Security and Privacy Protection Methods for Federated Learning for Big Data

Bin Zhou

School of Artificial Intelligence, Zibo Polytechnic University, Zibo, 255000, China
E-mail: 13589524832@163.com

Received 10 February 2026; Accepted 17 March 2026

Abstract

In the big data environment, federated learning faces multiple challenges such as privacy leakage, poisoning attacks, and communication overload. Existing methods, mostly functioning as point defenses, struggle to simultaneously balance security, efficiency, and utility. This research aims to construct a multi-layered federated learning security system encompassing "source protection, process defense, and global optimization". The study builds a cloud-edge-end collaborative architecture, integrating differential privacy with Shamir's secret sharing to achieve data source perturbation and gradient share transmission. Through Mixup data augmentation combined with gradient clustering, it proactively detects poisoning attacks and introduces federated unlearning to remediate malicious impacts post factum. Based on static Bayesian games, it dynamically allocates privacy budgets to achieve a Nash equilibrium between personalized privacy and model utility. Experiments conducted on the CIFAR-10 and FEMNIST datasets, using a convolutional neural network as the base model and comparing it with the Vanilla FL model and module ablation versions, demonstrate the following: the FAA achieves a communication overhead of only 91.9 MB with 75 clients. Mixup combined with gradient clustering maintains an accuracy of

Journal of Cyber Security and Mobility, Vol. 15_3, 603–628.
doi: 10.13052/jcsm2245-1439.1534

69.7% under 24% poisoning attacks. The game-theoretic framework attains a privacy–utility balance coefficient of up to 0.91. In complex dynamic scenarios, the multi-layered framework achieves an accuracy of 76.9%. This system exhibits robust security and adaptability under various attacks, providing a systematic solution for the practical deployment of federated learning.

Keywords: Federated learning, security defense, poisoning attacks, cloud-edge-end collaboration, differential privacy.

1 Introduction

With the rapid development of big data technology, federated learning, as an important paradigm of distributed machine learning, enables cross-institution collaborative training while protecting data privacy and has been widely applied in various fields such as healthcare, finance, and the Internet of Things (IoT) [1, 2]. Current data security research exhibits a diversified development trend, mainly focusing on encryption algorithm improvement, intrusion detection, blockchain integration, and game theory optimization. In the realm of encryption technology, researchers concentrate on optimizing traditional encryption algorithms and constructing hybrid encryption mechanisms to counter increasingly complex attack methods and enhance data transmission confidentiality [3, 4]. In the field of attack detection, deep learning methods, leveraging their powerful feature extraction capabilities, have been widely used for malware classification and network intrusion identification, significantly improving threat discovery efficiency [5]. Blockchain technology, due to its immutable and decentralized nature, is being explored for secure data sharing and verification in industrial IoT and cloud environments to enhance system trustworthiness [6, 7]. Game theory, by constructing attack–defense game models, is applied to the dynamic optimization of network attack and defense strategies, achieving rational allocation of security resources [8]. Existing research provides a technical foundation for federated learning security, but systematic protection solutions targeting its unique challenges, such as heterogeneous data processing, gradient leakage defense, and poisoning attack detection, still require in-depth exploration. Therefore, the research ensures data confidentiality at the terminal through noise addition via differential privacy and segmentation via secret sharing. During the training phase, Mixup data augmentation is employed to dilute malicious samples, while gradient clustering on the server identifies anomalies. Federated unlearning is

introduced to rapidly eliminate malicious influences, combined with dynamic optimization of privacy budgets based on game theory to achieve personalized utility balance. The study aims to construct a multi-layered federated learning security framework that covers data source protection, training process defense, and global utility optimization, enhancing the system's resilience against attacks, communication efficiency, and adaptability to personalized privacy in complex and dynamic scenarios, thereby providing reliable support for practical deployment.

The innovation of this research lies in the construction of an integrated cloud-edge-end collaborative federated learning security system across three dimensions. At the source and transmission layer, differential privacy and Shamir's secret sharing are integrated to achieve data perturbation and gradient share encryption. At the defense layer, Mixup data augmentation combined with gradient clustering enables active detection of poisoning attacks, while federated unlearning is introduced to remediate malicious impacts. At the optimization layer, a static Bayesian game is employed to dynamically allocate privacy budgets, achieving a Nash equilibrium between personalized privacy and model utility. The synergy of these three technologies ensures privacy, security, and efficiency simultaneously, eliminating the need for a trusted third party.

In summary, with the increasing stringency of global data privacy regulations (such as the GDPR and the Personal Information Protection Law), federated learning, as a key technology in privacy-preserving computing, has security implications that directly impact the compliance of sensitive industries such as healthcare and finance. However, existing methods primarily focus on defending against single types of attacks and lack a systematic solution. The multi-layered protection system proposed in this study not only theoretically addresses the shortcomings of current research but also provides an operational security assurance framework for the practical deployment of federated learning, offering significant theoretical innovation and engineering application value.

2 Related Works

Against the backdrop of a simultaneous surge in global data value and security risks, strict regulations and the concept of "proactive governance" are driving the security system to evolve towards a deep integration of management technology [9].

2.1 Data Encryption and Privacy Protection Technologies

In response to the problem of insufficient security of image data, Zeng et al. proposed a quantum-classical hybrid encryption method based on parameterized pixel ratio, which has high security and strong resistance to attacks [10]. To address the issue of insufficient security and access control for big data in the cloud environment, Singh et al. designed a cryptographic and access control model based on DNA computing. This scheme improves the security and anti-attack capability of big data in the cloud environment [11]. To enhance medical data security, Sharma et al. proposed a bit-level encryption method based on triple chaotic mapping and adaptive bald eagle search optimization, which can effectively resist various attacks and has high security and robustness [12]. Thenmozhi et al. proposed an attribute-based adaptive homomorphic encryption method for big data security and privacy protection. The method combines optimization algorithms to select key parameters and outperforms traditional methods in terms of encryption efficiency and key generation, and has the potential for real-time application [13].

Existing encryption methods perform well in protecting static data but mostly rely on trusted third parties or are suitable for centralized scenarios. This research adopts a combination of differential privacy and secret sharing, achieving lightweight perturbation and gradient segmentation transmission at the terminal, thereby blocking privacy leakage pathways without the need for a trusted third party.

2.2 Blockchain and Distributed Trust Mechanisms

In response to the problem of low data verification efficiency and high communication overhead of traditional blockchain in Industrial IoT (IIoT), Wang et al. proposed a secure storage mechanism based on partition vector commitment, which significantly reduces communication loss and optimizes storage space, effectively improving system security and stability [14]. To address the issues of low data verification efficiency and high proof overhead in traditional blockchains for IIoT, Li et al. proposed a secure storage method based on partition vector commitment. This method significantly reduces the data verification overhead and communication loss of traditional blockchains in IIoT and optimizes storage and security [15]. To address the issue of performance degradation caused by non-independent and identically distributed data in federated learning, Zhang et al. proposed a layered federated learning framework based on blockchain, which effectively improves

the performance of federated learning models under non-independent and identically distributed data [16].

