Frequency-Domain Signatures for Proactive Defense Against Model Poisoning Attacks in Federated Learning

Fangjie Hu, Aiqing Zhang, *Member, IEEE*, Meng Li, *Senior Member, IEEE*, and Chen Wang, *Senior Member, IEEE*

***Abstract*—Federated Learning enables decentralized model training without exposing raw data, but remains fundamentally vulnerable to poisoning attacks from malicious clients. Existing defenses rely heavily on passive anomaly detection, honest majority assumptions, or unrealistic statistical priors, making them ineffective against adaptive and stealthy adversaries. In this paper, we propose SpecShield, a proactive defense mechanism that actively probes client models through calibrated adversarial perturbations. By leveraging the Fast Gradient Sign Method on the server side, SpecShield elicits dynamic response patterns from each client. These responses are then analyzed in the frequency domain using the Discrete Wavelet Transform. These frequency-domain features uncover distinctive response patterns between benign and malicious clients, enabling robust detection of model poisoning attacks in both non-IID environments and Byzantine majority scenarios. We further derive theoretical upper bounds on perturbation magnitudes to guarantee detection accuracy while preserving benign client performance. Through extensive experiments conducted on real-world datasets under six state-of-the-art poisoning attacks, SpecShield consistently outperforms existing defenses in both detection accuracy and model robustness. Our results demonstrate that active perturbation-induced profiling provides a new dimension for securing federated learning against sophisticated adversarial threats.**

Received 8 August 2025; revised 26 March 2026; accepted 28 April 2026. Date of publication 4 May 2026; date of current version 8 May 2026. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 62372149, Grant U23A20303, Grant 62572168, and Grant 62272183; in part by the Key Research and Development Program of Hubei Province under Grant GJHZ202500049; in part by Anhui Provincial Natural Science Foundation under Grant 2508085MF151; in part by the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education of China, under Grant BigKEOpen2025-04; and in part by the Open Foundation of the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. The associate editor coordinating the review of this article and approving it for publication was Prof. Feng Lin. *(Corresponding author: Aiqing Zhang.)*

Fangjie Hu and Aiqing Zhang are with the School of Physics and Electronic Information, Anhui Normal University, Wuhu 241002, China (e-mail: fjhu2023@163.com; aqzhang2006@163.com).

Meng Li is with the School of Computer Science and Information Engineering and the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), Hefei University of Technology, Hefei, Anhui 230601, China, and also with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: mengli@hfut.edu.cn).

Chen Wang is with Hubei Key Laboratory of Internet of Intelligence, School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: chenwang@hust.edu.cn).

Digital Object Identifier 10.1109/TIFS.2026.3689720

***Index Terms*—Federated learning, discrete wavelet transform, model poisoning attack.**

I. Introduction

FEDERATED Learning (FL) [1] has emerged as a privacy-preserving paradigm that enables collaborative model training across decentralized entities (e.g., mobile devices, edge nodes, organizations) without exposing their private datasets. In FL, participants keep data locally, performing computations on their devices and sharing only model updates (gradients or weight parameters) with a central server for aggregation. This architecture reduces privacy risks while decreasing communication overhead and computational bottlenecks, making FL valuable for privacy-sensitive domains including healthcare [2], [3], finance, and edge computing [4].

Despite its decentralized strengths, FL's open architecture and statistical heterogeneity expose it to fundamental security vulnerabilities. Poisoning attacks [5] have been empirically demonstrated to pose a fundamental threat to FL systems. Malicious clients can execute model poisoning attacks by submitting manipulated model updates to the system. These attacks either cause systematic performance degradation [5], [6] on targeted tasks or introduce backdoor [7], [8] functionality that produces incorrect predictions when specific input triggers are present. The defense complexity against poisoning attacks is further exacerbated by the intrinsic characteristics of FL environments.

To address these security vulnerabilities, the research community has increasingly focused on fortifying FL systems without compromising model performance. Recent defensive approaches can be broadly categorized into three strategies: Impact Reduction [9], [10], Robust Aggregation [6], and Detection and Filtering [11], [12]. While these mechanisms have demonstrated partial robustness, they fundamentally operate within a reactive security paradigm that leaves FL systems vulnerable to adaptive adversaries.

The limitations of current defenses expose a critical research gap at the intersection of security and distributed learning. Traditional passive detection methods operate under the flawed assumption that malicious behavior will manifest as statistical outliers [13]. This assumption fails when confronted with sophisticated adversaries who employ gradient optimization techniques [14] or generative methods to craft updates that

statistically mimic benign contributions [5], [6], [15]. More concerning, state-of-the-art robust aggregation techniques [9], [11], [16], [17] collapse entirely under Byzantine majority scenarios, where malicious clients can collectively orchestrate their updates to manipulate the aggregation process while appearing individually legitimate.

A fundamental limitation underlying current approaches is their static, snapshot-based analysis of model parameters. This methodology ignores a crucial insight: benign and malicious models, though potentially indistinguishable in their static state, exhibit fundamentally different response dynamics when subjected to controlled perturbations. The necessity for a paradigm shift in FL security becomes evident when considering the evolving capabilities of adversaries. Rather than passively accepting submitted updates, a robust defense should actively probe clients through controlled perturbations, compelling models to expose latent behavioral discrepancies indicative of malicious intent.

To address these challenges, we propose SpecShield, a proactive defense mechanism by employing specifically designed techniques. Specifically, we employ the Fast Gradient Sign Method (FGSM) [18], originally designed for adversarial attacks, as a defensive tool to elicit informative client responses. This creates a controlled stress test that affects benign and malicious models differently. To magnify and reveal the behavioral differences elicited by perturbations, we design a novel detection mechanism based on wavelet-driven frequency-domain profiling. To the best of our knowledge, this is the first work leveraging Discrete Wavelet Transform (DWT) [19] to analyze client responses before and after FGSM-induced perturbations. It operates without assumptions about model architecture or data distribution characteristics, eliminating dependencies on restrictive statistical assumptions common in existing methods. These frequency-domain representations are further processed using density-based spatial clustering (DBSCAN) [20] to identify anomalous client behaviors without relying on prior distributional assumptions. In addition, our analysis yields a provable upper bound on perturbation strength, offering theoretical guidance for optimal perturbation and enhancing the defense's interpretability. Our key contributions can be summarized as follows:

- We propose SpecShield, a proactive defense framework that induces dynamic client responses via adversarial perturbations and analyzes their behavior in the frequency domain. This method maintains effectiveness even in Byzantine majority scenarios.
- We design a novel wavelet-based profiling mechanism, leveraging DWT to extract spectral features from client updates before and after controlled perturbation, which enables the capture of latent behavioral divergence between benign and malicious participants.
- We derive precise upper bounds for perturbation magnitudes in FGSM-based defenses, establishing a formal theoretical framework to guide their optimal perturbation. Furthermore, extensive experiments conducted on real-world datasets demonstrate that SpecShield achieves high detection accuracy and robust model performance under diverse poisoning scenarios, demonstrating the generality of the approach.

