institutional data distributions, and strict regulatory requirements governing customer privacy [21], [22]. These domain-specific characteristics require security mechanisms that simultaneously preserve privacy, maintain detection accuracy, and withstand adversarial manipulation. Existing federated learning frameworks generally address these requirements independently, resulting in trade-offs between privacy protection, robustness, and predictive performance.

Table 2.2 summarizes the major limitations identified in representative federated learning approaches.

Table 2.2 Research Gap Analysis

Existing Approach	Primary Strength	Major Limitation
Differential Privacy	Strong privacy guarantees	Performance degradation

The proposed framework presented in the next section is designed to address these research gaps and provide a comprehensive solution for secure federated learning in financial applications.

		due to fixed privacy budget
FedAvg	Communication efficiency	Highly vulnerable to poisoning attacks
Krum / Multi-Krum	Byzantine robustness	No adaptive trust evaluation
Trimmed Mean / Bulyan	Robust statistical aggregation	Limited consideration of historical client behavior
Secure Aggregation	Confidential communication	Cannot detect malicious model updates
Existing Financial FL Models	Collaborative fraud detection	Lack of integrated privacy, trust management, and adversarial defense

Based on these observations, there remains a clear need for a federated learning framework that integrates adaptive privacy preservation, continuous client trust assessment, adversarial update detection, and robust aggregation within a unified architecture. Motivated by these limitations, this paper proposes SecureFedShield, an adaptive security framework that combines these complementary mechanisms to enhance the reliability, privacy, and robustness of collaborative financial fraud detection systems. Unlike existing approaches, SecureFedShield continuously evaluates participant trustworthiness, dynamically adapts privacy protection, detects suspicious model updates before aggregation, and employs trust-aware robust aggregation to reduce the influence of malicious participants while maintaining high predictive performance.

3. Proposed SecureFedShield Framework

3.1 Framework Overview

The limitations identified in the previous section demonstrate that existing federated learning approaches independently address privacy preservation, robust aggregation, or adversarial defense but rarely integrate these capabilities into a unified framework suitable for financial fraud detection. To overcome these challenges, this paper proposes SecureFedShield, an adaptive privacy-preserving federated learning framework that combines trust-aware client evaluation, adversarial update detection, adaptive privacy protection, and robust aggregation into a single collaborative learning architecture.

The primary objective of SecureFedShield is to enable multiple financial institutions to collaboratively train an accurate fraud detection model without sharing sensitive customer information while maintaining resilience against malicious participants. Unlike traditional centralized learning, where all transaction records are collected into a single repository, SecureFedShield ensures that customer data remain within the local infrastructure of each participating institution throughout the entire training process. Only privacy-protected model updates are exchanged with the aggregation server, thereby reducing the risk of sensitive information disclosure and supporting compliance with financial data protection regulations [1], [2], [4].

The proposed framework consists of five functional components, each addressing a specific security challenge encountered during collaborative learning.

The first component is the Local Model Training Module, where every participating financial

institution independently trains a fraud detection model using its own transaction dataset. Since customer records never leave the organization's infrastructure, the likelihood of centralized data breaches and unauthorized disclosure is significantly reduced. Local model training also enables organizations to preserve data ownership while benefiting from collaborative intelligence generated across multiple institutions [1], [4].

The second component is the Adaptive Privacy Protection Module. Before locally trained model parameters are transmitted to the aggregation server, a privacy-preserving mechanism is applied to minimize the possibility of information leakage through communicated model updates. Unlike conventional differential privacy approaches that employ identical privacy settings for every participant, SecureFedShield adapts the level of privacy protection according to the reliability of participating clients. This adaptive strategy seeks to preserve model utility while maintaining strong privacy guarantees throughout federated training [2]–[4].

The third component is the Trust-Based Client Evaluation Module. Every client participating in federated learning is continuously monitored throughout successive communication rounds. Instead of evaluating model updates independently during each iteration, the framework maintains a dynamic trust profile that reflects historical participation, update consistency, and communication reliability. Institutions consistently producing reliable model updates gradually receive higher trust levels, whereas suspicious participants experience a reduction in influence during subsequent aggregation rounds. This continuous trust management enables the framework to identify gradually evolving attacks that may remain undetected by conventional statistical filtering techniques [4], [11]–[13].

