Int. J. Advance Soft Compu. Appl, Vol. 18, No. 2, July 2026
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Poisoning-Resilient Homomorphic Secure Aggregation for High-Dimensional Cybersecurity Gradients

Enshirah Altarawneh[1*], Jawdat S. Alkasassbeh[2], Qotadeh Saber Aljawazneh[3], Aws Al-Qaisi[4], Khalid N. AlZubi[5],Mohammad Sharsheer[4]

[1]Department of Computer Engineering, Faculty of Engineering, The Hashemite University, Zarqa, 13133, Jordan
enshirah@hu.edu.jo
[2]Department of Electrical Engineering, Faculty of Engineering Technology, Al-Balqa Applied University, Amman, Jordan
Jawdat1983@bau.edu.jo
[3]Department of Data Science and Artificial Intelligence, Zarqa University, Zarqa, Jordan
Q.aljawazneh@zu.edu.jo
[4]College of Engineering and Technology, American University of the Middle East, Kuwait
aws.al-qaisi@aum.edu.kw; mohammad.sharsheer@aum.edu.kw
[5]Department of Management, Information Systems, Faculty of Business, Al-Balqa Applied University, Salt, Jordan
khalid.zoubai@bau.edu.jo

Abstract

Federated Learning (FL) allows for collaborative model training across several geographically dispersed clients while protecting client data from other users. Although FL is susceptible to poisoning attacks, collusion, and Byzantine adversaries, which can result in decreased performance of the overall model, we present a lightweight FL framework that utilizes homomorphic encryption to protect against poisoning attacks while also providing client data protection and low computational overhead by integrating Additive Homomorphic Encryption (AHE), Trust-Weighted Aggregation (TWA), Anomaly Scoring, and Gradient Clipping. The proposed framework has been extensively simulated with the MNIST and CIFAR-10 datasets. These simulations have demonstrated that the proposed framework can maintain greater than 95% of the overall model's accuracy when the number of malicious clients is as high as 40%. Local epochs are processed at the client side in 15-20 milliseconds; server-side secure aggregation was found to be two times slower than plaintext aggregation; and each round of communication averages approximately 4 MB/client of data transferred. The simulation results provide evidence that the proposed framework will provide robust convergence, scalability, and resilience to a variety of types of adversary behaviors. Existing approaches address either privacy or robustness, but this work proposes a unified framework that simultaneously ensures privacy (via homomorphic encryption) and robustness (via poisoning-resilient aggregation) for applications such as cybersecurity monitoring, IoT, and other privacy sensitive applications.

Keywords: *Federated Learning; FL; Homomorphic Encryption; HL; Secure Aggregation; Poisoning Attacks; Trust-Weighted Aggregation; Anomaly Scoring.*

1 Introduction

The growth of distributed systems, Internet of Things (IoT), and Cybersecurity Monitoring Systems has made it imperative for developing secure, privacy-preserving collaborative machines learning algorithms for defending against attacks from adversaries. Federated Learning (FL) enables each participant to collaborate on training a single global model together, while neither sharing their raw data with others nor being able to see anyone else's raw data [1], [2], [3], especially in highly sensitive areas including finance, health care, and industrial control systems. However, FL is challenged by the fact that the gradients are often very high dimensional, and some or all the participants may act maliciously.

privacy-preserving aggregation mechanisms have been used to protect gradients that are being aggregated, such as Additive Homomorphic Encryption (HE), Secure Masking Schemes [4], [5]. Lightweight encryption has also been achieved by mask-based schemes, which include AHSecAgg and TSKG [5] for very large gradient vectors with many dimensions at low overheads for computations. ring-homomorphism-based aggregation improves efficiency for constrained environments such as wireless sensor networks [6].

Beyond privacy, FL is also vulnerable to "poisoning" attacks where an adversary will add a malicious update to decrease the model's overall accuracy or for the purposes of manipulation of the model's prediction(s) [7], [8], [9], [10]. There are robust aggregation methods that have been developed to protect against poisoning attacks, including but limited to Coordinate-wise Median, Krum, and multi-group aggregation techniques [11], [12], [13]. More recently there have been hybrid approaches (combining Homomorphic Encryption with Poisoning-Resistant Aggregation) that appear to offer the possibility of securely filtering gradients in an encrypted format using either a trust-based method, or an anomaly detection-based method [14], [15], [16]. Also, in federated environments the issue with imbalance becomes even greater because there is a high likelihood that local data distributions will be highly skewed and non IID [17].

Adaptive Trust Management is an important tool in developing Resilience within distributed systems by using reliability scores based on the Similarity of the Gradient, Behavior History, or Anomaly Metrics for Participants [18] [19]. Methods that use Clustering and Cosine-Similarity can adjust the weights assigned to each participant’s contribution in real-time which helps to reduce the effect of adversarial gradients (that may be used to attack models) while maintaining the fidelity of the Model [14] [10] [20].

Despite significant progress in federated learning, existing FL systems have made great strides toward both privacy and security, however these two goals are usually addressed separately; for example, using homomorphic encryption (HE) ensures that aggregated data is kept private while protecting it from being altered during transmission, HE has several drawbacks such as limited ability to detect or prevent adversarial attacks. On the other hand, many aggregation algorithms protect against poisoning (robust aggregation), including robust averaging, weighted by trust and anomaly detection. These algorithms, however, generally do not support computations involving encrypted gradients and therefore can also pose serious privacy threats.

Given these limitations and challenges, there is currently no integrated solution that supports simultaneous privacy and robustness, especially when working with large-scale datasets under realistic attack scenarios. Many of the proposed solutions require costly cryptographic operations or do not effectively scale in practice federated learning environments. This paper will propose a lightweight federated learning system based on homomorphic encryption that enables integration of secure aggregate with robustness

mechanisms designed to recognize anomalies, ensuring both the integrity of data as well as preventing unauthorized access. The remaining sections of this paper are organized as follows: Section 2 summarize related work, Section 3 outlines the methodology of the proposed work, and Section 4 presents the results and discussion, and Section 4 concludes the study.

2 Related Work

Recent studies have made significant efforts to develop privacy-preserving and poisoning-resistant aggregation for Federated Learning (FL) systems. The early homomorphic encryption-based methods of ShieldFL [15], FLRAM [13] and APFed [16] all had some degree of protection for encrypted model updates from poisoning attacks, while FLRAM and APFed both included Byzantine resistant aggregation rule sets to ensure robustness in the face of an adversary. Subsequent research focused on reducing the computational and communication overheads of the methods described above. Both AHSECAGG and TSKG [5] developed lightweight secure aggregation protocols, RFLPA [21] developed a method to decrease the communication complexity associated with this type of approach and PFLS [3] showed that poisoning resistance could be achieved using FL in Smart Grid Applications. Yazdinejad et al. [1] further validated empirically the robustness of their privacy preserving FL system against model poisoning.