II. Related Work

In this section, we review advanced defense strategies against poisoning attacks in FL.

Several methods focus on identifying anomalous gradients through statistical measures. Krum [21] selects the gradient closest to its neighbors based on Euclidean distance as the global update, while Chen et al. [22] propose a defense scheme that aggregates gradients using their median. However, these approaches share a fundamental limitation: they select a single device's gradient as the global update, compromising the collaborative essence of FL. Addressing this shortcoming, Bulyan [10] builds upon Krum by aggregating multiple gradient vectors based on proximity measures rather than selecting just one. Similarly, Yin et al. [9] introduce a strategy that eliminates extreme gradient vectors and averages the remaining ones to preserve the diversity of client contributions. Another category examines model parameters directly. FL-Defender [13] and CosDefense [23] analyze the weight parameters of the last layer using cosine similarity for client differentiation. RoseAgg [24] detects collusion by analyzing pairwise cosine similarity and incorporates personalized model integration to resist targeted attacks. In addition, FedRoLa [25] partitions updates by model depth and applies localized filtering, enhancing robustness against partial poisoning. Although these methods perform well in homogeneous settings, their effectiveness diminishes in non-IID scenarios, highlighting the difficulty of designing defenses robust across different data distribution scenarios.

Several techniques rely on server-side trusted data for validation. FLTrust [11] assumes the server maintains a clean root dataset and uses cosine similarity against this trusted data for anomaly detection. Similarly, FLShield [26] leverages validation performance on a held-out trusted dataset to dynamically penalize malicious clients. While potentially effective, these approaches impose strong assumptions about server-side data availability that fundamentally conflict with privacy-preserving FL principles. Addressing this shortcoming, FedDMC [27] detects malicious clients via a two-stage mechanism that combines historical consistency analysis and intra-round similarity filtering. Similarly, FreqFed [12] applies Discrete Cosine Transform to extract frequency features from model updates and performs unsupervised clustering to eliminate poisoned clients, which improves robustness without requiring a trusted validation set.

Clustering methods offer a more flexible detection paradigm. AUROR [28] employs K-means clustering to separate clients into two groups based on indicative features. However, this method requires prior knowledge of the original data distribution. Addressing this limitation, FLAME [29] further refines clustering techniques by combining cosine distance with HDBSCAN to identify poisoners. However, FLAME assumes that a majority of clients are benign. Complementing these approaches, Principal Component Analysis [24], [27], [30] has been utilized to extract clean components as criteria for distinguishing between benign and malicious updates,

offering a dimension-reduction perspective on the detection problem.

A more recent class of defense mechanisms shifts toward proactive detection and adversarial probing. SIREN+ [31] introduces a proactive alarming mechanism that flags anomalous client behavior based on round-wise loss fluctuation patterns. RECESS [32] takes this one step further by actively injecting small adversarial perturbations into the global model before dispatching it to clients, then observing the perturbed response to distinguish between benign and adversarial behavior. Despite these diverse approaches, most existing defenses still share fundamental limitations. They typically rely on the honest-majority assumption or require unrealistic knowledge of underlying data distributions.

Unlike prior works, we proactively elicit client-specific behaviors via calibrated perturbations and analyze their dynamic frequency-domain responses to identify malicious participants. Furthermore, we jointly consider adaptive attacks, non-IID data distributions, and high poisoning ratios without relying on statistical assumptions or trusted data. Our work addresses these limitations through a proactive defense paradigm that does not rely on these restrictive assumptions.

III. Background

In this section, we introduce the FGSM-based perturbation mechanism, and outline the use of DWT for frequency-domain analysis.

A. Federated Learning

We examine a standard FL setting where multiple clients jointly train a global model coordinated by a central server. The training follows an iterative optimization process defined as follows: Let $\mathcal{K} = \{1, 2, \ldots, K\}$ represent the set of participating clients. The central server initializes a global model θ_0 and sends the model parameters θ_t to each client at round t. Each client $k \in \mathcal{K}$ maintains a private dataset $\mathcal{D}_k$ and defines a local objective function $\mathcal{L}(\theta, \mathcal{D}_k)$. After receiving θ_t, each client performs E epochs of local optimization using stochastic gradient descent (SGD) to minimize its local objective and calculate:

$$\theta_{t+1}^k = \theta_t - \eta \nabla \mathcal{L}(\theta_t, \mathcal{D}_k), \tag{1}$$

where η is the learning rate. The client then sends the updated model θ_{t+1}^k back to the server. The server aggregates these updates using a predefined rule: $\theta_{t+1} = \sum_{k \in \mathcal{K}} p_k \theta_{t+1}^k$ where $p_k = \frac{|\mathcal{D}_k|}{\sum_{j \in \mathcal{K}} |\mathcal{D}_j|}$ is the weight assigned to each client's update, typically proportional to the local dataset size. This process repeats for multiple rounds until convergence, allowing the global model to benefit from all clients' data without direct access to it.

B. Fast Gradient Sign Method

FGSM [18] is a fundamental technique in adversarial machine learning that was originally developed to identify vulnerabilities in deep neural networks. FGSM creates adversarial examples by using gradient information to modify input data, causing misclassifications while making minimal changes to the original input. The mathematical formulation of FGSM is:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)), \tag{2}$$

where x is the original input, ϵ controls the magnitude of perturbation, and $\nabla_x J(\theta, x, y)$ represents the gradient of the model's loss function with respect to the input. While FGSM was initially designed as an attack vector, it has also proven valuable for evaluating model robustness and developing adversarial training methods. In our work, we repurpose FGSM from an offensive tool to a defensive instrument.

C. Discrete Wavelet Transform

Wavelet Transform (WT) [19] is a signal processing technique with strong capabilities for feature extraction and anomaly detection. Its primary advantage is its ability to capture localized signal variations across multiple scales. For FL security, DWT provides an analytical framework for examining model parameters. Through scale and translation invariance properties, DWT converts model parameters into the frequency domain, exposing variations that traditional statistical methods cannot detect. The mathematical representation of DWT is:

$$W_{j,k} = \sum_n f(n) \cdot \psi_{j,k}(n), \tag{3}$$

where j and k denote scale and displacement parameters respectively, and $\psi_{j,k}(n)$ is the wavelet basis function generated through scale and displacement transformations. DWT operates by decomposing signals through high-pass and low-pass filters, extracting both low-frequency approximation coefficients and high-frequency detail coefficients. It reveals key signal characteristics in the frequency domain.