The fourth component is the Adversarial Update Detection Module, which analyzes incoming client updates before they are incorporated into the global model. The objective of this module is to identify manipulated gradients generated through model poisoning, data poisoning, or coordinated adversarial behavior. Instead of relying solely on statistical outlier removal, the proposed framework evaluates update consistency using multiple complementary indicators, allowing suspicious updates to be identified before they affect collaborative model learning. This additional verification stage improves the resilience of the global model against sophisticated adversarial attacks reported in recent federated learning literature [7]–[13].

The final component is the Trust-Aware Robust Aggregation Module. Existing aggregation algorithms such as FedAvg perform weighted averaging of all received model updates without considering participant reliability, whereas robust aggregation methods such as Krum and Bulyan primarily rely on statistical characteristics observed during individual communication rounds [11]–[13]. SecureFedShield extends these approaches by incorporating continuously updated trust information into the aggregation process. Consequently, reliable participants contribute more significantly to the construction of the global fraud detection model, while the influence of suspicious clients is progressively reduced. This adaptive aggregation strategy improves model robustness without unnecessarily excluding legitimate participants that may exhibit natural data heterogeneity.

The operational workflow of the framework is summarized as follows:

1. The central server initializes the global fraud detection model and distributes it to all participating financial institutions.
2. Each institution independently trains the model using its local transaction data.
3. Privacy-preserving mechanisms are applied to local model updates before communication.
4. The aggregation server evaluates the reliability of participating clients using historical and behavioral information.

5. Incoming updates are analyzed to identify potential adversarial manipulation.
6. Trust-aware robust aggregation combines reliable model updates to generate an improved global model.
7. The updated global model is redistributed to participating institutions for the next communication round.
8. The collaborative learning process continues until the model converges or predefined stopping criteria are satisfied.

The layered design of SecureFedShield enables the framework to address three critical requirements of modern financial fraud detection systems simultaneously: customer privacy, adversarial robustness, and high predictive **performance**. By integrating adaptive privacy preservation with continuous trust evaluation and robust aggregation, the proposed framework overcomes several limitations identified in existing federated learning approaches. Consequently, SecureFedShield provides a practical and scalable architecture for secure collaborative machine learning in privacy-sensitive financial environments.

3.2 Novel Features of the Proposed Frameork

The proposed framework introduces several novel characteristics that distinguish it from existing federated learning approaches.

First, the framework integrates multiple security mechanisms into a unified architecture rather than addressing privacy preservation and adversarial robustness as independent problems. This holistic design enables SecureFedShield to provide comprehensive protection throughout the entire collaborative learning process.

Second, SecureFedShield employs adaptive privacy preservation, allowing privacy protection to evolve according to participant reliability instead of relying on static privacy settings. This adaptive approach aims to reduce unnecessary degradation of model accuracy while maintaining strong protection against information leakage.

Third, the framework continuously evaluates participant behavior through a dynamic trust management mechanism. By incorporating historical reliability into aggregation decisions, SecureFedShield is capable of identifying gradually evolving malicious behavior that may remain undetected using traditional statistical methods.

Fourth, the proposed framework introduces a dedicated adversarial update verification stage before model aggregation. This additional security layer minimizes the possibility that manipulated model updates influence the collaboratively learned fraud detection model.

Finally, SecureFedShield combines trust-aware decision making with robust aggregation, enabling reliable financial institutions to contribute more effectively to global model optimization while progressively reducing the influence of suspicious participants. This integrated strategy improves both the robustness and stability of collaborative financial fraud detection under adversarial conditions.

3.3 Adaptive Privacy Protection Module

Protecting sensitive financial information is one of the primary objectives of the proposed SecureFedShield framework. Although federated learning eliminates the need to exchange raw customer transaction records, recent studies have demonstrated that model parameters and gradients may still reveal confidential information through gradient inversion, membership inference, and reconstruction attacks
[7]–[10]. Consequently, additional privacy-preserving mechanisms are required before model updates are transmitted to the central aggregation server.