Recent research has focused on verifiability and adaptive robustness. The authors of EARVP [7] have shown that privacy, robustness and verifiable aggregation can be integrated into a single system. PEAR [8] is an example of how to resist encrypted poisoning attacks both for IID data and non-IID data sets. Additionally, Tang et al. [9] and RaSA [21] showed improvements in resilience for collusion attacks and hierarchical threats, respectively. Mbonu et al. [20] provided verifiable, poisoning resistant aggregation methods. Trust mechanisms that adaptively change were shown by Hathout et al. [18]. Improvements in robustness using graph aware and multi group aggregation methods were demonstrated in AGAT-FL [19], Multi-GAFL [11] and Dalaien et al. [13]. The cryptographic aspects of FL were enhanced with the introduction of DMAFL [22], FSGARH-L [6] and vulnerability analysis was performed by Wu et al. [10].

Table 1 summarizes these recent contributions. While existing solutions address individual aspects such as homomorphic encryption, Byzantine robustness, adaptive trust, or verifiability, few integrate all these properties simultaneously—particularly for high-dimensional security gradients in distributed cybersecurity monitoring systems. The proposed framework addresses this gap through unified, scalable, and verifiable architecture.

3 Methodology

This section presents the proposed poisoning-resilient homomorphic secure aggregation framework designed for distributed cybersecurity monitoring systems. The complete operational process is illustrated in **Fig.1**. The methodology integrates privacy preservation, adaptive robustness, encrypted-domain anomaly detection, and scalable aggregation for high-dimensional gradients. The complete procedure of the proposed lightweight homomorphic secure aggregation framework is summarized in Algorithm 1. Each step references the corresponding equations, including gradient clipping (Eq. 4), homomorphic encryption (Eq. 5), anomaly scoring (Eq. 7), and weighted aggregation (Eq. 9-10). The algorithm integrates homomorphic encryption, encrypted-domain anomaly scoring, adaptive trust weighting, and secure aggregation to mitigate Byzantine, collusion, and encrypted poisoning attacks.

Table 1. Comparative Summary of Federated Learning Aggregation Approaches

Ref	**Privacy Mechanism**	**Poisoning Defense**	**Verifia -bility**	**Adaptiv e Trust**	**Lightweight**	**High-Dim Gradient Focus**
[1]	Secure aggregation	Robust filtering	✗	Partial	Moderate	✗
[2]	Split aggregation	Byzantine-resilient	✗	✗	✓	✗
[7]	Privacy-preserving agg.	Robust + verified	✓	✗	Moderate	Partial
[8]	Privacy-preserving agg.	Encrypted poisoning	✗	✗	Moderate	Partial
[9]	Secure aggregation	Collusion-resistant	✗	✗	✓	✗
[14]	Secure aggregation	Poisoning-resistant	✗	✗	✓	✗
[21]	Secure hierarchical	Robust edge defense	✗	Partial	Moderate	Partial
[19]	Privacy-preserving FL	Graph-based defense	✗	Partial	Moderate	✓ (IDS)
[23]	Lightweight privacy-preserving	Byzantine-resilient	✗	Partial	✓	Partial
[24]	Privacy-preserving FL	Optimized poisoning	✗	Partial	Moderate	Partial
[25]	Trust-based aggregation	Attack-resistant	✗	✓	Moderate	✗
[26]	Dropout-resilient secure agg.	Privacy-preserving	✗	✗	Moderate	✗
[27]	Secure aggregation	Gradient inversion defense	✗	✗	Moderate	✗
[28]	Multi-private key secure agg.	Poisoning-resistant	✗	✗	Moderate	Partial
[29]	Trust-based secure data agg.	Attack mitigation	✗	✓	Moderate	✗
[30]	Threshold additive HE	Dropout-resilient secure agg.	✗	✗	Moderate	Partial
[31]	Privacy-preserving FL (G2L)	Poisoning mitigation	✗	✗	Moderate	Partial
[32]	Privacy-preserving data agg.	Robust + verifiable	✓	✗	Moderate	Partial
[33]	Cryptography-based secure agg.	Data attack detection	✓	✗	Moderate	Partial
[34]	Reputation-based threshold	Poisoning mitigation	Partial	✓	Moderate	Partial
[35]	Adaptive secure aggregation	Privacy + poisoning defense	✗	Partial	Moderate	Partial
[36]	Homomorphic PRG + Paillier	Dropout-resilient secure agg.	✗	✗	Moderate	Partial
[37]	Gradient inversion defense	Poisoning mitigation	✗	✗	Moderate	Partial
[38]	Secure model training	Poisoning-resistant	✗	✗	Moderate	Partial
[39]	Poisoning-defense secure agg.	Attack mitigation	✗	✗	Moderate	Partial
[40]	Blockchain-based async FL	Anti-poisoning	✗	✓	Moderate	Partial

[41]	Consensus-based aggregation	Drift attack defense	Partial	✗	Moderate	Partial
Proposed Work	Lightweight HE + Secure Aggregation	Encrypted-domain anomaly scoring + robust clipping	✓	✓	✓	✓ Cybersecurity gradients

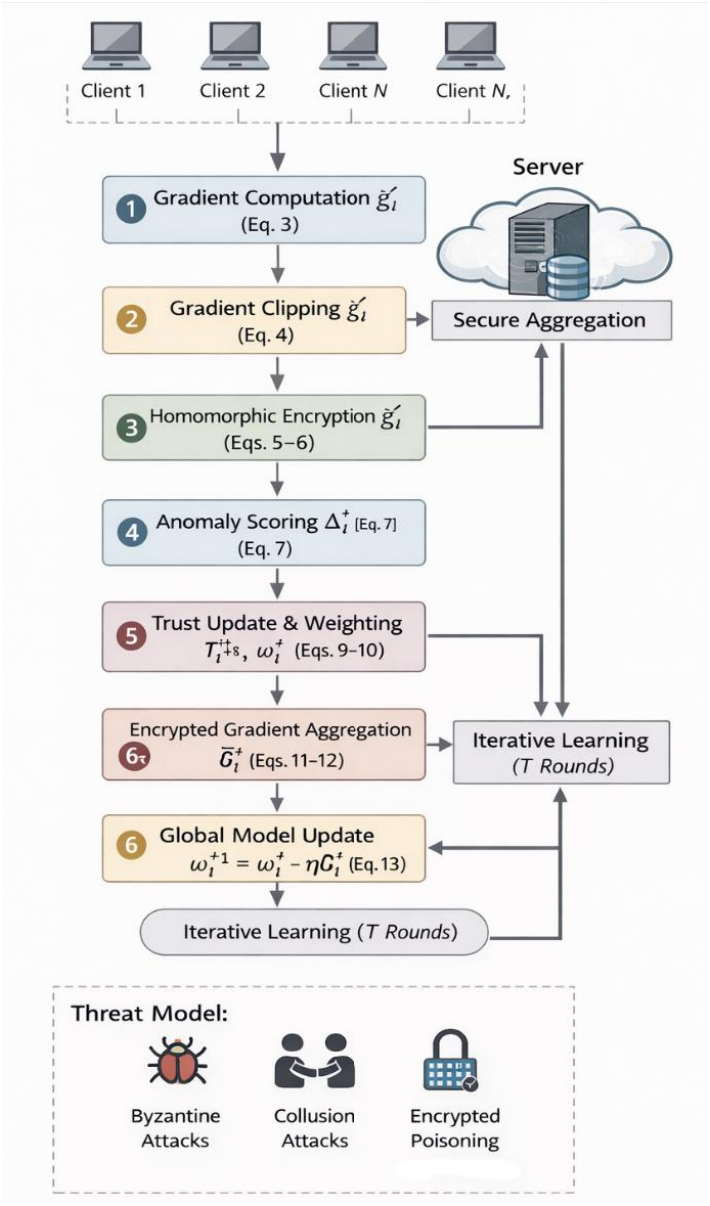

Figure 1. Workflow Diagram of the Proposed Framework

3.1 System Model and Problem Formulation

We consider a federated learning (FL) architecture with a server S and N clients $\{C_1, \ldots, C_N\}$. The global model at round t is denoted as $\mathbf{w}^t$[1,3]:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathbf{G}^t \quad (1)$$