IV. Problem Formulation

A. Threat Model

Building upon the classical FL architecture [1], the system comprises a central server and k clients. In FL, the primary security threats originate from potentially malicious clients and potentially untrusted servers. Given that our research focuses primarily on model poisoning attacks, we operate under the assumption that the server remains honest and trustworthy. In our paper, we consider a strong white-box adversary as described in [5], [6], [15], and [33]. It possesses access to both the global model parameters and the model updates submitted by other clients. Such an adversary can construct threat models using locally held clean data to approximate and analyze the behavior of benign clients, thereby crafting highly adaptive and deceptive poisoning strategies. To rigorously evaluate the robustness and generalizability of our proposed defense mechanism, we adopt the more challenging white-box setting throughout our experiments.

B. Defender's Capability and Knowledge

- *Defender's Capability:* The defender is the honest and trusted aggregation server. In each round, the server has access to the current global model, distributes the

TABLE I
MATHEMATICAL NOTATION TABLE

Symbol	Description
θ_i	Model parameters of client i
θ_i'	Uploaded model of client i after local training
$\Delta\theta_i$	Perturbation applied to client i
ϵ	Perturbation magnitude factor
$\mathcal{W}(\cdot)$	DWT operator
ℓ	Wavelet decomposition level
$A_i^{(L)}$	Low-frequency components of client i at level L
$B_i^{(\ell)}$	High-frequency components of client i at level ℓ
$E_i^{(\ell)}$	Spectral energy feature at level ℓ
$H_i^{(\ell)}$	Frequency entropy at level ℓ
D_i	Differential feature vector of client i
C_i	The i-th DBSCAN cluster
B_t	Benign client set at round t
$\bar{D}_{\text{ben}}$	Mean benign differential feature vector

global model to selected clients, and collects their locally updated models. Beyond standard aggregation, SpecShield allows the server to inject bounded perturbations into the model before transmission, so that client behavior can be actively probed.

- *Defender's Knowledge:* The defender does not observe clients' raw training data or local data distributions. It also does not rely on any trusted validation data or any auxiliary clean reference set. Instead, the defense operates solely on observable model parameters and their frequency-domain response differences under server-induced perturbations. Importantly, SpecShield does not require an honest-majority assumption.

C. Design Goals

Our research establishes specific design criteria for the development of an effective proactive defense mechanism.

- ***Precise Identification of Malicious Clients***: A fundamental objective of a defense mechanism is to achieve superior detection accuracy for malicious clients, even under extreme adversarial conditions.
- ***Enhanced Global Model Robustness***: The defense mechanism should demonstrate strong cross-scenario transferability, maintaining robustness across diverse data distributions and attack intensities.
- ***Efficient Defense Overhead Management***: The defense mechanism must carefully balance security enhancements against resource consumption, ensuring that additional computational and communication overhead remains within acceptable bounds.

We provide a comprehensive mathematical notation table in Tab. I.

V. SPECSHIELD

In this section, we present our new defense against poisoning attacks, SpecShield.

A. Overview

The high-level overview of our framework SpecShield is shown in Fig. 1 and Algorithm 1. In each round, the server proactively crafts personalized perturbations for each client's model parameters θ_i using the FGSM. Upon receiving these perturbed parameters, clients perform local training and return the updated models θ_i'. Based on the collected parameter pairs (θ_i, θ_i'), the server applies DWT to each client's model parameters both before and after introducing controlled perturbations. It extracts both low-frequency and high-frequency components, then derives key frequency-domain metrics. These frequency-domain features are then fed into the DBSCAN clustering algorithm. After grouping clients based on density in the frequency domain, the server computes the mean value of frequency domain differential features for each cluster and selects the one with the highest representative value for secure aggregation.

Algorithm 1 SpecShield

Require: Initial global model θ_0, total rounds T, selected client set $\mathcal{S}_t$
Ensure: Final global model θ_T

for each round $t \in [1, T]$ **do**
  if detection mode is **True then**
  /* Construct induced perturbation */
    for each $i \in \mathcal{S}_t$ **do**
      $\Delta\theta_i \leftarrow \epsilon \cdot \text{sign}\big(\nabla_{\theta_i} L_{\text{induced}}\big)$
      $\theta_i' \leftarrow \theta_i + \Delta\theta_i$
    end for
  /* DWT and feature extraction */
    for each $i \in \mathcal{S}_t$ **do**
      $W(\theta_i), W(\theta_i') \leftarrow \text{DWT}(\theta_i, \theta_i')$
      $D_i \leftarrow \text{ExtractFeature}\big(W(\theta_i), W(\theta_i')\big)$
    end for
  /* DBSCAN clustering */
    $(C_1, C_2, \ldots, C_L) \leftarrow \text{DBSCAN}(\text{PCA}(\{D_i\}))$
    for each cluster C_ℓ **do**
      $R(C_\ell) \leftarrow \frac{1}{|C_\ell|} \sum_{j \in C_\ell} \|D_j\|_2^2$
    end for
    $C^* \leftarrow \arg\max_{C_\ell} R(C_\ell)$
    Identify benign clients: $\mathcal{B}_t \leftarrow \{i \mid C_i = C^*\}$
  end if
  Aggregate benign updates
end for

B. Induced Parameter Generation

The server generates personalized perturbations for each client using the FGSM with a unified loss function. This process is mathematically expressed as:

$$\Delta\theta_i = \epsilon \cdot \text{sign}\left(\nabla_{\theta_i} L_{\text{induced}}\right), \tag{4}$$

where $\Delta\theta_i$ represents the generated inducement parameters, ϵ controls perturbation intensity, and $\nabla_{\theta_i} L_{\text{induced}}$ denotes the gradient of the loss function with respect to model parameters. The server distributes these personalized perturbations $\theta_i + \Delta\theta_i$ to respective clients. They perform local training and update their models upon receiving these perturbed parameters.