In SecureFedShield, an Adaptive Privacy Protection Module (APPM**)** is incorporated

between the local training stage and the communication stage. The objective of this module is to reduce information leakage while preserving the quality of model updates contributed by participating institutions. Unlike conventional privacy-preserving approaches that apply identical protection to every client, the proposed framework adapts the degree of privacy preservation according to the reliability of each participating institution. This adaptive strategy is intended to balance two competing objectives: maintaining strong customer privacy and preserving model accuracy during collaborative learning [2]–[4].

The adaptive privacy mechanism operates immediately after local model training is completed. Before transmitting model parameters to the aggregation server, each participating institution applies privacy protection to its locally trained model update. The protected update is then securely transmitted to the aggregation server instead of the original model parameters. Since only privacy-preserving updates are exchanged, the probability of exposing sensitive financial information through communication is significantly reduced.

Another important characteristic of the proposed module is its ability to account for the dynamic behavior of participating clients. During successive communication rounds, institutions demonstrating consistent and trustworthy behavior are allowed to contribute model updates with relatively lower perturbation, thereby preserving useful learning information. Conversely, updates received from suspicious or unreliable participants undergo stronger privacy protection, limiting their ability to reveal sensitive information or adversely influence collaborative learning. This adaptive behavior distinguishes SecureFedShield from conventional differential privacy mechanisms that rely on fixed privacy configurations regardless of client reliability [3], [4].

The Adaptive Privacy Protection Module also supports compliance with modern financial data protection regulations. Banking organizations are required to protect customer information while simultaneously enabling secure data-driven decision making. By ensuring that sensitive transaction records remain within institutional boundaries and only privacy-preserving model updates are exchanged, the proposed framework aligns with the principles of data minimization and confidentiality emphasized by regulatory frameworks such as GDPR and similar financial privacy regulations [2], [4].

Overall, the Adaptive Privacy Protection Module represents the first layer of defense in the SecureFedShield framework. It minimizes information leakage during communication while preserving the utility of collaborative model learning, thereby providing a secure foundation for subsequent trust evaluation and adversarial detection.

3.4 Trust-Based Client Evaluation Module

One of the fundamental assumptions made by many existing federated learning algorithms is that participating clients behave honestly throughout the training process. In practical financial environments, however, participating institutions may become compromised because of insider attacks, malware infections, system vulnerabilities, or intentionally malicious behavior. As a result, treating all clients equally during model aggregation can significantly reduce the robustness of the global fraud detection model [11]–[13].

To address this limitation, SecureFedShield introduces a Trust-Based Client Evaluation Module (TCEM**)** that continuously assesses the reliability of every participating institution throughout federated training. Instead of evaluating model updates independently during each communication round, the proposed framework maintains a dynamic trust profile that evolves according to the historical behavior of each participant.

Following every communication round, the aggregation server analyzes the received model updates and evaluates their consistency with the overall collaborative learning process. Clients that repeatedly submit reliable and consistent updates gradually receive higher trust scores, whereas institutions exhibiting abnormal or suspicious behavior experience a progressive reduction in trust. This continuous evaluation allows the framework to distinguish between temporary deviations caused by data heterogeneity and persistent malicious behavior resulting from adversarial attacks.

The Trust-Based Client Evaluation Module considers multiple characteristics when assessing participant reliability. These include the consistency of model updates across communication rounds, similarity between local and global learning behavior, historical participation records, and the frequency of abnormal updates detected during training. By combining these indicators, the framework provides a more comprehensive assessment of client reliability than conventional aggregation algorithms that rely solely on statistical analysis of current model updates [4], [11].

Unlike traditional Byzantine-robust aggregation methods that immediately discard suspicious updates, SecureFedShield adopts a progressive trust management strategy. Institutions identified as potentially unreliable are not permanently removed from the collaborative learning process after a single abnormal update. Instead, their contribution to the global model is gradually reduced while their future behavior continues to be monitored. This design minimizes false-positive decisions and improves model stability, particularly in financial environments characterized by heterogeneous transaction distributions and naturally varying local model updates.

The Trust-Based Client Evaluation Module also enables the proposed framework to defend against stealthy adversarial attacks. Since trust is continuously updated over multiple communication rounds, malicious participants attempting gradual model manipulation become easier to identify than in conventional approaches that evaluate only instantaneous statistical anomalies. Consequently, historical behavioral analysis enhances the overall resilience of collaborative learning against long-term poisoning strategies reported in recent federated learning literature [12], [13].