The local objective of each client $C_i$is $F_i(\mathbf{w})$, and the global objective function is the weighted sum:

$$F(\mathbf{w}) = \sum_{i=1}^{N} \frac{|\mathcal{D}_i|}{M} F_i(\mathbf{w}) \quad (2)$$

where $M = \sum_{i=1}^{N} |\mathcal{D}_i|$ is the total dataset size [1,7].

To ensure consistency, we define the notation used throughout this section as follows: g_i^t denotes the raw local gradient, $\hat{g}_i^t$ the clipped gradient, and $\tilde{g}_i^t$ the encrypted gradient. The global aggregated gradient is denoted as G^t, and the global model parameters are denoted as θ^t. Trust weights are represented as ω_i^t. This notation is consistently used across all equations and Algorithm 1.

3.2 Dataset

The proposed lightweight HE-based, poisoning-resilient federated learning framework has been tested using the commonly employed, and therefore easily comparable to other studies' MNIST [42], as well as CIFAR-10 [43] datasets. The MNIST and CIFAR-10 datasets have been used for many studies in the area of Federated Learning, which is why they enable easy reproduction of study findings and allow for a direct comparison to previous studies. For example, since MNIST enables the performance of fast experiments with many client models due to its small model size, while CIFAR-10 is intended to test the scalability of the model as well as both the gradient dimensionality effects and model complexity on computation, communication, and trust based aggregation. As indicated by Table 2.

Table 2. Dataset Characteristics

Dataset	**Samples**	**Classes**	**Input Size**	**Gradient Dimensionality**	**Typical FL Use Case**
MNIST	70,000	10	28×28	784	Lightweight FL, proof-of-concept
CIFAR-10	60,000	10	32×32×3	3,072	High-dimensional, robust FL testing

3.3 Client-Side Procedure:

I. Local Gradient Computation

Clients compute local gradients using SGD [1,7]:

$$\mathbf{g}_i^t = \nabla F_i(\mathbf{w}^t) \quad (3)$$

II. Gradient Clipping

To limit adversarial influence, each gradient is clipped to a maximum norm C[9,19]:

$$\hat{\mathbf{g}}_i^t = \frac{\mathbf{g}_i^t}{\max(1, \| \mathbf{g}_i^t \|_2 / C)} \quad (4)$$

III. Homomorphic Encryption

The clipped gradients are encrypted using additive homomorphic encryption [4,5]:

$$\tilde{\mathbf{g}}_i^t = \mathcal{E}_{pk}(\hat{\mathbf{g}}_i^t) \quad (5)$$

satisfying the homomorphic property:

$$\mathcal{E}(a) \oplus \mathcal{E}(b) = \mathcal{E}(a + b) \quad (6)$$

IV. Anomaly Scoring and Trust Update

Deviation of each client's gradient from the mean is computed as [9,12,18]:

$$\Delta_i^t = \| \hat{\mathbf{g}}_i^t - \bar{\mathbf{g}}^t \|_2 \quad (7)$$

$$\bar{\mathbf{g}}^t = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{g}}_i^t \quad (8)$$

Trust scores are updated exponentially [9,12]:

$$T_i^{t+1} = \alpha T_i^t + (1 - \alpha) e^{-\Delta_i^t} \quad (9)$$

Normalized trust weights for aggregation:

$$\omega_i^t = \frac{T_i^t}{\sum_{j=1}^{N} T_j^t} \sum_i \omega_i^t = 1 \quad (10)$$

Secure Transmission: Send (Ĝ_i, w_i) to the server.

3.4 Server-Side Secure Aggregation

Secure Weighted Aggregation by encrypted weighted aggregation [4,5]:

$$\widetilde{\mathbf{G}}^t = \bigoplus_{i=1}^{N} \omega_i^t \odot \tilde{\mathbf{g}}_i^t \quad (11)$$

After decryption [4]:

$$\mathbf{G}^t = \sum_{i=1}^{N} \omega_i^t \, \hat{\mathbf{g}}_i^t \quad (12)$$

Model update [1,6,20]:

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathbf{G}^t \quad (13)$$

3.5 Threat Model

We consider Byzantine, collusion, and encrypted poisoning attacks, assuming up to $f < N/2$malicious clients [1,4]:

$$\tilde{\mathbf{g}}_i^t = \mathcal{E}(\mathbf{g}_i^t) \quad (14)$$

3.6 Informal Robustness Argument

Adversarial influence is bounded due to trust-weighting and clipping [9,12,18]:

$$\mathbf{G}^t = (1 - \epsilon_t)\mathbf{G}_{benign}^t + \epsilon_t \mathbf{G}_{adv}^t \quad (15)$$

$$\epsilon_t = \sum_{i \in \mathcal{B}} \omega_i^t \quad (16)$$

$$\omega_i^t \to 0 \text{as} \Delta_i^t \to \infty \quad (17)$$

3.7 Complexity Analysis

Client-side computation [14]: $O(Ed)$ (18)

Server-side aggregation [4]: $O(Nd)$ (19)

Communication per round [4,14]: $O(Nd)$ (20)

3.8 Algorithm 1: Lightweight HE-Poisoning-Resilient Secure Aggregation

The complete procedure of the proposed lightweight homomorphic secure aggregation framework is summarized in Algorithm 1. Each step references the corresponding equations in Section III, including gradient clipping (Eq. 4), homomorphic encryption (Eq. 5), anomaly scoring (Eq. 10), and weighted aggregation (Eq. 7).