To ensure controlled impact on benign clients, we rigorously constrain the perturbation intensity of inducement parameters through theoretical derivation (detailed in Sec. VI). This

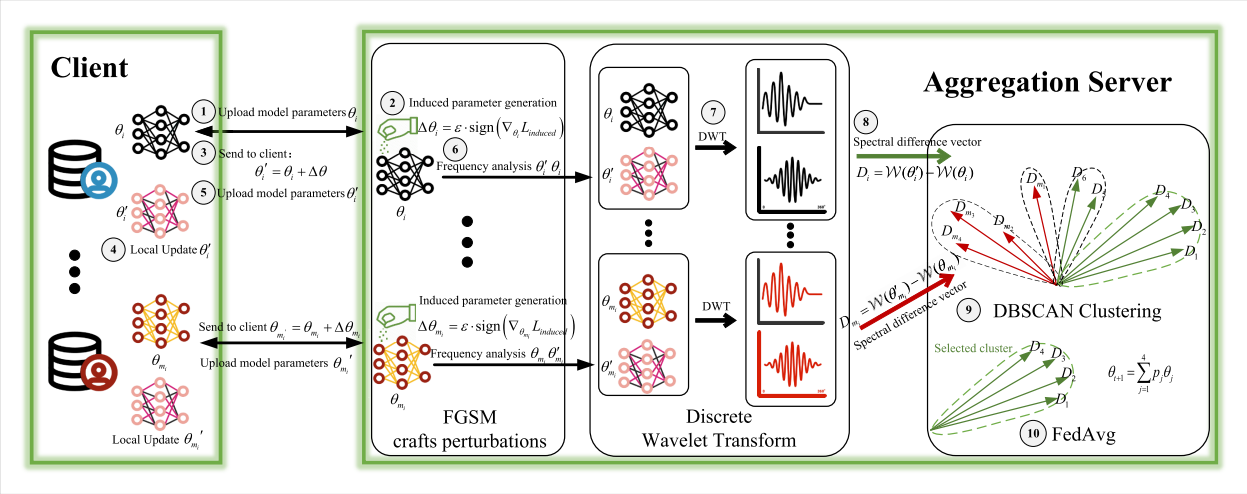


Fig. 1. Overview of SpecShield.

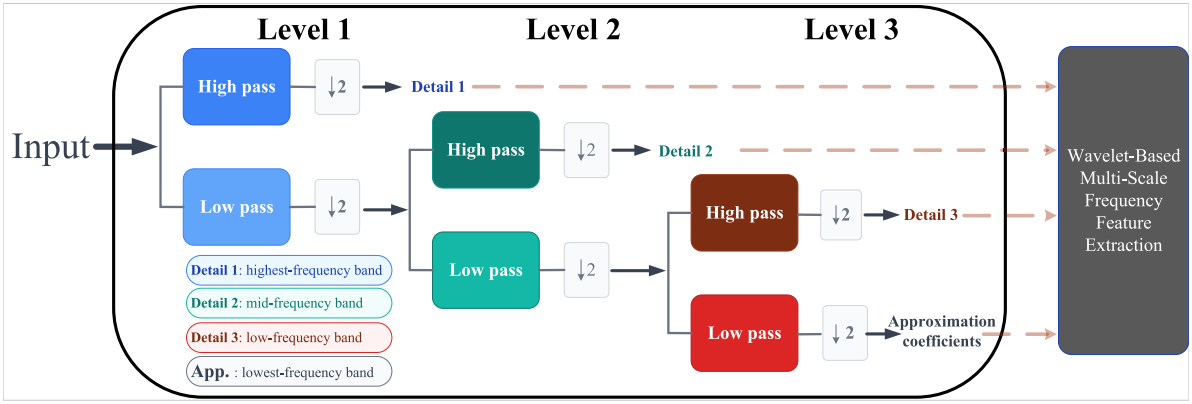


Fig. 2. Three-Level Daubechies-4 wavelet decomposition for frequency domain feature extraction.

design guarantees the amplification effect on malicious client behavior without introducing excessive interference to benign clients. Intuitively, malicious clients, having crafted adversarial updates, will likely try to counteract the perturbations. Alternatively, they may exploit the perturbations to enhance their attack and maintain their objectives. These behavioral adaptations generate distinctive pattern signatures in the uploaded model updates. They can be effectively identified in subsequent analysis phases. In the following sections, we elaborate on how these amplified differences enable efficient detection and defense against malicious clients.

C. Frequency Domain Analysis

In neural network models, weights represent connection strengths between neurons. During training, their distribution and corresponding energy evolve dynamically. This process progressively reduces the gap between model predictions and the ground truth. However, when malicious clients inject poisoning attacks, this natural evolution is artificially altered. Malicious clients manipulate their local training processes to capture specific artificial correlations in model weights that do not exist in benign data. These artificial correlations manifest as local or global biases in model updates. Adversaries use optimization strategies to disguise malicious updates as benign contributions, allowing them to evade traditional anomaly detection methods.

To overcome these challenges, we propose a frequency-domain feature extraction method based on wavelet transform. Specifically, clients perform local training and upload post-perturbation (θ_i') model parameters to the server. The server first normalizes these uploads to eliminate range variations:

$$\tilde{\theta}_i = \frac{\theta_i}{||\theta_i||_2}, \quad \tilde{\theta}_i' = \frac{\theta_i'}{||\theta_i'||_2}, \tag{5}$$

then applies DWT using Daubechies wavelets as basis functions to both sets of parameters, as shown in Fig. 2. The parameter vector is decomposed with three levels of wavelet decomposition. This decomposes the parameters into multi-scale representations, extracting both low and high-frequency components:

$$W_i = \mathcal{W}(\tilde{\theta}_i) = \left\{A_i^{(L)}, B_i^{(1)}, B_i^{(2)}, \ldots, B_i^{(L)}\right\}, \tag{6}$$

$$W_i' = \mathcal{W}(\tilde{\theta}_i') = \left\{A_i'^{(L)}, B_i'^{(1)}, B_i'^{(2)}, \ldots, B_i'^{(L)}\right\}. \tag{7}$$

where $A_i^{(L)}$ and $A_i'^{(L)}$ represent the low-frequency approximation components at the highest level L of the wavelet decomposition, and $B_i^{(\ell)}$ and $B_i'^{(\ell)}$ denote the high-frequency

detail components at each level ℓ. The server extracts the spectral energy distribution $E_i^{(\ell)}$, frequency entropy $H_i^{(\ell)}$, and scale variation characteristics as frequency-domain features from the pre-perturbation and post-perturbation parameter sets, and constructs differential feature vectors D_j based on their differences. Here, the frequency entropy $H_i^{(\ell)}$ is defined as the Shannon entropy of the normalized squared wavelet coefficients at level ℓ:

$$p_{i,k}^{(\ell)} = \frac{|B_{i,k}^{(\ell)}|^2}{\sum_m |B_{i,m}^{(\ell)}|^2}, \tag{8}$$

$$H_i^{(\ell)} = -\sum_k p_{i,k}^{(\ell)} \log p_{i,k}^{(\ell)}, \tag{9}$$

where $B_{i,k}^{(\ell)}$ denotes the k-th wavelet coefficient at level ℓ, and $p_{i,k}^{(\ell)}$ represents its normalized energy contribution.

$$D_i = \left[E_i^{(\ell)} - E_i'^{(\ell)}, H_i^{(\ell)} - H_i'^{(\ell)}, \ldots \right], \quad \forall \ell \in \{1, \ldots, L\}. \tag{10}$$

By transforming differential information into frequency-domain signals, we can more efficiently capture behavioral differences between clients before and after inducement.