By integrating continuous trust assessment into the collaborative learning process, SecureFedShield improves both aggregation reliability and adversarial resilience. The computed trust profiles are subsequently utilized by the robust aggregation module to determine the influence of individual participants during global model construction, thereby strengthening the overall security of the proposed framework.

3.5 Adversarial Update Detection Module

Although federated learning enables collaborative model training without sharing raw financial data, it remains vulnerable to malicious participants that intentionally manipulate model updates before transmission to the aggregation server. Such adversarial behavior can significantly degrade the performance of the global fraud detection model or implant hidden backdoors that remain undetected during normal operation [7]–[13]. Consequently, identifying malicious updates before model aggregation is essential for maintaining the reliability of collaborative financial learning systems.

To address this challenge, SecureFedShield incorporates an Adversarial Update Detection Module (AUDM**)** that analyzes every client update before it is included in the global aggregation process. The objective of this module is to identify suspicious model updates while minimizing the rejection of legitimate updates generated from heterogeneous financial datasets.

Unlike conventional approaches that rely solely on statistical outlier detection, the proposed module performs a multi-stage evaluation of

received model updates. During each communication round, the aggregation server compares incoming updates with the expected learning behavior of the global model and evaluates their consistency with updates received from other participating institutions. Multiple characteristics, including update consistency, directional similarity, historical client reliability, and abnormal behavioral patterns, are jointly considered during the evaluation process. This comprehensive assessment enables the framework to distinguish naturally heterogeneous updates from deliberately manipulated ones [11]–[13].

The proposed detection module is designed to identify several categories of adversarial attacks commonly reported in federated learning literature. These include model poisoning attacks**,** where malicious participants intentionally modify model parameters before communication; data poisoning attacks, in which corrupted local datasets generate misleading model updates; and backdoor attacks, where attackers embed hidden triggers into locally trained models while maintaining acceptable performance on normal samples [8], [11], [13]. The framework also considers privacy inference attacks such as gradient inversion and membership inference, which attempt to extract confidential information from exchanged model updates [7]–[10].

Once an abnormal update is detected, the framework does not immediately remove the corresponding client from the collaborative learning process. Instead, the detected anomaly is incorporated into the client's trust profile, allowing future aggregation decisions to consider both current and historical behavior. This progressive strategy reduces the likelihood of falsely identifying honest participants as malicious, particularly in financial environments characterized by non-independent and identically distributed (Non-IID) transaction data [4], [26].

The output of the Adversarial Update Detection Module is subsequently forwarded to the Trust-Aware Aggregation Module, where detected anomalies influence the contribution of individual participants during global model construction. This layered defense mechanism enhances the robustness of SecureFedShield while preserving collaborative learning among legitimate financial institutions.

3.6 Trust-Aware Hybrid Robust Aggregation

Model aggregation is one of the most critical stages in federated learning because the quality of the global model directly depends on the reliability of the received client updates. Conventional aggregation algorithms, such as Federated Averaging (FedAvg), combine all client updates using weighted averaging without evaluating their trustworthiness [1]. Consequently, malicious participants can substantially influence the global model by submitting manipulated updates during collaborative training.

Several robust aggregation algorithms, including Krum, Multi-Krum, Trimmed Mean, and Bulyan, have been proposed to reduce the impact of Byzantine attacks by filtering statistically abnormal model updates [11]–[13]. Although these approaches improve robustness, they primarily evaluate updates within individual communication rounds and do not incorporate long-term participant behavior into aggregation decisions. As a result, sophisticated adversaries capable of gradually manipulating model updates may still influence the learning process.

To overcome these limitations, SecureFedShield introduces a Trust-Aware Hybrid Robust Aggregation (THRA**)** strategy. Instead of relying exclusively on statistical filtering, the proposed aggregation mechanism combines information obtained from the Trust-Based Client Evaluation Module and the Adversarial Update Detection Module. Before aggregation, every participating institution is assigned a reliability score based on its historical contribution, consistency of model updates, and previously detected adversarial behavior. Clients demonstrating stable and reliable participation contribute more

significantly to the construction of the global fraud detection model, whereas suspicious participants receive progressively lower influence during aggregation.