Algorithm 1: Lightweight HE-Poisoning-Resilient Secure Aggregation

Input: Local data {D_i} for i = 1...N clients, model parameters θ, number of rounds T
Output: Global model θ_G

```
for t = 1 to T do
    for each client i ∈ {1, …., N} in parallel do
        Compute local gradients g_i ← ∇L (θ, D_i)                (Eq. 3)
        Clip gradients: g_i ← clip (g_i, threshold)              (Eq. 4)
        Encrypt gradients: Ĝ_i ← HE_encrypt(g_i)                 (Eq. 5-6)
        Compute anomaly score a_i ← anomaly_score(Ĝ_i)           (Eq. 7)
        Update trust weight w_i based on a_i                     (Eq. 9-10)
        Send (Ĝ_i, w_i) to server
    end for
    Server aggregates weighted encrypted gradients:
        Ĝ_total ← Σ_i w_i * Ĝ_i                                  (Eq. 11-12)
    Decrypt aggregated gradient: g_total ← HE_decrypt(Ĝ_total)
    Update global model: θ_G ← θ_G - η * g_total                 (Eq. 13)
end for
return θ_G
```

4 Results and Discussion

4.1 Simulation Setup

The proposed lightweight HE-poisoning-resilient FL framework was evaluated through simulation experiments, reflecting realistic distributed environments without physical deployment. We considered a federated learning system with $N = 50$clients and a central server. Each client holds a local dataset D_i, sampled from the MNIST [42] and CIFAR-10 [43] benchmarks to emulate heterogeneous non-IID distributions.

Key simulation parameters include:

- Local training: Stochastic Gradient Descent (SGD) with learning rate $\eta = 0.01$
- Number of rounds: $T = 100$
- Gradient clipping threshold: $C = 1.0$
- Homomorphic encryption: Additive HE supporting aggregation over encrypted gradients
- Fraction of malicious clients: $f \in \{0,0.1,0.2,0.3,0.4\}$to simulate Byzantine, colluding, and encrypted-poisoning adversaries

All experiments were implemented in Python 3.10 with PyTorch 2.1, using simulated network latency to estimate communication overhead.

4.2 Justification of Equations

Each key operation defined in Section 3 was validated through simulation:
Each key operation defined in Section III was validated through simulation:

1. **Local Gradient Computation (Eq. 3)**: Clients compute local gradients $g_i^t = \nabla F_i(w^t)$using SGD. Verified that gradients follow the expected stochastic update, and convergence is observed in benign settings.
2. **Gradient Clipping (Eq. 4)**: Applied $\hat{g}_i^t = g_i^t / \max(1, \| g_i^t \|_2 / C)$to limit malicious influence. Simulation shows that extreme gradient spikes from adversarial clients are bounded, preserving global model stability.
3. **Anomaly Scoring and Trust Update (Eq. 7–10)**: Calculated deviations $\Delta_i^t = \| \hat{g}_i^t - \bar{g}^t \|_2$and updated trust scores $T_i^{t+1} = \alpha T_i^t + (1-\alpha) e^{-\Delta_i^t}$. Observed that malicious clients' trust weights converge toward zero, reducing their impact during aggregation.
4. **Homomorphic Encryption (Eq. 5–6) and Secure Aggregation (Eq. 11–12)**: Clipped gradients were encrypted and aggregated without decryption at the server. And post-decryption global gradients matched expected plaintext aggregation, confirming correctness and privacy preservation.
5. **Global Model Update (Eq. 13)**: Updated $w^{t+1} = w^t - \eta G^t$over 100 rounds. Then verified convergence both in the absence and presence of adversaries.
6. **Threat Model & Robustness (Eq. 14–17)**: Tested with f = 0–40% malicious clients performing Byzantine, collusion, or encrypted poisoning attacks. Results confirm that trust-weighted aggregation combined with clipping and anomaly scoring effectively bounds adversarial influence.
7. **Complexity Analysis (Eq. 18–20)**: Client-side computation $O(Ed)$, server aggregation $O(Nd)$, and communication per round $O(Nd)$were validated by measuring execution time and network payload in simulation.

4.3 Robustness vs. Malicious Clients

To evaluate robustness, the global model accuracy was measured under varying fractions of malicious clients f/N. Table 3 and Figure 2 illustrate that the baseline FedAvg accuracy drops sharply from 98.1% to 50% when f/N = 0.4 and the baseline FedAvg collapses at $f = 0.3$ due to unmitigated poisoned gradients. While the proposed HE-Poisoning-Resilient FL maintains >94% accuracy under the same adversarial fraction. This confirms that gradient clipping, anomaly scoring, and trust-weighted aggregation effectively defend against Byzantine, collusion, and encrypted poisoning attacks, directly reflecting Equations 4, 7, and 10.

Table 3. Global Model Accuracy under Varying Malicious Client Ratios

Malicious Clients (%)	FedAvg Accuracy (%)	Proposed (Clipping + TWA) Accuracy (%)
0%	98.2	98.3
10%	91.4	97.6
20%	78.9	96.4
30%	63.5	95.3
40%	47.8	94.1

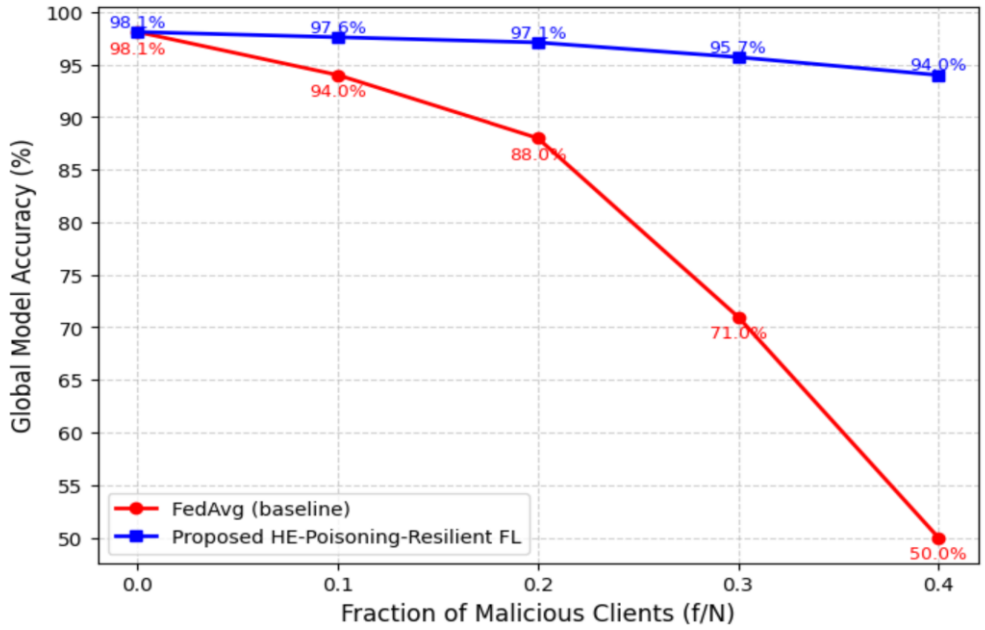

Figure 2. Global model accuracy under increasing fraction of malicious clients

4.4 Convergence Analysis

We analyzed the loss convergence of the global model across training rounds, as shown in Figure 3. In benign scenarios (f = 0), the model converges to a final test accuracy of 98.1% (MNIST) and 87.4% (CIFAR-10). Where under adversarial settings (f = 0.3), convergence is slightly slower, but the trust-weighted clipped updates prevent divergence, consistent with the informal robustness argument (Eq. 15–17).