D. Secure Aggregation

To effectively group and filter clients, we first standardize D_i using z-score normalization, clip the values to the range of $[-5, 5]$, and project them into a 2-dimensional PCA space. Subsequently, we employ the DBSCAN clustering algorithm to identify distinct client groups. It automatically organizes the data based on density reachability principles. Following the clustering process, the server conducts further analysis on each cluster to quantify its trustworthiness. Specifically, the server calculates the mean value of frequency-domain differential features D_j for all clients within each cluster, defined as the cluster's representative value. This representative value is calculated as:

$$R(C_i) = \frac{1}{|C_i|} \sum_{j \in C_i} \|D_j\|_2^2, \tag{11}$$

where $R(C_i)$ denotes the representative value of cluster C_i, $|C_i|$ represents the number of clients within the cluster, and D_j is the frequency-domain differential feature value of client j. This representative value reflects the overall intensity of the response to inducement parameters among clients in the cluster.

A higher representative value means that clients in the cluster showed more significant differences after being influenced by the inducement parameters. This pronounced difference usually suggests that these clients did not try to counteract the perturbation. Instead, they faithfully reflected the impact of the perturbation on their model training process. Consequently, clusters with larger representative values are more likely to consist of benign clients. Based on this principle, the server selects the cluster with the highest representative value as the group of benign clients.

$$C^* = \arg\max_{C_i \in C} R(C_i). \tag{12}$$

Our method clusters frequency-domain differences before and after perturbations for each client. This measure is inherently independent of other clients' behaviors or data distributions, enabling accurate differentiation between normal benign model parameters and malicious model parameters, even in situations with severely imbalanced data distributions or high proportions of malicious clients.

VI. Theoretical Derivation of Perturbation Upper Bound

This section provides a rigorous mathematical foundation for determining the optimal perturbation magnitude that maximizes adversary detection while ensuring minimal impact on benign clients' performance.

A. Preliminaries and Problem Setup

In our FL setting, let $\theta^k \in \mathbb{R}^d$ denote the global model at round k. In conventional FL, client i initiates local training from θ^k and performs E local updates using stochastic gradient descent. The final local model after E updates is:

$$\theta_i' = \theta^k - \eta \sum_{t=0}^{E-1} \nabla F_i(\theta_i^{(t)}, \zeta_t). \tag{13}$$

In our proactive defense framework, the server introduces a client-specific perturbation δ_i at the initialization of each client's local training:

$$\delta_i = \epsilon \cdot \text{sign}(\nabla F_i(\theta^k)). \tag{14}$$

This perturbation satisfies:

$$\|\delta_i\|_\infty = \epsilon, \quad \|\delta_i\|_2 \le \epsilon \sqrt{d}. \tag{15}$$

The client now starts local training from the perturbed model $\theta^k + \delta_i$, resulting in:

$$\theta_i'(\delta_i) = \theta^k + \delta_i - \eta \sum_{t=0}^{E-1} \nabla F_i(\theta_i^{(t)}, \zeta_t). \tag{16}$$

Our objective is to determine the maximum perturbation magnitude $\epsilon_{\max}$. A critical challenge is to balance perturbation magnitude: insufficient perturbation may fail to reveal adversarial behavior, while excessive perturbation risks degrading the performance of benign models.

B. Upper Bound Analysis Under Strong Convexity

We first analyze the scenario where the loss function $F_i(\theta)$ satisfies standard optimization assumptions.

Assumption 1 (μ-strong convexity): The loss function $F_i(\theta)$ is μ-strongly convex for $\mu > 0$, if

$$F_i(\theta') \ge F_i(\theta) + \nabla F_i(\theta)^\top (\theta' - \theta) + \frac{\mu}{2} \|\theta' - \theta\|_2^2. \tag{17}$$

Assumption 2 (L-smoothness): The loss function $F_i(\theta)$ is L-smooth, if

$$\|\nabla F_i(\theta) - \nabla F_i(\theta')\|_2 \le L \|\theta - \theta'\|_2. \tag{18}$$

Under these assumptions, we can characterize how the perturbation impacts training convergence.

Theorem 1: Let $F_i(\theta)$ be μ-strongly convex and L-smooth. If client i performs E local gradient descent steps with step size

$\eta < \frac{2}{L+\mu}$ from an initial point perturbed by δ_i where $\|\delta_i\|_2 \leq \epsilon\sqrt{d}$, then the additional loss incurred due to the perturbation is bounded by:

$$F_i(\theta_i'(\delta_i)) - F_i(\theta_i') \leq \epsilon\sqrt{d}L(1-\mu\eta)^E, \tag{19}$$

where θ_i' denotes the model obtained after E local updates without perturbation.

Proof: We begin by analyzing the convergence behavior of gradient descent under strong convexity. For the unperturbed training trajectory, after E updates, the residual loss compared to the client's local optimum θ_i^* is:

$$R_0 = F_i(\theta_i') - F_i(\theta_i^*) \leq (1-\mu\eta)^E[F_i(\theta^k) - F_i(\theta_i^*)], \tag{20}$$

where $R_0 = F_i(\theta_i') - F_i(\theta_i^*)$ is the residual loss after E steps of unperturbed training.

For the perturbed trajectory, we need to bound the parameter distance between the perturbed and unperturbed final models. Since gradient descent with step size $\eta < \frac{2}{L+\mu}$ is a contraction mapping under μ-strong convexity, we have:

$$\|\theta_i'(\delta_i) - \theta_i'\|_2 \leq (1-\mu\eta)^E\|\delta_i\|_2 \leq \epsilon\sqrt{d}(1-\mu\eta)^E. \tag{21}$$

Using the L-smoothness of F_i, we can bound the loss difference:

$$|F_i(\theta_i'(\delta_i)) - F_i(\theta_i')| \leq L\|\theta_i'(\delta_i) - \theta_i'\|_2 \leq \epsilon\sqrt{d}L(1-\mu\eta)^E. \tag{22}$$

This completes the proof. □

To ensure benign clients experience only acceptable performance degradation, we impose the constraint:

$$F_i(\theta_i'(\delta_i)) - F_i(\theta_i^*) \leq \Delta F_{\max}, \tag{23}$$

where $\Delta F_{\max}$ represents the maximum tolerable loss increase. Using our results:

$$\begin{aligned} &F_i(\theta_i'(\delta_i)) - F_i(\theta_i^*) \\ &= [F_i(\theta_i') - F_i(\theta_i^*)] + [F_i(\theta_i'(\delta_i)) - F_i(\theta_i')] \\ &\leq R_0 + \epsilon\sqrt{d}L(1-\mu\eta)^E \leq \Delta F_{\max}. \end{aligned} \tag{24}$$

Solving for ϵ, we obtain the ϵ upper bound $\epsilon_{\max}$:

$$\epsilon_{\max} = \frac{\Delta F_{\max} - R_0}{L(1-\mu\eta)^E\sqrt{d}}. \tag{25}$$

C. Generalization to Non-Convex Settings via PL Condition

Since deep learning models in FL often operate in non-convex settings, we extend our analysis using the Polyak-Lojasiewicz (PL) condition, which is weaker than strong convexity but still ensures linear convergence.