Blockchain technology enhances system credibility through decentralization, but its high latency and computational overhead limit its deployment at the edge. This research adopts a cloud-edge-end collaborative architecture, shifting privacy protection to the terminal end, with the edge side undertaking aggregation tasks, thereby reducing latency while ensuring security and avoiding the additional overhead of blockchain.

2.3 Attack Detection and Intrusion Prevention

To address the threat of network attacks to data security in IoT systems, Liang et al. proposed a multi-level intrusion detection model that integrates multi-wavelet learning and Transformer. Experiments show that it has good predictive performance in both cloud and edge terminal layers [17].

Existing intrusion detection methods perform well in IoT scenarios, but most assume data is independent and identically distributed, making them susceptible to interference in non-IID scenarios. This research introduces a linked detection mechanism combining Mixup data augmentation and gradient clustering, effectively mitigating the interference caused by data heterogeneity through sample fusion and update trajectory analysis.

2.4 Federated Learning-specific Security Frameworks

In response to the challenges of interpretability and data privacy faced by federated learning in smart healthcare, Zhao et al. proposed a scheme that balances interpretability and security aggregation, which performs well in terms of interpretability, security, and efficiency [18].

Existing federated learning security solutions have made progress in privacy protection but suffer from issues such as delayed poisoning detection and static budget allocation. This research constructs a multi-level closed-loop system, achieving real-time detection through gradient clustering, rapid remediation through federated unlearning, and dynamic budget allocation through game theory, thereby realizing systematic protection throughout the entire lifecycle.

In summary, unlike existing research that predominantly focuses on defending against single types of attacks (e.g., encryption or detection), this study constructs a full-lifecycle protection system spanning from data sources to global optimization. For instance, the hybrid quantum-classical encryption

method proposed by Zeng et al. enhances image data security but does not account for communication efficiency or adaptability to heterogeneous scenarios. The PSFL framework introduced by Li et al. ensures privacy and model security but relies on static privacy budget allocation. The explainable federated learning approach by Zhao et al. performs well in medical data sharing but lacks active detection of poisoning attacks. In contrast, this study advances privacy protection through cloud-edge-end collaboration, employs a combined Mixup and gradient clustering mechanism for active poisoning attack detection, and introduces game theory for dynamic privacy budget optimization. This achieves a synergistic enhancement across privacy, security, and utility. Experimental results demonstrate that the proposed method attains an accuracy of 76.9% in complex dynamic scenarios, significantly outperforming single-module solutions.

3 Methods and Materials

3.1 Federated Learning Methods Based on Privacy Enhancement of the Data Source and Transmission Process

Faced with the increasingly complex deployment of federated learning in the big data environment, its security and privacy face multiple challenges [19]. Existing research focuses on defending against single attacks or using isolated techniques, lacking a systematic solution for dynamic data streams, heterogeneous terminals and persistent threats [20, 21]. Especially at the data source, traditional methods are difficult to simultaneously protect the real-time protection of original sensitive information and the efficiency of global model training, resulting in privacy leakage risks throughout the entire distributed learning process. To this end, a cloud-edge-end collaborative architecture was designed. By implementing lightweight perturbations on the terminal and performing secure aggregation on the edge, privacy protection is moved forward and computational load is optimized, laying a reliable foundation for subsequent federated training. The federated GAN with collaborative privacy enhancement (GAN-CPE) architecture is shown in Figure 1.

In Figure 1, the GAN-CPE works collaboratively around a three-layer architecture of cloud, edge, and terminal. The terminal device, as the data source, first performs privacy enhancement processing on the original data. It adds appropriate noise to generate perturbation data through a differential privacy Gaussian mechanism, thereby reducing the risk of privacy leakage from the root. The formula for the differential privacy Gaussian mechanism

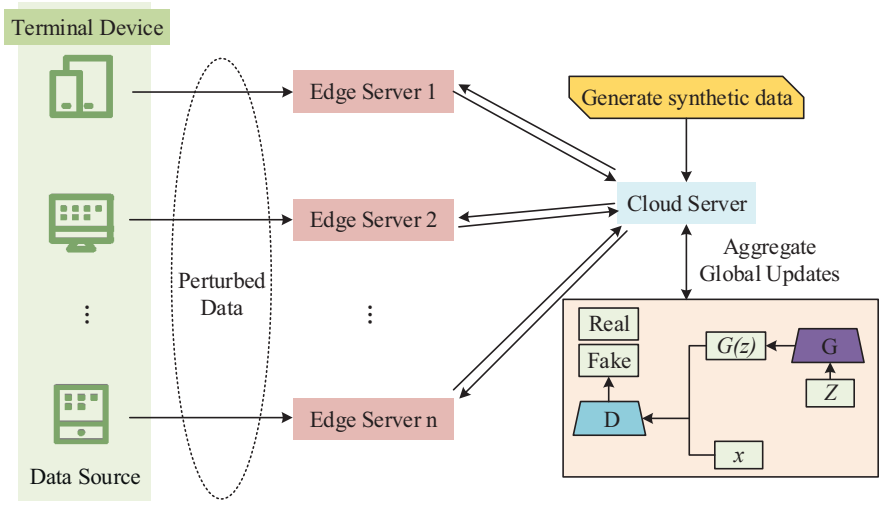


Figure 1 GAN-CPE architecture.

is shown in Equation (1).

$$M(D) = f(D) + \Re(0, \sigma^2 S^2) \tag{1}$$

In Equation (1), $M(D)$ represents the perturbed data, $f(D)$ represents the query function on the original data D, $\Re(0, \sigma^2 S^2)$ represents Gaussian noise, σ represents the noise scale parameter, and S represents the sensitivity of the query function f. Data protected by differential privacy is uploaded to the edge server for aggregation to form a privacy dataset. The server uses edge computing power to train the generator and discriminator locally to share the computing pressure of the terminal. By implementing differential privacy perturbation at the terminal, raw data remains local, effectively blocking pathways for privacy leakage. Edge-side aggregation reduces the computational burden on the cloud, improves training efficiency, and provides a more secure intermediate representation for subsequent federated learning. This cloud-edge-end collaborative design achieves a dual optimization of privacy protection and computational efficiency.

After local training is completed, the edge nodes upload the gradients to the cloud. The cloud aggregates all gradients and updates the global model, and then sends the new parameters to each node for the next round of federated learning iteration [22]. To reduce the communication burden and privacy risks of gradient transmission, the research proposes the federated-aware algorithm (FAA), which integrates secret sharing encryption and adaptive

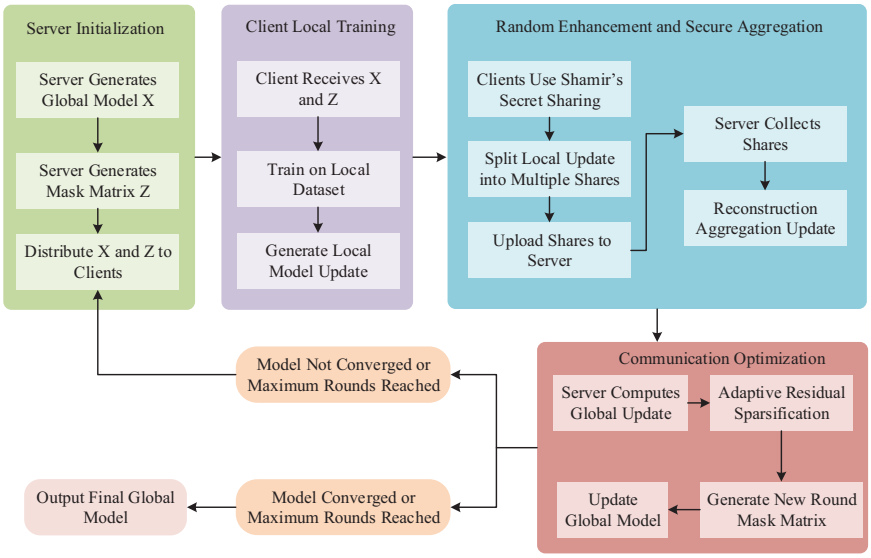


Figure 2 FAA for communication optimization.

sparsity compression technology. The framework of this FAA is shown in Figure 2.