The aggregation procedure follows three sequential stages. First, the trust information generated during previous communication rounds is examined to identify participants exhibiting persistent abnormal behavior. Second, the adversarial detection results obtained during the current communication round are incorporated to evaluate newly received model updates. Finally, only reliable updates are combined to construct the next global model, while suspicious updates receive reduced influence or are excluded when their behavior consistently violates predefined reliability criteria. This progressive decision-making process provides stronger robustness than traditional one-step aggregation algorithms while minimizing unnecessary exclusion of legitimate participants.

Another important advantage of the proposed aggregation strategy is its adaptability to heterogeneous financial datasets. Financial institutions often differ considerably in transaction volume, customer demographics, and fraud patterns, resulting in naturally diverse local model updates. Conventional statistical aggregation methods may incorrectly classify these legitimate variations as malicious behavior. By incorporating historical trust information in addition to current update characteristics, SecureFedShield reduces false-positive detections and maintains stable model convergence even in highly heterogeneous collaborative environments [4], [26].

Compared with existing federated learning approaches, the proposed Trust-Aware Hybrid Robust Aggregation mechanism provides three key advantages. First, it improves resistance against coordinated poisoning attacks by reducing the influence of unreliable participants. Second, it preserves collaborative learning among trustworthy institutions instead of aggressively removing clients after isolated abnormal updates. Third, it enhances the long-term stability of global model training by continuously adapting aggregation decisions according to evolving participant behavior.

Collectively, the Adaptive Privacy Protection Module, Trust-Based Client Evaluation Module, Adversarial Update Detection Module, and Trust-Aware Hybrid Robust Aggregation Module constitute the core components of the proposed SecureFedShield framework. Their integration provides a comprehensive security architecture capable of simultaneously preserving customer privacy, detecting adversarial behavior, and maintaining robust collaborative fraud detection performance in distributed financial environments.

3.7 Advantages of the Proposed SecureFedShield Framework

The proposed SecureFedShield framework offers several practical advantages over conventional federated learning architectures. First, it enables collaborative fraud detection without requiring financial institutions to exchange sensitive customer transaction records, thereby supporting compliance with modern privacy regulations. Second, the adaptive privacy mechanism minimizes information leakage while preserving model utility during collaborative learning. Third, the trust-based evaluation mechanism continuously monitors participant behavior, enabling early identification of compromised or malicious clients. Fourth, the adversarial detection module strengthens the resilience of the framework against poisoning, backdoor, and inference attacks reported in recent literature. Finally, the trust-aware aggregation strategy improves the robustness and stability of global model training by integrating both historical reliability and current behavioral analysis into aggregation decisions.

These integrated capabilities distinguish SecureFedShield from existing federated learning approaches and establish the proposed framework as a comprehensive solution for

secure, privacy-preserving financial fraud detection under adversarial environments.

4. Proposed Methodology

4.1 Research Methodology

The proposed SecureFedShield framework is designed to evaluate the effectiveness of adaptive privacy preservation and adversarial defense in federated financial fraud detection. The experimental methodology follows a federated learning workflow in which multiple financial institutions collaboratively train a fraud detection model without exchanging raw customer transaction data. Each institution independently performs local model training using its own dataset while only privacy-preserving model updates are communicated to the central aggregation server.

The experimental workflow consists of six major stages:

1. **Data Preparation:** Financial transaction datasets are preprocessed by removing missing values, encoding categorical attributes, normalizing numerical features, and addressing class imbalance using appropriate sampling techniques.
2. **Local Model Training:** Each participating client independently trains the fraud detection model using its local dataset without transferring customer information outside the organization.
3. **Adaptive Privacy Protection:** Before communication, local model updates are processed by the Adaptive Privacy Protection Module to reduce the possibility of information leakage during transmission.
4. **Trust Evaluation and Adversarial Detection:** The aggregation server evaluates the reliability of participating clients and analyzes received model updates for suspicious behavior indicative of poisoning or backdoor attacks.
5. **Trust-Aware Robust Aggregation:** Reliable model updates are aggregated using the proposed SecureFedShield aggregation strategy to construct an updated global model.
6. **Performance Evaluation:** The global model is evaluated using multiple classification and security metrics, and the results are compared with existing federated learning approaches.