Definition 1 (PL Condition): A differentiable function F_i satisfies the Polyak-Lojasiewicz condition with constant $\mu > 0$ if:

$$\frac{1}{2}\|\nabla F_i(\theta)\|_2^2 \geq \mu(F_i(\theta) - F_i(\theta_i^*)), \tag{26}$$

where F_i^* is the global minimum value of F_i.

Theorem 2: Suppose F_i satisfies the PL condition with constant $\mu > 0$ and is L-smooth. If client i performs E local gradient descent steps with step size $\eta < \frac{1}{L}$ from an initial point perturbed by δ_i where $\|\delta_i\|_2 \leq \epsilon\sqrt{d}$, then with high probability (due to stochastic gradients), the additional loss incurred due to the perturbation is bounded by:

$$F_i(\theta_i'(\delta_i)) - F_i(\theta_i') \leq \epsilon\sqrt{d}\,\sqrt{2\mu R_0}\cdot(1-\mu\eta)^E. \tag{27}$$

Proof: Under the PL condition, the convergence analysis proceeds similarly to the strongly convex case, but with different constants:

$$F_i(\theta_i^{(t+1)}) - F_i(\theta_i^*) \leq (1-\mu\eta)(F_i(\theta_i^{(t)}) - F_i(\theta_i^*)). \tag{28}$$

After E iterations without perturbation, we have:

$$R_0 = F_i(\theta_i') - F_i(\theta_i^*) \leq (1-\mu\eta)^E(F_i(\theta^k) - F_i(\theta_i^*)). \tag{29}$$

For the perturbed trajectory, we decompose the total loss difference:

$$\begin{aligned} &F_i(\theta_i'(\delta_i)) - F_i(\theta_i^*) \\ &= [F_i(\theta_i') - F_i(\theta_i^*)] + [F_i(\theta_i'(\delta_i)) - F_i(\theta_i')] \\ &= R_0 + \Delta F_\delta. \end{aligned} \tag{30}$$

Using a first-order Taylor expansion around θ_i':

$$\Delta F_\delta \approx \nabla F_i(\theta_i')^\top(\theta_i'(\delta_i) - \theta_i'). \tag{31}$$

By Cauchy-Schwarz inequality:

$$\Delta F_\delta \leq \|\nabla F_i(\theta_i')\|_2 \cdot \|\theta_i'(\delta_i) - \theta_i'\|_2. \tag{32}$$

From the PL condition:

$$\|\nabla F_i(\theta_i')\|_2 \leq \sqrt{2\mu R_0}. \tag{33}$$

And since $\|\theta_i'(\delta_i) - \theta_i'\|_2 \leq \epsilon\sqrt{d}(1-\mu\eta)^E\|\delta_i\|_2 \leq (1-\mu\eta)^E$, we get:

$$\Delta F_\delta \leq \epsilon\sqrt{d}\,\sqrt{2\mu R_0}\cdot(1-\mu\eta)^E. \tag{34}$$

This completes the proof. □

To ensure benign client performance remains acceptable, it should satisfy:

$$R_0 + \sqrt{2\mu R_0}\cdot\epsilon\sqrt{d}(1-\mu\eta)^E \leq \Delta F_{\max}. \tag{35}$$

Solving for ϵ, we obtain the ϵ upper bound $\epsilon_{\max}$:

$$\epsilon_{\max} = \frac{\Delta F_{\max} - R_0}{\sqrt{2\mu R_0}\cdot(1-\mu\eta)^E\sqrt{d}}. \tag{36}$$

D. Accounting for Stochasticity in Client Updates

In practice, clients use stochastic gradient descent (SGD) with mini-batch sampling, introducing variance in the model updates. We extend our bound to accommodate this stochasticity using concentration inequalities.

Theorem 3: When clients use SGD with bounded gradient variance σ^2, the perturbation-induced loss deviation satisfies with probability at least $1-\delta$:

$$F_i(\theta_i'(\delta_i)) - F_i(\theta_i') \leq \epsilon\sqrt{d}L + t(\delta), \tag{37}$$

where $t(\delta) = \sqrt{2\eta^2\sigma^2E\ln(1/\delta)} + \frac{1}{3}B\ln(1/\delta)$ and B is an upper bound on individual SGD update magnitudes.

Proof: We model the client's SGD updates as a martingale difference sequence:

$$S_E = \sum_{t=0}^{E-1} X_t, \quad \text{where} \quad X_t = -\eta\nabla F_i(\theta_i^{(t)}, \zeta_t). \tag{38}$$

TABLE II
EXPERIMENT DATASETS AND FL SETTINGS

Dataset	Size	Dimension	Clients	Batch Size	Optimizer	Learning Rate	Epochs	Local Epochs
MNIST	60,000	28×28	50	128	SGD	0.01 → 0.001	30	3
Fashion-MNIST	70,000	28×28	50	64	SGD	0.01 → 0.001	30	3
CIFAR-10	60,000	32×32	50	32	SGD	0.01 → 0.001	100	10
Tiny-ImageNet	110,000	64×64	50	64	SGD	0.05 → 0.01	150	3

With cumulative variance $V_E = \sum_{t=0}^{E-1} \mathbb{E}[\|X_t\|^2]$. Applying Freedman's inequality:

$$\|S_E\| \leq \sqrt{2V_E \ln(1/\delta)} + \frac{1}{3}B\ln(1/\delta). \tag{39}$$

This bounds the deviation due to gradient noise. Combined with the deterministic bound on perturbation effects, we obtain:

$$\|\theta_i'(\delta_i) - \theta_i'\| \leq \|\delta_i\| + t(\delta). \tag{40}$$

Using L-smoothness:

$$F_i(\theta_i'(\delta_i)) - F_i(\theta_i') \leq L(\|\delta_i\| + t(\delta)) \leq \epsilon\sqrt{d}L + Lt(\delta). \tag{41}$$

For simplicity, we absorb L into $t(\delta)$, giving our final bound.□

E. Dynamic Determination of Tolerance Threshold

Rather than assuming a fixed tolerance threshold $\Delta F_{\max}$, we propose deriving it from intrinsic statistical properties of the FL process.