As shown in Figure 2, the FAA is based on the federated learning paradigm. The client completes model training locally, and the original sensitive data is not uploaded, reducing the risk of privacy leakage from the source. To address gradient leakage attacks, the client encrypts local model updates using secret sharing technology, transmitting only the segmented secret share instead of the gradient plaintext, ensuring the confidentiality of transmission. To address the issue of some devices being offline or asynchronous, the Shamir thresholding scheme is adopted, as shown in Equation (2).

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1} \tag{2}$$

In Equation (2), $P(x)$ represents the polynomial used to generate the secret share, a_0 represents the secret value, a_1 represents the randomly selected coefficient, x represents the share index, and k represents the mini number of shares required to reconstruct the secret. Global model updates can be reconstructed once the number of participating clients exceeds a threshold, ensuring training continuity. For honest but curious servers, they can only obtain the aggregated global results and cannot separate and identify specific updates from individual clients. The algorithm incorporates adaptive sparsity

techniques, as shown in Equation (3).

$$ML(g) = \{g_i \,|\, |g_i| \in ML(|g|)\} \tag{3}$$

In Equation (3), g represents the gradient vector, $ML(g)$ represents the L elements with the largest absolute value in the vector g, and i represents the index of the element in the gradient vector g. Selective transmission of key parameters and indices of the model significantly reduces communication overhead and achieves a balance between communication efficiency and model performance while strengthening privacy protection, providing reliable technical support for scenarios such as personal big data health monitoring [23].

3.2 Federated Learning-based Poisoning Attack Defense Method Based on Active Detection and Post-remediation

Current defense strategies for federated learning mainly focus on anomaly detection on the server side, but these methods are often ineffective in the face of highly covert data poisoning attacks, especially in non-independent and identically distributed scenarios [24]. Malicious clients use carefully designed local data poisoning models, making it difficult for traditional defenses to maintain the robustness of the global model while protecting privacy, and the coordination of security and performance faces challenges [25]. To this end, this study proposes a defense mechanism that links client-side data augmentation with server-side gradient clustering to build an end-to-end active protection system from local to global. The data augmentation-driven collaborative defense (DACD) process is shown in Figure 3.

As shown in Figure 3, all clients perform data augmentation operations before local training, using Mixup technology to perform linear interpolation on the local data. Mixup data augmentation is shown in Equation (4).

$$\begin{cases} \tilde{m} = \lambda m_i + (1-\lambda) m_j \\ \tilde{n} = \lambda n_i + (1-\lambda) n_j \end{cases} \tag{4}$$

In Equation (4), $\tilde{m}$ represents the mixed sample data, $\tilde{n}$ represents the mixed label, m_i and m_j represent the two original sample data, n_i and n_j represent the corresponding original labels, λ represents the mixing coefficient, and i and j represent the index of the element. The client uses Mixup data augmentation to generate a fused dataset, which alleviates the effects of non-independent and identically distributed data and suppresses

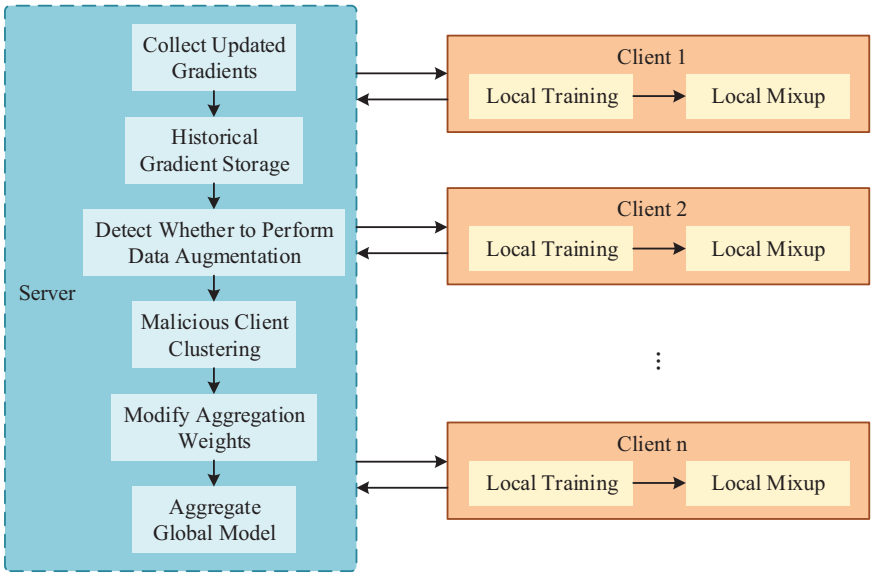


Figure 3 DACD mechanism.

the interference of poisoned data. After training, the client only uploads gradient updates and retains the original data locally [26]. The server builds a gradient history library, uses cluster analysis to update the trajectory to identify malicious nodes, and dynamically reduces their aggregation weights to avoid detection behavior exposure [27]. The resulting collaborative defense system can effectively improve the ability of federated learning to resist poisoning attacks. In view of the continuous harm of data poisoning in the global model, the study proposes an active defense method based on approximate forgetting, which achieves deep security protection by accurately identifying and eliminating the influence of malicious clients. The complete process of this two-stage federated unlearning (TS-FU) is shown in Figure 4.