The overall methodology ensures that privacy preservation, adversarial robustness, and fraud detection performance are evaluated simultaneously under collaborative learning conditions

4.2 Experimental Dataset

The proposed framework is intended to be evaluated using publicly available financial fraud detection datasets that have been extensively used in previous machine learning and federated learning studies.

The PaySim Financial Dataset represents simulated mobile money transactions generated from real financial transaction patterns. The dataset contains multiple transaction types, including cash transfers, payments, cash withdrawals, debit operations, and fraudulent transactions. It has become one of the most widely used benchmark datasets for evaluating fraud detection algorithms because it accurately represents class imbalance commonly observed in real-world financial systems [22].

To further validate the robustness of the proposed framework, experiments may also be conducted using the IEEE-CIS Fraud Detection Dataset, which contains anonymized e-commerce transaction records collected from large-scale financial systems. The dataset includes both transaction-level features and identity-related attributes, providing a realistic benchmark for evaluating fraud detection performance under heterogeneous data distributions [21].

4.3 Experimental Configuration

The experiments are designed to simulate a collaborative federated learning environment involving multiple participating financial institutions. Each institution acts as an independent client possessing its own local transaction dataset.

The experimental configuration includes:

- Multiple federated learning clients representing independent financial organizations.
- A centralized aggregation server responsible for coordinating model updates.
- Non-IID data distribution across participating institutions.
- Secure communication between clients and the aggregation server.
- Iterative global model updates until convergence.

To evaluate adversarial robustness, selected clients are assumed to behave maliciously by introducing poisoned model updates during specific communication rounds. This configuration enables the proposed SecureFedShield framework to be evaluated under realistic adversarial scenarios reported in recent federated learning research [11]–[13].

4.4 Performance Evaluation Metrics

The effectiveness of SecureFedShield is evaluated using both classification and security-related performance metrics.

Classification Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

These metrics measure the capability of the proposed framework to correctly identify fraudulent financial transactions while minimizing false alarms [21], [22].

Security Metrics

To evaluate adversarial robustness and privacy preservation, the following security metrics are considered:

- Attack Success Rate (ASR)
- Model Robustness under Poisoning Attacks
- Communication Overhead
- Training Convergence Rate
- Privacy Leakage Resistance

Together, these metrics provide a comprehensive assessment of both predictive performance and system security.

5. Expected Results and Discussion

The proposed SecureFedShield framework is expected to achieve improved fraud detection performance while maintaining stronger privacy guarantees than conventional federated learning algorithms. By integrating adaptive privacy protection with trust-aware aggregation and adversarial update detection, the framework is anticipated to reduce the impact of malicious participants without significantly affecting model convergence.

Compared with FedAvg, the proposed framework is expected to demonstrate higher robustness against poisoning attacks because suspicious model updates receive reduced influence during aggregation. Similarly, compared with Krum, Multi-Krum, Trimmed Mean, and Bulyan, SecureFedShield is expected to achieve more stable learning by incorporating historical client reliability rather than relying solely on statistical filtering [11]–[13].

The adaptive privacy mechanism is also expected to preserve model utility more effectively than

conventional fixed-budget differential privacy techniques by reducing unnecessary perturbation for trustworthy participants while maintaining stronger protection for potentially unreliable clients [2]–[4].

Table 5.1 presents the anticipated qualitative comparison between the proposed framework and representative federated learning approaches.

Table 5.1 Expected Performance Comparison

Method	Priva cy	Poisoni ng Resistan ce	Trust Evaluati on	Adapti ve Privac y
FedAvg	Low	Low	No	No
Krum	Low	Moderat e	No	No
Multi-Krum	Low	High	No	No
Bulyan	Low	High	No	No
SecureFedSh ield (Proposed)	High	High	Yes	Yes

The expected outcomes indicate that integrating adaptive privacy preservation, continuous trust management, adversarial detection, and robust aggregation can substantially improve the security and reliability of collaborative financial fraud detection systems.