Theorem 4 (Probabilistic Tolerance Threshold): Consider a FL system where each client i optimizes a local loss function $F_i(\theta)$ that is L-smooth and satisfies the PL condition with parameter μ. Let client i perform E local SGD steps with learning rate $\eta < \frac{1}{L}$ starting from a perturbed initialization $\theta^k + \delta_i$ where $\|\delta_i\|_2 \leq \epsilon\sqrt{d}$. Then, for any benign client i participating honestly in the federated objective, the following holds with probability at least $1-\delta$:

$$F_i(\theta_i'(\delta_i)) - F_i(\theta_i^*) \leq \Delta F_{\max} = R_0 + \epsilon\sqrt{d}L + t(\delta). \tag{42}$$

Proof: We decompose the total loss deviation into three independent components and bound each separately.

Step 1: Loss Decomposition. For any client i receiving perturbation δ_i, we can write Eq. 29. The first term R_0 represents the natural optimization gap that arises from the finite number of local epochs. Under the PL condition, this satisfies Eq. 28. For the second term ΔF_δ, we leverage the parameter distance bound from Theorem 2. Under L-smoothness:

$$|\Delta F_\delta| = |F_i(\theta_i'(\delta_i)) - F_i(\theta_i')| \leq L\|\theta_i'(\delta_i) - \theta_i'\|_2. \tag{43}$$

From our previous analysis, we have $\|\theta_i'(\delta_i) - \theta_i'\|_2 \leq \epsilon\sqrt{d}(1-\mu\eta)^E$. Under the conservative bound $(1-\mu\eta)^E \leq 1$:

$$|\Delta F_\delta| \leq L \cdot \epsilon\sqrt{d}. \tag{44}$$

Step 2: The SGD process introduces additional randomness. $t(\delta)$ quantifies the stochastic deviation from mini-batch SGD and has been established in Theorem 3 via Freedman's inequality.

Combining all three bounds using the union bound and independence of error sources, we obtain:

$$\mathbb{P}[F_i(\theta_i'(\delta_i)) - F_i(\theta_i^*) \leq R_0 + \epsilon\sqrt{d}L + t(\delta)] \geq 1-\delta. \tag{45}$$

This completes the proof. □

VII. EVALUATION

A. Experiment Settings

1) Datasets: We use four datasets in our experiments: MNIST, Fashion-MNIST, CIFAR-10, and Tiny-ImageNet. We create non-IID data distributions among local clients using the Dirichlet distribution. Specifically, we extract $q^j \sim DirN(\beta)$ from a Dirichlet distribution and assign a q_i^j percentage of class j examples to client i. The concentration parameter β determines client similarity, ranging from 0 to 1. We set the non-IID degree β to 0.5 unless otherwise noted. Tab. II summarizes the parameter settings used in our evaluation.

2) Metrics: We evaluate our defense mechanism using standard metrics that align with the design objectives outlined in Sec. IV-C. We define the evaluation terms as follows: ***Test Accuracy (Acc)*** represents the proportion of test instances correctly classified by the global model. ***False Positive Rate (FPR)*** indicates a benign client incorrectly identified as malicious. ***True Positive Rate (TPR)*** represents a malicious client correctly identified as malicious. ***Norm of Eigenvector*** refers to the ℓ_2 norm of the leading eigenvector derived from each client's update perturbation response.

3) Baselines: We compare SpecShield against state-of-the-art defense mechanisms: Mkrum [21], Median [9], Trmean [9], Bulyan [10], FLTrust [11], DnC [6], FreqFed [12], FedDMC [27], RECESS [32]. To ensure fair evaluation, we maintain consistent parameter settings with previous work. Additionally, we consider six recent poisoning techniques: Fang Attack [5], LIE [15], AGR-tail [6], AGR-max [6], and AGR-sum [6], FL-MMR [33]. In line with previous research, we assess SpecShield's defense effectiveness against FedAvg without any Byzantine attacks. Unless otherwise specified, we assume 20% of clients are malicious.

4) Setup: We conducted simulation experiments of the FL system on a server with the PyTorch 1.10.0 framework, version 3.8 of Python that uses Cuda 11.3 to accelerate computation. The experiments were run on a server equipped with an RTX 4090D (24G) GPU with a 16-core Intel Xeon Platinum 8481C CPU and 80GB of main memory.

B. Test Accuracy of SpecShield

Our rigorous evaluation benchmarks SpecShield against nine state-of-the-art defense mechanisms across four standard