As shown in Figure 4, the process begins with the client completing local model training and uploading updated parameters. The server collects and stores this updated information and analyzes it in conjunction with the global model test results. The server first locates the key layers in the model based on the gradient weight norm, then calculates the distribution distance of each client in that layer and identifies potential malicious clients by combining the anomaly threshold. Once detected, the server calculates the gradient residual

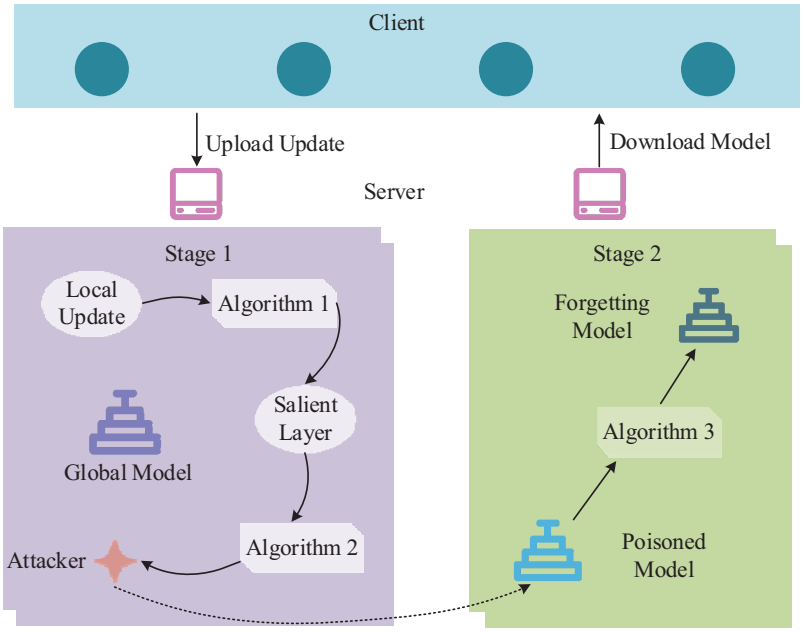


Figure 4 TS-FU flowchart.

of the removed client based on the historical gradient update and adjusts the global model weights proportionally. The approximate forgetting gradient update is shown in Equation (5).

$$w_{t+1} = w_t - \eta \cdot \sum_{a \in C,E} \nabla \ell(w_t, D_a) + \Re(0, \sigma^2) \tag{5}$$

In Equation (5), w represents the global model weights, t represents the update round, η represents the learning rate, C represents the set of clients participating in the current round of training, E represents the set of malicious clients, and $\nabla \ell(w_t, D_a)$ represents the gradient of client a on the original data D_a. Simultaneously, Gaussian noise that meets the requirements of differential privacy is introduced to generate an approximate forgotten model that is statistically indistinguishable from the retrained model. If no malicious client is detected, the server directly outputs the aggregated global model for the next round of training. This method ensures the model's security and robustness by identifying attacks in real time and eliminating malicious influences.

3.3 A Federated Learning Privacy and Security Framework Based on System Balancing and Integration Optimization

With the deepening application of federated learning in dynamic heterogeneous big data environments, its security and privacy protection face new challenges [28]. Existing methods mostly adopt static and uniform protection strategies, which are difficult to adapt to the changing privacy needs and data sensitivity differences of different users over time, often leading to an imbalance between model utility and individual privacy rights [29]. To balance personalized privacy and model utility, this study introduces game theory ideas for dynamic optimization and achieves the optimal protection and utility balance acceptable to both parties by simulating the policy interaction between users and servers. The game theory-based personalized privacy balancing (GT-PPB) architecture is shown in Figure 5.

As shown in Figure 5, the client sets the protection level independently according to its own privacy preferences. The local perturbator matches the corresponding privacy budget and perturbs the local model update through a personalized local differential privacy mechanism. The personalized local differential privacy is shown in Equation (6).

$$\hat{g}_a = g_a + \mathrm{Lap}\left(\frac{\Delta f}{\epsilon_a}\right) \tag{6}$$

In Equation (6), $\hat{g}_a$ represents the gradient after client perturbation, g_a represents the original gradient, Lap represents Laplace noise, Δf represents the sensitivity of the gradient function, and ϵ_a represents the client's privacy budget. After generating the perturbation report, it is uploaded to the server.

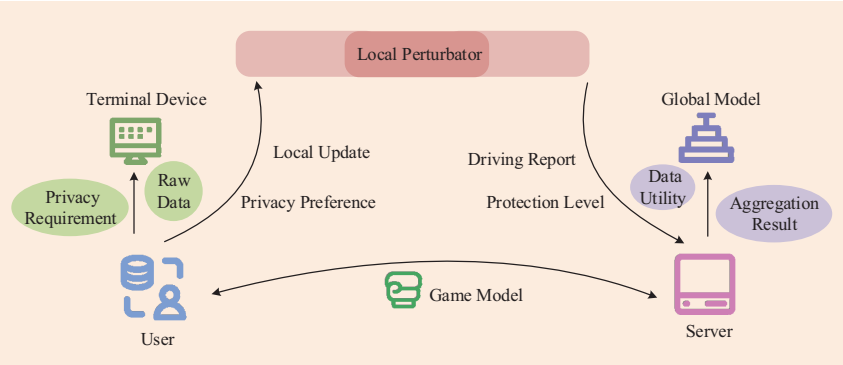


Figure 5 GT-PPB architecture.

The server does not need to be assumed to be trustworthy; it can only receive the perturbation report and cannot obtain the original data, thus blocking the risk of privacy inference from the transmission link. The server first clusters and grades the user's privacy preferences, calculates the proportion of each level and allocates the global privacy budget, and then processes the perturbation report according to the corresponding level decoding rules. The decoded update is aggregated to generate a new round of global model and fed back to the client. The algorithm coordinates the distribution of benefits between the two parties based on static Bayesian game, and the Bayesian game model is shown in Equation (7).

$$U_a(\epsilon_a, \theta) = \alpha \cdot \mathrm{Utility}(\theta) - \beta \cdot \mathrm{PrivacyCost}(\epsilon_a) \tag{7}$$

In Equation (7), U_a represents the client's utility function, θ represents the global model performance index, $\mathrm{Utility}(\theta)$ represents the utility brought by model performance, $\mathrm{PrivacyCost}(\epsilon_a)$ represents the privacy protection cost, and α and β represent trade-off coefficients, balancing utility and cost. The optimal balance between privacy protection and model utility is achieved using Nash equilibrium, satisfying users' personalized privacy needs, avoiding budget waste, and enabling the server to achieve better model performance, providing a flexible and efficient privacy–utility balancing solution for multi-scenario federated learning. To this end, based on a hierarchical design from data source to global governance, this study integrates the above key technologies and constructs a collaborative enhancement federated learning research system. Its complete architecture and module interaction relationships are shown in Figure 6.

As shown in Figure 6, this framework constructs a hierarchical federated learning security enhancement system. The bottom layer is responsible for data and communication privacy, employing dynamic differential privacy protection at the data source and combining secret sharing and adaptive sparsity techniques to ensure the confidentiality and efficiency of gradient transmission. The middle layer is the core training and active defense layer, which, on top of basic aggregation, integrates real-time detection and a deep repair mechanism based on salient layer identification and near-forgetting to form a closed loop for monitoring and eliminating poisoning attacks. The top layer is a privacy utility dynamic balancing layer, which uses a game theory model to dynamically allocate the privacy budget according to user preferences, systematically reconciling the contradiction between model utility and privacy protection. The technologies of each layer work synergistically to form a complete solution that balances privacy, security, and performance.