6. Conclusion and Future Work

This paper presented SecureFedShield, an adaptive privacy-preserving federated learning framework for secure financial fraud detection under adversarial environments. The proposed framework integrates adaptive privacy protection, continuous trust-based client evaluation, adversarial update detection, and trust-aware robust aggregation into a unified collaborative learning architecture. Unlike conventional federated learning approaches that independently address privacy preservation or adversarial robustness, SecureFedShield combines these complementary mechanisms to enhance both model security and predictive performance.

A comprehensive review of existing literature revealed that current federated learning algorithms remain vulnerable to poisoning attacks, privacy inference attacks, and the absence of continuous client reliability assessment. These limitations motivated the development of SecureFedShield as an integrated solution capable of protecting sensitive financial information while maintaining robust collaborative learning among multiple financial institutions.

The proposed methodology establishes a foundation for evaluating privacy-preserving financial fraud detection using benchmark datasets such as PaySim and IEEE-CIS Fraud Detection. Performance evaluation is planned using standard classification metrics together with security-oriented measures that quantify adversarial robustness and privacy preservation.

Future research will focus on implementing the proposed framework in a real federated learning environment, evaluating its performance against advanced adaptive adversarial attacks, optimizing communication efficiency, and extending the framework to cross-device federated learning scenarios. Additional research may also investigate blockchain-assisted trust management, explainable artificial intelligence techniques, and privacy-preserving federated learning for other financial cybersecurity applications.

References

[1] H. 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 Int. Conf. Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[2] C. Dwork, "Differential Privacy," in Proc. 33rd Int. Colloquium on Automata, Languages and Programming (ICALP), Venice, Italy, 2006, pp. 1–12.

[3] C. Dwork and A. Roth, The Algorithmic Foundations of Differential Privacy. Boston, MA, USA: Now Publishers Inc., 2014.

[4] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends® in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021.

[5] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), Dallas, TX, USA, 2017, pp. 1175–1191.

[6] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), Dallas, TX, USA, 2017, pp. 603–618.

[7] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning," in Proc. IEEE Symp. Security and Privacy (SP), San Francisco, CA, USA, 2019, pp. 739–753.

[8] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 2019.

[9] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting Gradients – How Easy Is It to Break Privacy in Federated Learning?" in Advances in Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 2020.

[10] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning," in Proc. IEEE Symp. Security and Privacy (SP), San Francisco, CA, USA, 2019, pp. 691–706.

[11] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in Advances in Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 2017.

[12] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in Proc. 35th Int. Conf. Machine Learning (ICML), Stockholm, Sweden, 2018, pp. 5650–5659.

[13] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, "The Hidden Vulnerability of Distributed Learning in Byzantium," in Proc. 35th Int. Conf. Machine Learning (ICML), Stockholm, Sweden, 2018, pp. 3521–3530.

[14] B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," Pattern Recognition, vol. 84, pp. 317–331, Dec. 2018.

[15] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in Proc. IEEE Symp. Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 39–57.

[16] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in Proc. Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, 2015.

[17] A. Shamir, "How to Share a Secret," Communications of the ACM, vol. 22, no. 11, pp. 612–613, Nov. 1979.

[18] C. Gentry, A Fully Homomorphic Encryption Scheme. Stanford University, Stanford, CA, USA, Ph.D. dissertation, 2009.

[19] O. Goldreich, S. Micali, and A. Wigderson, "How to Play Any Mental Game," in Proc. 19th ACM Symp. Theory of Computing (STOC), New York, NY, USA, 1987, pp. 218–229.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[21] Vesta Corporation and IEEE Computational Intelligence Society, "IEEE-CIS Fraud Detection Dataset," Kaggle. Available: https://www.kaggle.com/competitions/ieee-fraud-detection

[22] E. A. Lopez-Rojas and S. Axelsson, "PaySim: A Financial Mobile Money Simulator for Fraud Detection Research," in The 28th European Modeling and Simulation Symposium (EMSS), Larnaca, Cyprus, 2016.

[23] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 2017.

[24] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency," arXiv preprint arXiv:1610.05492, 2016.

[25] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60,
May 2020.

[26] 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, Jan. 2019.

[27] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), Denver, CO, USA, 2015, pp. 1310–1321.