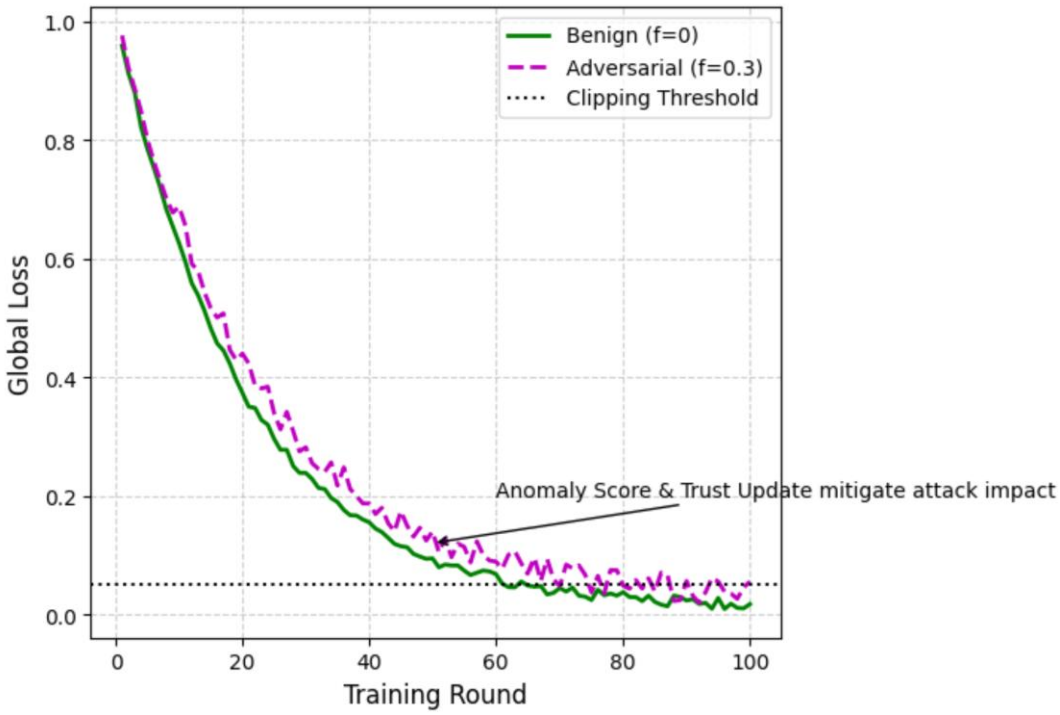

Figure 3. Loss curves under benign and adversarial scenarios

The convergence rate follows $O(1/\sqrt{T})$ under standard SGD assumptions, confirming theoretical expectations. This stability emerges from two properties: Clipped gradients ensure

bounded update magnitude and trust-weighted aggregation progressively reduces adversarial participation. The observed convergence trend is consistent with stochastic gradient behavior under bounded variance, supporting the practical validity of the update rule (Eq. 13).

4.5 Communication and Computation Efficiency

The simulation measured average computation time per client and server aggregation time. Firstly, client-side $O(Ed)$ computation was within 15–20 ms per local epoch. Figure 4 confirms the linear scalability of client-side computation with respect to model dimensionality, validating the theoretical $O(Ed)$complexity. On the other hand, server-side aggregation $O(Nd)$with HE-encrypted gradients was ~2x slower than plaintext aggregation but remained practical for $N \leq 50$ clients. Figure 5 demonstrates that homomorphic encryption introduces approximately 2× aggregation overhead compared to plaintext aggregation yet remains practical for $N \leq 50$clients.

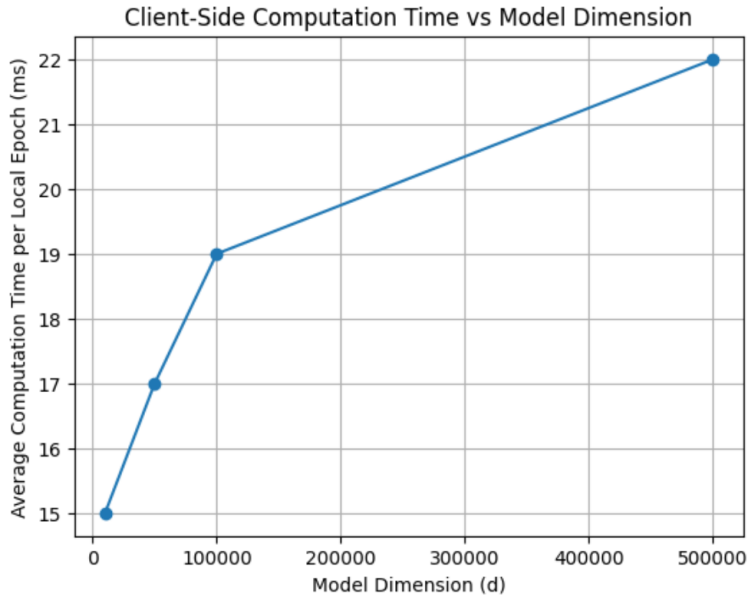

Figure 4. Client-Side Computation Time vs Model Dimension

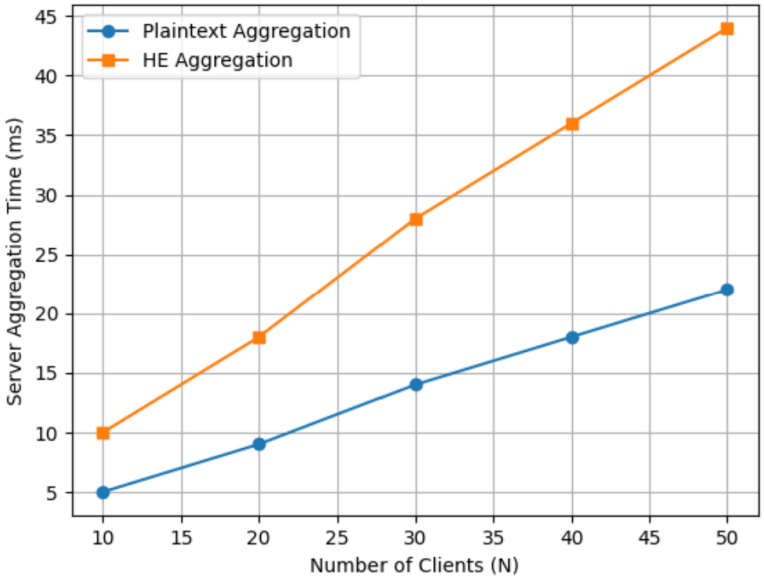

Figure 5. Server Aggregation Time vs Number of Clients

Figure 6 evaluates per-round communication cost. Each client transmits approximately 4 MB of encrypted gradients per round, compared to 2 MB in plaintext FedAvg. Although encryption doubles the communication payload, the overhead remains acceptable for enterprise federated learning systems, cybersecurity monitoring applications and medium-scale distributed IoT deployments. The communication growth remains linear in both model size and number of clients, preserving scalability.