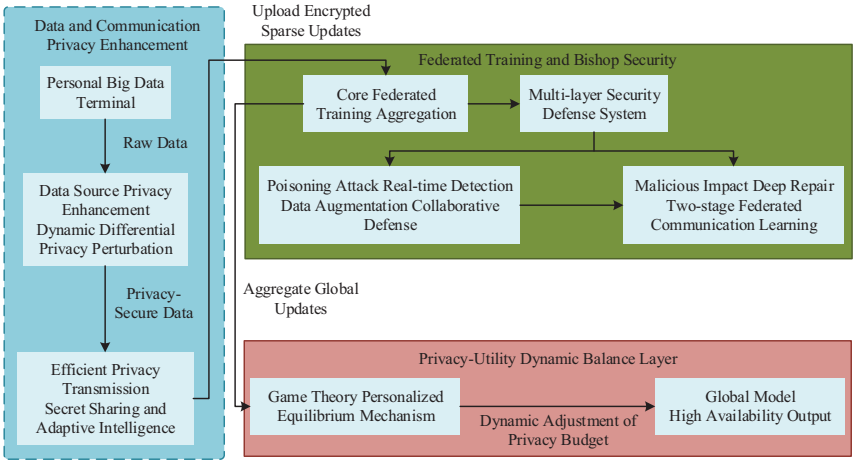


Figure 6 Multi-level federated learning security framework.

The research hypothesizes two types of threats in federated learning systems: first, external attackers stealing private information through eavesdropping or gradient reconstruction; second, internal malicious clients corrupting the global model by crafting poisoned data. The poisoning attack model considered in this research is as follows: an attacker controls a subset of malicious clients who tamper with local training data or gradient updates, aiming to degrade the accuracy of the global model or inject backdoors. It is assumed that the proportion of malicious clients does not exceed 30%, and the server is "honest-but-curious", i.e., it follows the protocol but attempts to infer private information from the received messages.

The system environment exhibits typical heterogeneity, including the non-independent and identically distributed (Non-IID) nature of client data and variations in computational capabilities. The proposed framework, through a cloud-edge-end collaborative architecture, supports clients dynamically joining and leaving, and is theoretically scalable to thousands of nodes.

4 Results

4.1 Performance Testing of Privacy Enhancement Methods at the Data Source and During Transmission

To assess the security and privacy protection methods of federated learning, this study selected the following open-source datasets for experiments:

Table 1 Detailed information of the testing platform

	Configure	Parameter
Hardware configuration	CPU	Intel Xeon Gold 63xx
	GPU	NVIDIA RTX A6000
	Storage	2TB NVMe SSD+4TB HDD
Software configuration	Operating system	Ubuntu Server 22.04 LTS
	Virtualization & container tools	Docker 24.x + Kubernetes 1.28
	Federated learning framework	PySyft 0.7.x/TensorFlow Federated 0.56/ FATE 2.0
	Programming language	Python 3.10 + PyTorch 2.0/TensorFlow 2.12
	Attack simulation tools	ART (Adversarial Robustness Toolbox)/ FLSim (Facebook)
	Privacy computing library tools	OpenDP Library (v0.9.0)/PySyft (v0.8.0)/TenSEAL (v0.3.0)

CIFAR-10 (University of Toronto): containing 60,000 32 $\times$ 32 color images, suitable for privacy and communication benchmark testing; FEMNIST (derived from the LEAF benchmark): containing 800,000 handwritten character images, divided among 3500 real users, exhibiting natural non-independent and identically distributed characteristics, suitable for personalized privacy and heterogeneous training verification. Experimental parameters were uniformly set as follows: learning rate 0.01, batch size 32, total training epochs 50, and client-side local training epochs 5. Privacy protection strength was measured by the ε value, data heterogeneity was quantified by the degree of non-independent and identically distributed characteristics, and the poisoning ratio was adjusted by the proportion of malicious clients to reflect real-world scenarios. The experiments were conducted on a simulation platform built on Ubuntu Server 22.04 LTS system, with specific configurations shown in Table 1. All datasets used were publicly available resources.

To ensure the reliability and statistical significance of the experimental results, all experiments were independently repeated five times, with the average value taken as the final result, and the standard deviation was calculated to measure the range of fluctuation. The data points in the figures represent the mean values. Comparative results were validated for significance using a paired t-test ($p < 0.05$). All experiments were conducted with fixed random seeds to ensure reproducibility.

The gradient leakage attack employed the deep gradient leakage method, where the attacker optimized the input to make the reconstructed gradient match the true gradient, thereby recovering the original data. The privacy protection success rate was defined as the proportion of instances where the attacker failed to successfully reconstruct the original data, with a threshold set at a peak signal-to-noise ratio (PSNR) of below 20 dB between the reconstructed and original images. Each attack scenario was repeated 10 times, and the results were averaged.

To test the comprehensive performance of GAN-CPE and FAA methods in terms of privacy protection, communication efficiency, and resistance to gradient leakage attacks, a comparison was made with traditional federated learning (Vanilla FL). The experiment tested model accuracy, communication overhead, and the privacy protection success rate by adjusting variables such as privacy protection strength (ε value), number of clients, and number of attack iterations. The test results are shown in Figure 7.

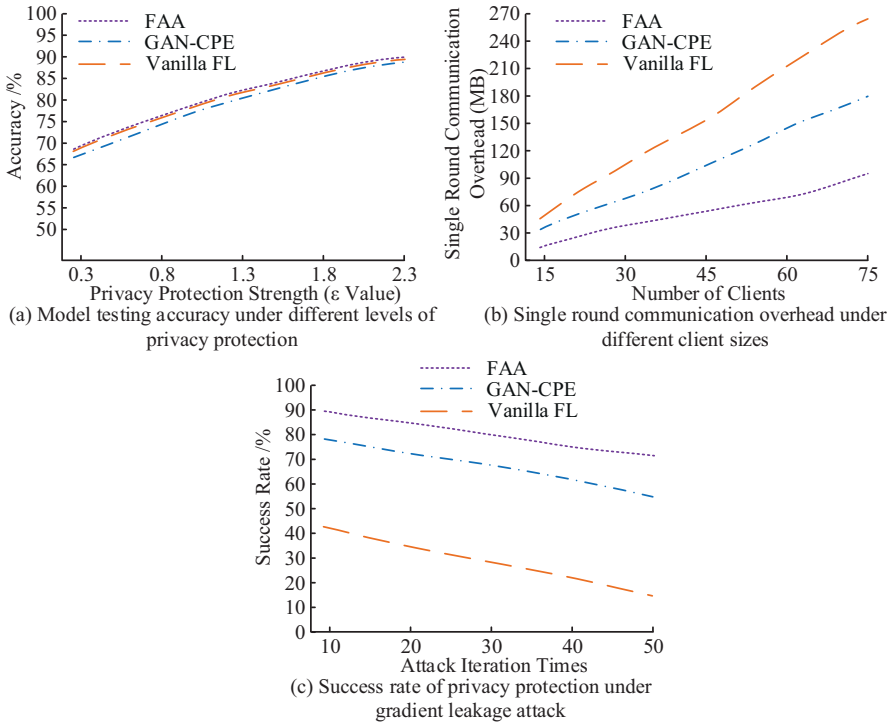


Figure 7 Privacy enhancement method performance test.