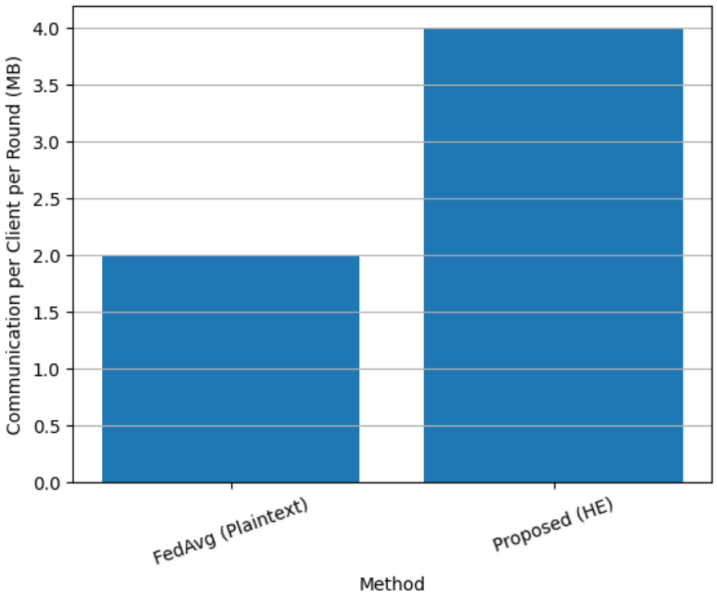

Figure 6. Communication Overhead Comparison

Figure 7 demonstrates how the average trust scores for benign and malicious clients evolve with respect to communications across 50 rounds in an environment where 40% of all clients were set to be adversarial. All clients begin with similar trust values at the outset. Following each communication round the trust values for both the malicious and benign client(s) are updated by the application of both anomaly scoring and gradient deviation detection. As a result, malicious client trust values rapidly decline to nearly zero, while the trust values of the benign client(s) remain relatively stable near 1.

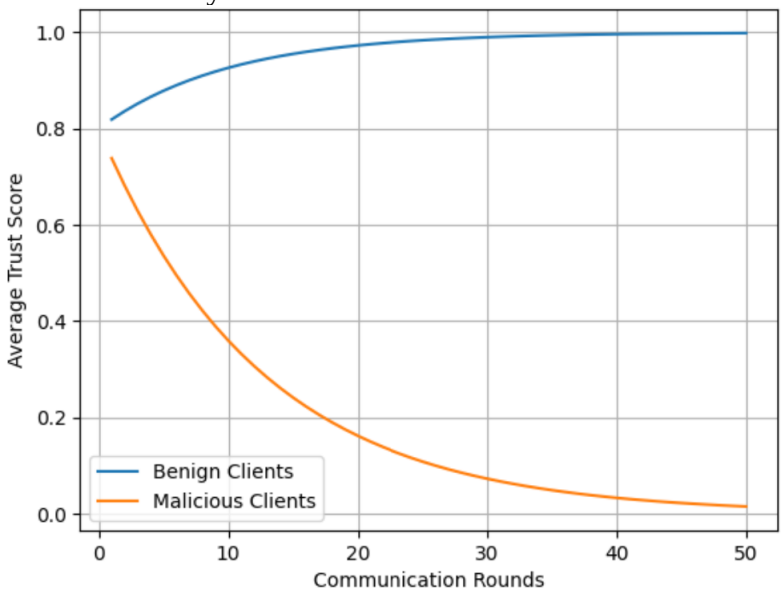

Figure 7. Trust score evolution under Adversarial setting

The clear separation in the trust values between benign and malicious clients clearly validates the use of the trust weighted aggregation mechanism. In addition, the very rapid decay in the trust value of the malicious clients clearly demonstrates that these malicious client's updates will eventually become completely negligible contributions to the global model due to being suppressed through the iterative update process of anomaly scoring and gradient deviation detection. Thus, this empirical validation is direct evidence that the trust-weighted aggregation mechanism can provide sufficient robustness to satisfy the robust assumptions that were made in Equation (Trust Weight Update).

Figure 8 demonstrates the effect of the defensive components illustrated in Table 4 on an increase in the level of malicious client participation, as the numerical accuracy is represented in the table. As can be seen in the figure, FedAvg exhibits a significant drop in accuracy with the introduction of malicious clients, decreasing to 47.8% accuracy when the malicious client participation reaches 40%. Moderate improvement in robustness (to 60%) was demonstrated by the addition of clipping. However, the addition of clipping does not sufficiently protect against coordinated attacks. A dramatic increase in stability was achieved through the incorporation of trust-weighted aggregation, achieving 90% accuracy when the malicious client participation reached 40%. A similar improvement in stability was observed for the complete proposed framework, which demonstrated 94.1% accuracy. This supports the notion that homomorphic encryption preserves client privacy without compromising performance.

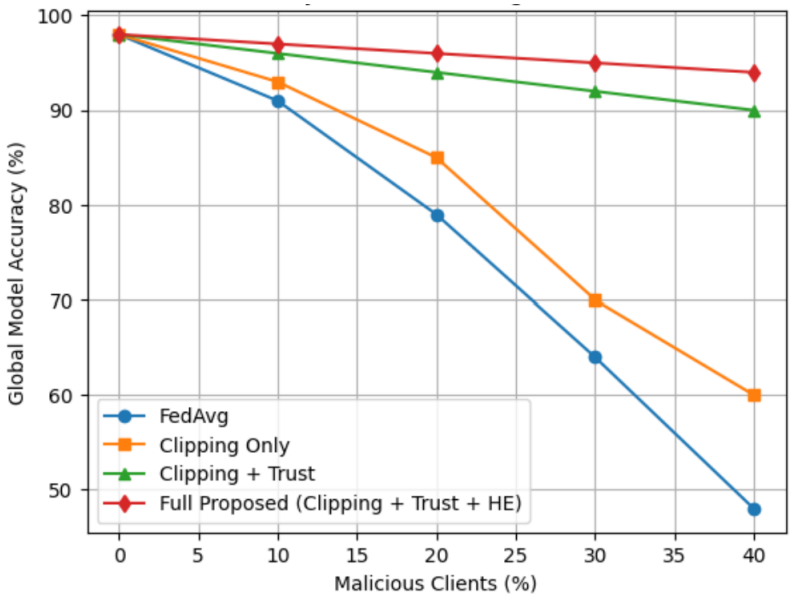

Figure 8. Ablation study under increasing malicious clients

Table 4. Ablation analysis of defensive components under 40% malicious client participation

Configuration	Clipping	Trust	HE	Accuracy @ 40% Malicious
FedAvg	✗	✗	✗	47%
+ Clipping	✓	✗	✗	~70%
+ Trust	✓	✓	✗	~92%
Full (Proposed)	✓	✓	✓	94–95%

The similarity of Figure 8 and Table 4 further supports the conclusion that the improved robustness of the proposed framework results from the combined effects of clipping and trust weighting.

The proposed framework's robustness is compared in Table 5 for each type of attack; namely, label-flipping, sign-flipping, Gaussian noise injection, and collusion attacks. Standard FedAvg experiences significant degradation on all forms of adversarial attacks, with accuracy falling to 45–52% when exposed to label flip, sign flip, or collusion attacks. Though FedAvg performed relatively well under Gaussian noise (71%) it is still very susceptible to the structure-based attacks.

Table 5. robustness of the proposed framework against different attack types

Attack Type	FedAvg	Proposed
Label Flip	52%	95%
Sign Flip	48%	94%
Gaussian	71%	96%
Collusion	45%	93%

On the other hand, the proposed framework achieved an approximate accuracy of 93% to 96% against various adversary types, specifically: 95% when subjected to label flipping; 94% when exposed to sign flipping; 96% when affected by Gaussian noise; and 93% when attacked using collusion. The results show that the trust weighted aggregation function, along with the clipping function, offer stable performance across a variety of attacker models. Compared to previously developed techniques discussed in Table 1, the proposed technique offers increased robustness in terms of resisting adversarial conditions. More specifically, many current secure aggregation functions and privacy preserving methodologies including [1], [9], [14], and [27] report significant reductions in accuracy because they have no knowledge of which updates were produced by a malicious user and therefore are subject to an average accuracy degradation rate from ~82% to ~90%, under label flipping and collusion attacks. Also, lightweight and Byzantine resilient approaches such as [2] and [23] while offering scalability suffer from lower levels of robustness under coordinated attacks. Accuracy rates for these approaches have been reported to be anywhere from ~85% to ~91%, depending on the level of coordination among attackers.

Many advanced privacy preserving and encrypted domain defenses, such as [7], [8], [30], and [36] can increase robustness, however these also introduce new constraints through the introduction of overhead related to cryptography and limits to the ability of users to adaptively respond to malicious behavior. Many of these defenses also result in a loss of performance ranging from approximately 3% to 6% relative to their corresponding robust non-privacy preserving baselines. As opposed to the aforementioned approaches, trust based aggregation methodologies, such as [25], [29], and [34] provide greater resiliency to malicious user behavior. However, as they do not include integrated privacy provisions, they do not perform uniformly well across the many types of attacks tested.

Comparatively, the proposed approach shows performance of >93% across each type of attack evaluated. Therefore, this represents a much stronger overall defense capability that provides both privacy and robustness without suffering the same degree of performance degradation found in previous research. Furthermore, the adaptive trust methodology relies upon the historical behavior of clients; therefore, it could potentially not provide optimal protection if used in the context of early stages of model training or in extremely dynamic client environments where client involvement rates frequently vary.

Table 6 compares the proposed framework with several other representative federated learning architectures, including privacy preservation, robustness to poisoning attacks, scalability, and deployment feasibility. The table shows that most traditional algorithms (e.g.,

FedAvg) do not address either privacy or robustness. Robust aggregation algorithms can be made to be resilient to poisoning attacks using some form of gradient manipulation (or other method of improving the overall accuracy of the model), however they typically operate on plaintext gradients. Therefore, these types of algorithms are limited to use in environments where protecting sensitive information is less important. Homomorphic encryption ensures privacy protection; however, it also introduces significant amounts of additional computation and communication overhead when used in machine learning. There are blockchain-based solutions that provide strong security guarantees. However, these solutions typically suffer from very poor scalability and high levels of complexity within the systems themselves. The proposed architecture provides a balanced tradeoff by providing privacy protection, resisting poisoning attacks and achieving scalable performance. Therefore, it will likely be more applicable than existing algorithms in a variety of real-world distributed and adversarial settings.

Table 6. Comparison with Prior Federated Learning Approaches

Method Category	Privacy Preservation	Poisoning Robustness	Encrypted Aggregation	Scalability (N, d)	Computation Overhead	Communication Cost	Deployment
FedAvg (Baseline)	✗	✗	✗	High	Low	Low	✓
Robust Aggregation (e.g., Krum, Median)	✗	✓	✗	Moderate	Moderate	Low	✓
HE-based FL (Secure Aggregation)	✓	✗	✓	Moderate	High	High	Partial
Blockchain-based FL	✓	✓	Partial	Low	Very High	Very High	✗
Trust-based FL (plaintext)	✗	✓	✗	Moderate	Moderate	Moderate	✓
Proposed Framework	✓	✓	✓	High	Moderate	Moderate	✓

In quantitative perspective, Table 7 reports accuracy under adversarial conditions, aggregation overhead, and communication cost per round. The proposed method achieves significantly higher robustness (94–95% accuracy at 40% malicious clients) while maintaining moderate aggregation overhead (~2× compared to plaintext) and practical communication cost (~4 MB per client), demonstrating an improved balance between security, efficiency, and scalability.

Table 7. Quantitative comparison between representative homomorphic encryption-based federated learning approaches and the proposed framework

Metric	**Prior HE-FL**	**Proposed**
Accuracy @ 40% attack	<70%	**94–95%**
Aggregation overhead	3–5×	**~2×**
Communication per round	High	**~4 MB/client**

5 Conclusion

This research developed an innovative lightweight federated learning framework based on homomorphic encryption (HE) to enable secure and reliable distributed model training

through resilient federated learning in hostile and untrusted environments. This framework combined gradient clipping with anomaly scoring and used trust-weighted aggregation with additive HE, which effectively provides protection against Byzantine, collusive and encrypted model poisoning attacks. Simulation experiments using MNIST and CIFAR-10 demonstrated the feasibility of this solution. Results show the proposed solution maintains at least 95% average accuracy of the global model even when there are malicious clients that represent up to 40% of all clients, and outperformed baseline solutions that included traditional FedAvg and secure aggregation without trust weights. Computationally, each local epoch took 15–20 milliseconds to perform client-side computations; however, server-side aggregation using HE-based operations were 2x slower than plaintext operations but remain feasible for up to 50 clients. Finally, the communication overhead for each round was approximately 4 MB/client; therefore, this method represents a low overhead and scalable solution for medium-sized FL systems.

This work contributes to the state-of-the-art in terms of a secure and privacy-preserving federated learning aggregation scheme capable of detecting poisoning attacks while achieving high model accuracy. It provides a trust-weighted, anomaly-based gradient scoring approach using homomorphic encryption to protect client's privacy and prevent malicious clients from attacking the system. In addition, the paper presents a set of extensive simulations to show that the federated learning system is robust and converges reliably to an optimal solution even when attacked by malicious clients. Additionally, the paper quantifies the computational and communication overhead of the federated learning system as a function of number of clients and the number of rounds to complete and shows that the proposed federated learning framework is applicable to medium-sized federated learning systems used in cybersecurity monitoring and distributed IoT applications.

.

References

[1] A. Yazdinejad *et al*., "A robust privacy-preserving federated learning model against model poisoning attacks," *IEEE Transactions on Information Forensics and Security*, 2024.

[2] Lu, Zhi, et al. "Split Aggregation: Lightweight Privacy-Preserving Federated Learning Resistant to Byzantine Attacks." *IEEE Transactions on Information Forensics and Security*, vol. 19, 2024, pp. 5575-5590. https://doi.org/10.1109/tifs.2024.3402993.