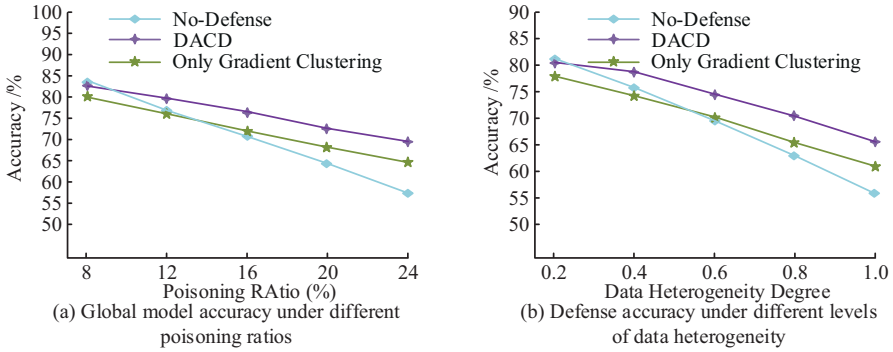
As shown in Figure 7(a), the accuracy of all three methods steadily increased with increasing ε. When $\varepsilon = 2.3$, the accuracy of the FAA reached 89.8%, and that of GAN-CPE was 87.5%, with a difference of only 2.3 percentage points, which conforms to the privacy-accuracy trade-off. Figure 7(b) shows that when the number of clients was 75, the overhead of Vanilla FL, collaborative federated learning, and FAAs were 264.1 MB, 178.0 MB, and 91.9 MB, respectively; when the number of clients was 15, the overhead of FAA was only 34.8% of that of Vanilla FL, highlighting its sparse transmission advantage. According to Figure 7(c), as the number of attack iterations increased, the privacy protection success rate of Vanilla FL decreased the fastest, reaching only 16.8% after 50 iterations. Collaborative federated learning maintained 55.7% by relying on differential privacy noise; while FAA achieved a success rate of 71.2% through secret sharing encryption, reaching as high as 89.2% after 10 iterations, demonstrating stronger resistance to attacks.

4.2 Performance Testing of Poison Attack Active Detection and Post-repair Defense Methods

The poisoning attack adopts a label-flipping strategy, where malicious clients randomly alter the labels of their local training data to other classes (non-original labels), aiming to degrade the accuracy of the global model. The poisoning proportion refers to the percentage of malicious clients relative to the total number of clients. For gradient clustering on the server side, the DBSCAN algorithm is used, with the anomaly threshold set at twice the standard deviation.

The experiment aimed to verify the defense effectiveness of DAD under different poisoning ratios and data heterogeneity, comparing it with no defense mechanism. The experiment tested the global model and defense accuracy by adjusting the proportion of malicious clients (poisoning ratio) and the degree of non-independent identical distribution (data heterogeneity). The test results are shown in Figure 8.

As shown in Figure 8(a), the accuracy of the unprotected model dropped significantly with increasing poisoning ratio, reaching only 57.6% at 24% poisoning. In contrast, the DCD method achieved an accuracy of 82.9% with 8% poisoning and maintains 69.7% with 24% poisoning, a 12.1 percentage point improvement over the unprotected model. This improvement is attributed to the dilution effect of the Mixup technique on the poisoned samples and the identification of malicious updates by gradient clustering.



(a) Global model accuracy under different poisoning ratios

(b) Defense accuracy under different levels of data heterogeneity

Figure 8 Active detection defense method performance test.

Figure 8(b) shows that the unprotected accuracy dropped to 56.7% with heterogeneity increasing to 1.0. The DCD accuracy reached 80.7% with heterogeneity of 0.2 and still maintained 66.8% with heterogeneity of 1.0, a 10.1 percentage point improvement over the unprotected model. This demonstrates that the method effectively mitigates the impact of non-independent and identically distributed data, improving the stability of defenses in heterogeneous environments.

To assess the performance of TS-FU in terms of model repair accuracy, privacy security, and repair efficiency, a comparison was made with traditional model retraining. By adjusting the number of malicious clients, the number of repair iterations, and the proportion of contaminated samples, the model repair accuracy, privacy leakage risk, and repair time were tested. The test results are shown in Figure 9.

As shown in Figure 9(a), the model's repair accuracy decreased with the increase in the number of malicious clients: when there were 3 malicious clients, the repair accuracy of traditional retraining was 83.5%, which is 2.3 percentage points higher than TS-FU. TS-FU adjusted the weights by gradient residuals to approximate the forgetting of malicious influences, avoiding full data retraining while maintaining slightly lower accuracy, thus balancing performance and efficiency. Figure 9(b) shows that TS-FU (including noise) had an accuracy of 2.3% after 1 iteration, which is 2.8 percentage points lower than Retrain. After 5 iterations, TS-FU had an accuracy of 1.7%, which is 2.6 percentage points lower than Retrain. Differential privacy noise further blocked the privacy inference path, demonstrating the privacy protection advantage of TS-FU. Figure 9(c) shows that when the pollution ratio was 10%, TS-FU took 18.6 seconds, only 41.1% of Retrain. When the pollution

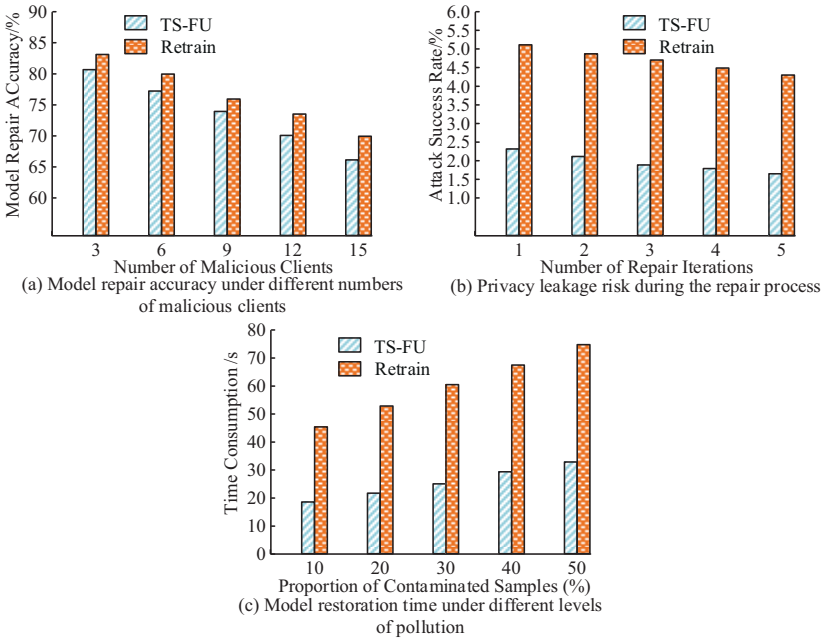


Figure 9 Post-repair defense method performance test.

ratio increased to 50%, TS-FU took 33.2 seconds, while Retrain took 74.8 seconds, highlighting the significant efficiency advantage of this method.

4.3 Privacy-utility Balance Framework and Overall Performance Testing

To validate the performance of GT-PPB in terms of both privacy budget utilization efficiency and privacy–utility balance, a comparison was made with a fixed privacy budget framework. By adjusting the total privacy budget and the number of collaborative learning rounds, the budget utilization efficiency and privacy–utility balance coefficient were tested respectively. The test results are shown in Figure 10.

As shown in Figure 10(a), the budget utilization efficiency of all three schemes gradually improved with the increase of the total privacy budget, demonstrating the significant advantage of the game theory approach. When the total privacy budget was 5.0, the budget utilization efficiency of game theory was 88.7%, which is 26.4 percentage points higher than that of a fixed privacy budget. When the budget was 15.0, the efficiency of game

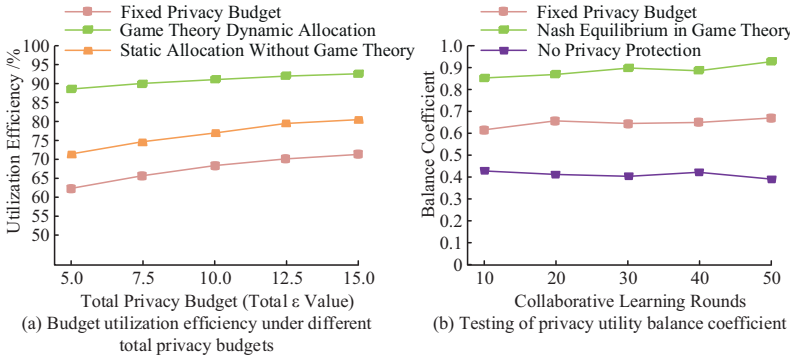


Figure 10 Game theory-based personalized privacy balancing performance test.

theory was 93.1%, while that of a fixed privacy budget was 71.8%. Game theory significantly improved utilization efficiency by dynamically allocating the budget and accurately matching user privacy preferences. As shown in Figure 10(b), the privacy–utility balance coefficient of the game theory framework continuously increased and remained at its highest level as the number of collaborative learning rounds increased. The coefficient was 0.85 at round 10, which is 0.23 higher than that of the fixed budget scheme and reached 0.91 at round 50. This framework achieved a long-term optimal balance by dynamically coordinating the utility and privacy costs of the model through Nash equilibrium. The experiment compared the comprehensive performance of the multi-layered federated learning security framework (integrating privacy enhancement, poisoning defense, and privacy–utility balance mechanism) with traditional federated learning (Vanilla FL), privacy enhancement module only (PE-module), and poisoning defense module only (AD-module) in different complex scenarios.

The model accuracy was tested in the following scenarios: a conventional scenario (no attack + uniform data distribution), a privacy attack scenario (gradient leakage + heterogeneous data), a poisoning attack scenario (20% poisoning + heterogeneous data), a mixed attack scenario (gradient leakage + 15% poisoning), and a complex dynamic scenario (hybrid attack + dynamic changes in user privacy preferences). The test results are shown in Table 2.

According to Table 2, the accuracy of all four schemes decreased with increasing scenario complexity and attack intensity, but the multi-layered security framework consistently achieved the highest accuracy. In conventional scenarios, the Vanilla FL achieved an accuracy of 89.6%,

Table 2 Multi-level federated learning performance metrics under different test scenarios

Test Scenario	Vanilla FL	PE-module	AD-Module	Multi-layer Framework
Conventional scenario	89.6	87.3	88.5	88.9
Privacy attack scenario	62.4	81.7	60.3	83.2
Poisoning attack scenario	58.9	56.4	79.2	80.5
Mixed attack scenario	53.7	72.5	71.8	78.6
Complex dynamic scenario	49.2	68.3	67.5	76.9

slightly higher than the multi-layered framework's 88.9%. However, in complex and dynamic scenarios, the multi-layered framework achieved an accuracy of 76.9%, which was 27.7, 8.6, and 9.4 percentage points higher than Vanilla FL, the privacy enhancement module alone, and the poisoning defense module alone, respectively. This framework integrates the collaborative advantages of multiple modules, using privacy enhancement to resist gradient leakage, poisoning defense to address data pollution, and a game theory framework to balance privacy and utility, thereby achieving comprehensive protection in complex scenarios and demonstrating the technical advantages of hierarchical integration.

5 Conclusion

This study addresses the challenges of federated learning in big data environments, including privacy breaches, poisoning attacks, high communication overhead, and the imbalance between privacy and utility. It constructs a multi-layered security system based on a "protection–defense–optimization" approach. The research has established three major technological modules: achieving source-level privacy enhancement and communication optimization through FAA and GAN-CPE, detecting and repairing poisoning attacks via DACD and TS-FU, and dynamically balancing privacy and utility using GT-PPB. In terms of technology selection, emphasis is placed on scenario adaptability: the differential privacy Gaussian mechanism is chosen for its ability to achieve lightweight perturbations on terminals, meeting local deployment requirements; Shamir's secret sharing is adopted to enable gradient segmentation transmission without relying on a trusted third party, thereby blocking privacy leakage paths at the source; Mixup data augmentation combined with gradient clustering linkage detection is introduced to effectively mitigate false positives in poisoning attacks under non-IID scenarios through sample fusion and update trajectory analysis; privacy

budgets are dynamically allocated based on game theory to overcome the utility loss caused by static allocation. The federated learning security system constructed in this study achieves synergistic optimization in three aspects: privacy enhancement, attack defense, and utility balance. Experiments validate its robustness and efficiency in complex dynamic scenarios: with 75 clients, communication overhead is only 91.9 MB; under 24% poisoning attacks, accuracy remains at 69.7%; privacy leakage risk is minimized to 1.7%; and the privacy–utility balance coefficient reaches 0.91. The research demonstrates that this system provides systematic technical support for the practical deployment of federated learning in sensitive domains such as healthcare and finance. However, the performance of existing methods still has room for improvement in extremely heterogeneous scenarios, and issues related to communication scheduling and edge computing load become increasingly prominent when facing a thousand-node scale. Future work will focus on optimizing the noise adaptation mechanism to enhance robustness in extreme scenarios, introducing hierarchical aggregation and asynchronous update strategies combined with dynamic optimization of sparsification compression thresholds to reduce the overhead of edge nodes and improve scalability in large-scale dynamic networks. Simultaneously, we will explore integration with blockchain technology to further enhance system decentralization and tamper resistance, promoting the secure and reliable deployment of federated learning in a broader range of scenarios.

References

[1] Jin X, Ma C, Luo S, Zeng P, Wei Y. Two-stage client selection scheme for blockchain-enabled federated learning in IoT. Computers, Materials & Continua, 2024, 81(11): 2317–2336. DOI:10.32604/cmc.2024.055344.

[2] Tang Y. Risk assessment of distributed network data security based on Simhash algorithm. International Journal of Industrial Engineering-Theory Applications and Practice, 2024, 31(5): 950–966. DOI:10.23055/ijietap.2024.31.5.9995.

[3] Somsuk K. Enhanced Algorithm for Recovering RSA Plaintext when Two Modulus Values Share at least One Common Prime Factor. Journal of Cyber Security and Mobility, 2025, 14(2): 433–456. DOI:10.13052/jcsm2245-1439.1427.

[4] Huang P, Liao G, Ren J. The application of AES-SM2 hybrid encryption algorithm in big data security and privacy protection. International

Journal of Advanced Computer Science and Applications, 2024, 15(6): 989–997. DOI:10.14569/IJACSA.2024.01506101.