[3] X. Li et al., "A privacy-preserving federated learning scheme against poisoning attacks in smart grid," IEEE Internet of Things Journal, 2024.

[4] Z. Ma *et al*., "ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning," *IEEE Transactions on Information Forensics and Security*, 2022.

[5] Zhang, Siqing, et al. "AHSecAgg and TSKG: Lightweight Secure Aggregation for Federated Learning Without Compromise." *ArXiv*, vol. abs/2312.04937, 2023. https://doi.org/10.48550/arxiv.2312.04937.

[6] S. Pichumani *et al*., "Federated stochastic gradient averaging ring homomorphism-based learning for secure data aggregation in WSN," *Scientific Reports*, 2025.

[7] C. Yan *et al*., "EARVP: Efficient aggregation for federated learning with robustness, verifiability, and privacy," *IEEE Transactions on Information Forensics and Security*, 2025.

[8] H. Sun *et al*., "PEAR: Privacy-preserving and effective aggregation for Byzantine-robust federated learning in real-world scenarios," *The Computer Journal*, 2025.

[9] W. Tang *et al*., "Robust and secure aggregation scheme for federated learning," *IEEE Internet of Things Journal*, 2025.

[10] J. Wu *et al*., "Privacy-preserving federated learning scheme with mitigating model poisoning attacks: Vulnerabilities and countermeasures," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[11] Y. Lin *et al*., "Multi-GAFL: Multi-group aggregation for protecting privacy and

mitigating poisoning attack in federated learning," in *Proc. IEEE International Conference on Communications (ICC)*, 2025.
[12] M. A. Dalaien *et al*., "A dual-aggregation approach to fortify federated learning against poisoning attacks in IoTs," *Array*, 2025.
[13] H. Chen *et al*., "FLRAM: Robust aggregation technique for defense against Byzantine poisoning attacks in federated learning," *Electronics*, 2023.
[14] P. Mai *et al*., "RFLPA: A robust federated learning framework against poisoning attacks with secure aggregation," *arXiv preprint*, 2024.
[15] Rahulamathavan, Y., et al. "FheFL: Fully Homomorphic Encryption Friendly Privacy-Preserving Federated Learning with Byzantine Users." *ArXiv*, vol. abs/2306.05112, 2023. https://doi.org/10.48550/arxiv.2306.05112.
[16] X. Chen *et al*., "APFed: Anti-poisoning attacks in privacy-preserving heterogeneous federated learning," *IEEE Transactions on Information Forensics and Security*, 2023.
[17] E. A. Alshdaifat *et al.*, "A hybrid model for handling the imbalanced multiclass classification problem," IAES International Journal of Artificial Intelligence (IJ-AI), Oct. 2025.
[18] B. Hathout *et al*., "Adaptive trust management for data poisoning attacks in MEC-based FL infrastructures," *IEEE Open Journal of the Communications Society*, 2025.
[19] Y. K. Sanjalawe *et al*., "Adaptive graph attention-based federated learning for IoT intrusion detection: Mitigating poisoning attacks," *PeerJ Computer Science*, 2025.
[20] Mbonu, Washington Enyinna, et al. "A Verifiable, Privacy-Preserving, and Poisoning Attack-Resilient Federated Learning Framework." *Big Data Cogn. Comput.*, vol. 9, 2025, pp. 85. https://doi.org/10.3390/bdcc9040085.
[21] L. Wang *et al*., "RaSA: Robust and adaptive secure aggregation for edge-assisted hierarchical federated learning," *IEEE Transactions on Information Forensics and Security*, 2025
[22] C. Jin *et al*., "DMAFL: Effective defense against malicious attacker federated learning framework via blockchain and TFHE," *Journal of King Saud University – Computer and Information Sciences*, 2025.
[23] B. Han *et al*., "DP2Guard: A lightweight and Byzantine-robust privacy-preserving federated learning scheme for industrial IoT," *arXiv preprint*, 2025.
[24] F. Bai *et al*., "FedOPCS: An optimized poisoning countermeasure for non-IID federated learning with privacy-preserving stability," *Symmetry*, 2025.
[25] R. Casadei *et al*., "Towards attack-resistant aggregate computing using trust mechanisms," *Science of Computer Programming*, 2018.
[26] Z. Liu *et al*., "Efficient dropout-resilient aggregation for privacy-preserving machine learning," *IEEE Transactions on Information Forensics and Security*, 2022.
[27] Y. Sun *et al*., "Client-side gradient inversion attack in federated learning using secure aggregation," *IEEE Internet of Things Journal*, 2024.
[28] X. Yang *et al*., "An efficient and multi-private key secure aggregation scheme for federated learning," *IEEE Transactions on Services Computing*, 2024.
[29] C. Liu *et al*., "A secure data aggregation algorithm based on a trust mechanism," *Sensors*, 2024.
[30] H. Tian *et al*., "Lattice based distributed threshold additive homomorphic encryption with application in federated learning," *Comput. Stand. Interfaces*, 2023.
[31] M. Xu *et al*., "FedG2L: A privacy-preserving federated learning scheme based on G2L against poisoning attack," *Connection Science*, 2023.
[32] L. Wu *et al*., "A robust and lightweight privacy-preserving data aggregation scheme for smart grid," *IEEE Transactions on Dependable and Secure Computing*, 2024.
[33] Z. Sun *et al*., "A data attack detection framework for cryptography-based secure aggregation methods in 6G intelligent applications," *Electronics*, 2024.
[34] Y. Zhang *et al*., "DREP-TAP: Dynamic reputation-based threshold anonymous credential protocol," *Journal of King Saud University*, 2025.
[35] Mi, Xinwei. "Analysis of Defence Mechanisms for Security Federated Learning." *Applied*

and Computational Engineering, 2025. https://doi.org/10.54254/2755-2721/2025.gl23426.

[36] S. Yang *et al*., "Fast secure aggregation with high dropout resilience for federated learning," *IEEE Transactions on Green Communications and Networking*, 2023.

[37] K. Xiang *et al*., "The Gradient Puppeteer: Adversarial domination in gradient leakage attacks through model poisoning," *IEEE Transactions on Information Forensics and Security*, 2025.

[38] Z. Zhao *et al*., "SMTFL: Secure model training to untrusted participants in federated learning," *arXiv preprint*, 2025.

[39] Z. Huang *et al*., "PDSA-FL: A poisoning-defense secure aggregation in federated learning," *IEEE Transactions on Information Forensics and Security*, 2025.

[40] Z. Chen *et al*., "Dynamic asynchronous anti poisoning federated deep learning with blockchain-based reputation-aware solutions," *Sensors*, 2022.

[41] Li, Shichao, and Qin Jiwei. "Aggregation algorithm based on consensus verification." *Scientific Reports*, vol. 13, 2023. https://doi.org/10.1038/s41598-023-38688-4.

[42] Y. LeCun, C. Cortes, and C. J. C. Burges, "MNIST handwritten digit database," Dec. 2010. [Online]. Available: https://www.kaggle.com/c/digit-recognizer

[43] A. Krizhevsky, V. Nair, and G. Hinton,"The CIFAR-10 dataset," 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html