[5] Dai X, Yu Z, Liang C, Gao C, He Q, Wu D, et al. Detecting novel malware classes with a foundational multi-modality data analysis model. Data Intelligence, 2024, 6(4): 968–993. DOI:10.3724/2096-7004.di.2024.0056.

[6] Mohammed M A, Wahab H B A. Enhancing IoT data security with lightweight blockchain and Okamoto Uchiyama homomorphic encryption. Computer Modeling in Engineering & Sciences, 2024, 138(2): 1731–1748. DOI:10.32604/cmes.2023.030528.

[7] Huang W, Chen Y, Jing D, Feng J, Han G, Zhang W. A multicloud collaborative data security sharing scheme with blockchain indexing in industrial internet environments. IEEE Internet of Things Journal, 2024, 11(16): 27532–27544. DOI:10.1109/JIOT.2024.3398774.

[8] Bai X, Bai Y. Equilibrium Strategy of Attack and Defense in Computer Networks Based on Markov Signal Game Theory. Journal of Cyber Security and Mobility, 2025, 14(1): 127–154. DOI:10.13052/jcsm2245-1439.1416.

[9] Odeh A, Taleb A A. Robust network security: a deep learning approach to intrusion detection in IoT. Computers, Materials & Continua, 2024, 81(12): 4149–4169. DOI:10.32604/cmc.2024.058052.

[10] Zeng L, Chang Y, Zhang X, Xue W, Zhang S, Yan L, et al. Cryptographic enhancement of image data security through quantum-classical hybrid encryption with parameterized pixel ratios. Quantum Information Processing, 2024, 23(7): 1–25. DOI:10.1007/s11128-024-04431-9.

[11] Singh A, Kumar A, Namasudra S. DNACDS: Cloud IoE big data security and accessing scheme based on DNA cryptography. Frontiers of Computer Science, 2024, 18(1): 157–170. DOI:10.1007/s11704-022-2193-3.

[12] Sharma S R, Singh B, Kaur M. Improvement of medical data security using SABES optimization algorithm. The Journal of Supercomputing, 2024, 80(9): 12929–12965. DOI:10.1007/s11227-024-05937-w.

[13] Thenmozhi R, Shridevi S, Mohanty S N, Garcia Diaz V, Gupta D, Tiwari P, et al. Attribute-based adaptive homomorphic encryption for big data security. Big Data, 2024, 12(5): 343–356. DOI:10.1089/big.2021.0176.

[14] Wang J, Huang G, Sherratt R S, Huang D, Ni J. Data secure storage mechanism for IIoT based on blockchain. Computers, Materials & Continua, 2024, 78(3): 4029–4048. DOI:10.32604/cmc.2024.047468.

[15] Li J, Tian Y, Zhou Z, Xiang A, Wang S, Xiong J, et al. PSFL: ensuring data privacy and model security for federated learning. IEEE Internet of Things Journal, 2024, 11(15): 26234–26252. DOI:10.1109/JIOT.2024.3394168.

[16] Zhang F, Zhang Y, Ji S, Han Z. Secure and decentralized federated learning framework with non-IID data based on blockchain. Heliyon, 2024, 10(5): e27176. DOI:10.1016/j.heliyon.2024.e27176.

[17] Liang P, Yang L, Xiong Z, Zhang X, Liu G. Multilevel intrusion detection based on transformer and wavelet transform for IoT data security. IEEE Internet of Things Journal, 2024, 11(15): 25613–25624. DOI:10.1109/JIOT.2024.3369034.

[18] Zhao L, Xie H, Zhong L, Wang Y. Explainable federated learning scheme for secure healthcare data sharing. Health Information Science and Systems, 2024, 12(1): 1–14. DOI:10.1007/s13755-024-00306-6.

[19] Shastri V H, Pragathi C. Data security using crypto bipartite graph theory with modified Diffie-Hellman algorithm. Wireless Personal Communications, 2024, 139(4): 1905–1926. DOI:10.1007/s11277-024-11679-y.

[20] Lill B, Sauerwein C, Mexis N, K. Langner. A Comprehensive Review of Information Security Research regarding SMEs and Future Directions. Journal of Cyber Security and Mobility, 2025, 14(5): 1245–1288. DOI:10.13052/jcsm2245-1439.1459.

[21] Wang J. Identification of SQL Injection Security Vulnerabilities in Web applications Based on Binary Code Similarity. Journal of Cyber Security and Mobility, 2024, 13(6): 1239–1262. DOI:10.13052/jcsm2245-1439.1361.

[22] Muthubalaji S, Muniyaraj N K, Rao S P V S, Kavitha T, Mohan P R, Somasundaram T, et al. An intelligent big data security framework based on AEFS-KENN algorithms for the detection of cyber-attacks from smart grid systems. Big Data Mining and Analytics, 2024, 7(2): 399–418. DOI:10.26599/BDMA.2023.9020022.

[23] Basha U S, Gupta S K, Alawad W, Kim S, Bharany S. Fortifying healthcare data security in the cloud: a comprehensive examination of the EPM-KEA encryption protocol. Computers, Materials & Continua, 2024, 79(5): 3397–3416. DOI:10.32604/cmc.2024.046265.

[24] Alyoubi A A. Enhancing data security in mobile ad-hoc network (MANETs) using trust-based approach with RSSI and fuzzy logic. Mobile Networks and Applications, 2024, 29(SI6): 2030–2046. DOI:10.1007/s11036-024-02336-6.

[25] Sudarsa D, Rao A N, Sivakumar A P. Data security optimization at cloud storage using confidentiality-based data classification. International Journal of Advanced Computer Science and Applications, 2024, 15(5): 699–709. DOI:10.14569/IJACSA.2024.0150570.

[26] Lu H, Chen W, Zhou C, Wu H, Lyu F, Shen X. A two-dimensional hybrid federated learning framework for secure data cooperation of multiple network service providers. IEEE Wireless Communications, 2024, 31(5): 215–222. DOI:10.1109/MWC.018.2300534.

[27] Verma P, Tripathi V, Pant B. Secure hashgraph for healthcare: strengthening privacy and data security in patient records. IEEE Transactions on Consumer Electronics, 2024, 70(1): 1205–1213. DOI:10.1109/TCE.2024.3370737.

[28] Waqas M, Naseem A. Artificial intelligence in sustainable industrial transformation: a comparative study of industry 4.0 and industry 5.0. FinTech and Sustainable Innovation, 2025, 1: A2–A2. DOI:10.47852/bonviewFSI52025321.

[29] Rehman M U. Quantum-enhanced chaotic image encryption: strengthening digital data security with 1-D sine-based chaotic maps and quantum coding. Journal of King Saud University-Computer and Information Sciences, 2024, 36(3): 101980. DOI:10.1016/j.jksuci.2024.101980.

Biography

Bin Zhou, female, born in January 1981, from Zibo City, Shandong Province, Han ethnicity, received her bachelor's degree in computer science and technology from Harbin University of Commerce in 2003, and went on to earn a master's degree in software engineering from Shandong University in 2006, with a research focus on software technology and big data technology. From 2003 to 2025 she was employed at Zibo Vocational College. From 2025 to present she has been employed at Zibo Polytechnic University. She has published 4 academic papers and 1 academic textbook.