TABLE III
COMPARISON OF FL ACCURACY WITH VARIOUS DEFENSES AGAINST POISONING ATTACKS

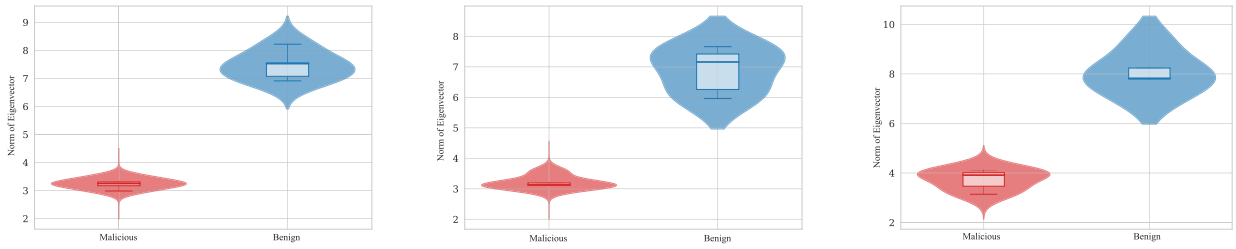
Dataset	Defense	Attack Type						
		No Attack	Fang Attack	LIE Attack	AGR-tail	AGR-max	AGR-sum	FL-MMR
MNIST	Mkrum	96.17±0.08	84.66±0.21	90.00±2.18	85.50±2.20	80.73±2.01	83.53±2.07	71.86±1.71
	Median	93.29±0.16	91.53±0.25	91.33±1.41	91.70±1.93	90.17±0.48	91.24±1.75	69.34±2.11
	Trmean	96.23±0.34	92.41±0.35	91.21±2.12	89.69±1.73	88.72±1.63	90.87±1.17	54.32±1.98
	Bulyan	95.49±0.13	86.12±0.14	90.56±1.36	87.23±1.81	90.14±1.36	87.74±1.97	78.46±3.74
	FLTrust	97.87±0.31	96.13±0.08	95.25±2.28	94.27±0.73	**95.23**±1.93	95.17±1.34	64.32±3.26
	DnC	96.83±0.27	96.17±0.21	96.07±2.08	94.41±1.92	93.51±1.07	**95.73**±2.26	78.89±2.37
	FreqFed	97.81±0.04	97.01±1.09	95.79±1.79	94.07±1.07	91.37±2.37	94.17±0.56	89.32±1.17
	FedDMC	**98.05**±0.07	96.35±0.41	96.31±2.37	92.31±1.42	92.64±0.99	93.20±1.32	87.64±2.16
	RECESS	97.92±0.09	96.88±0.36	95.96±1.24	93.84±1.41	93.12±1.08	94.68±1.17	90.76±1.83
	SpecShield	95.41±0.05	**97.12**±0.43	**96.45**±1.18	**94.89**±1.78	93.57±0.63	94.37±1.37	**94.49**±1.67
Fashion-MNIST	Mkrum	82.85±0.13	61.04±1.04	75.17±2.02	39.53±0.93	40.08±1.14	25.17±1.23	50.94±2.30
	Median	86.97±0.24	83.66±1.15	71.37±2.43	58.97±1.79	47.34±2.48	60.56±1.48	56.75±2.11
	Trmean	87.23±0.27	83.76±0.94	72.33±1.92	66.76±2.11	62.17±1.37	51.51±1.53	60.34±1.84
	Bulyan	83.14±0.30	74.65±1.14	66.71±1.30	46.54±2.27	67.43±2.20	67.46±2.18	63.75±2.10
	FLTrust	**89.75**±0.11	83.23±0.80	83.01±1.33	**86.41**±0.71	81.93±0.84	82.18±1.37	35.46±2.73
	DnC	82.74±0.14	75.21±1.16	71.82±0.90	63.93±0.80	76.66±1.73	70.37±2.42	59.64±2.27
	FreqFed	86.31±0.20	82.16±0.69	83.76±1.16	80.31±0.97	81.46±1.36	79.97±2.18	**81.36**±0.97
	FedDMC	88.76±0.18	85.46±1.32	82.17±0.66	83.46±1.59	82.67±0.92	80.43±0.89	75.09±2.42
	RECESS	89.14±0.12	86.73±0.91	83.94±0.88	84.21±1.03	83.48±0.95	82.96±1.12	78.64±1.74
	SpecShield	86.08±0.04	**87.66**±0.94	**84.72**±0.46	85.16±1.09	**84.29**±1.30	**84.54**±0.63	80.34±1.59
CIFAR-10	Mkrum	67.73±1.37	55.28±0.40	45.12±0.79	31.23±1.43	33.11±1.78	34.35±1.41	47.34±1.85
	Median	64.47±1.24	54.53±0.43	56.35±1.03	26.50±0.78	29.34±0.27	34.92±0.93	41.38±1.38
	Trmean	73.76±0.43	54.36±1.03	60.63±0.91	31.55±1.16	27.73±1.44	43.81±1.63	43.76±2.13
	Bulyan	64.41±1.31	61.95±0.98	36.48±1.38	23.49±1.18	29.48±1.21	25.51±1.20	38.76±2.34
	FLTrust	71.54±1.05	66.16±1.95	63.19±0.79	55.34±1.29	53.03±1.67	50.94±1.46	17.36±4.76
	DnC	77.43±1.28	64.44±2.12	61.01±1.77	60.66±0.56	57.71±2.34	56.76±0.66	39.54±3.64
	FreqFed	77.31±1.75	73.61±1.15	73.44±1.66	74.46±1.84	72.34±1.18	71.34±1.29	54.76±4.65
	FedDMC	79.76±1.48	72.16±1.76	74.86±1.73	75.34±0.88	73.79±1.29	74.36±1.21	59.63±3.45
	RECESS	79.92±0.97	76.85±1.14	76.14±1.28	77.08±1.05	74.83±1.37	75.62±1.21	66.48±2.56
	SpecShield	**80.46**±0.86	**79.64**±0.87	**78.34**±1.54	**80.94**±1.11	**76.12**±1.57	**78.13**±1.27	**72.34**±2.03
Tiny-ImageNet	Mkrum	41.86±0.93	23.74±1.28	21.63±1.11	17.82±0.97	16.94±1.06	18.36±1.14	20.15±2.47
	Median	43.52±0.88	25.68±1.17	24.47±1.26	18.95±1.08	19.63±0.95	21.24±1.19	22.38±2.31
	Trmean	45.06±0.91	27.91±1.14	26.35±1.02	20.73±0.88	20.14±1.16	23.46±1.08	24.12±2.08
	Bulyan	42.74±0.96	24.36±1.21	20.84±1.35	16.28±1.02	18.47±1.11	17.96±1.24	21.43±2.27
	FLTrust	47.38±0.82	35.62±1.46	33.91±1.37	30.47±1.18	29.86±1.23	31.18±1.29	12.74±3.84
	DnC	46.91±0.95	33.84±1.71	31.77±1.53	29.63±1.07	28.44±1.31	27.95±1.18	24.68±2.96
	FreqFed	49.26±0.87	43.17±1.16	42.34±1.24	43.96±1.08	41.72±1.33	40.85±1.27	31.64±2.78
	FedDMC	50.18±0.79	41.93±1.28	43.28±1.11	44.17±1.02	42.63±1.26	43.54±1.09	34.92±2.41
	RECESS	50.64±0.84	45.82±1.09	44.96±1.22	45.38±1.16	43.91±1.34	44.27±1.18	37.15±2.43
	SpecShield	**51.07**±0.76	**47.26**±1.03	**46.58**±1.15	**48.42**±0.94	**45.76**±1.21	**46.83**±1.07	**40.64**±2.18

datasets under six adaptive poisoning attack vectors. Tab. III systematically quantifies the security-utility tradeoffs.

Statistical aggregations Median and Trimmed Mean often underperform in image classification tasks, particularly under LIE and AGR attacks. These results demonstrate that simple median- or mean-based defenses remain vulnerable to Byzantine attacks. FLTrust generally achieves higher accuracy by suppressing large-magnitude gradients through normalization, thereby outperforming median methods in most cases. In contrast, SpecShield consistently achieves higher accuracy across all benchmark datasets, while competing defenses often degrade substantially under strong poisoning attacks. We attribute this advantage to the DWT-based response profiling adopted in SpecShield. Unlike FreqFed, which applies DCT to static updates, our method analyzes perturbation-response patterns that are often multi-scale and locally distributed. DWT is better suited to capture such variations, producing more discriminative features for separating benign and malicious updates. Interestingly, we observe that under certain attacks, the global model trained with SpecShield achieves even higher accuracy than in the no attack baseline. This indicates that some malicious clients may retain residual utility. It highlights that our clustering aggregation introduces minimal degradation to the convergence behavior of the FL system.

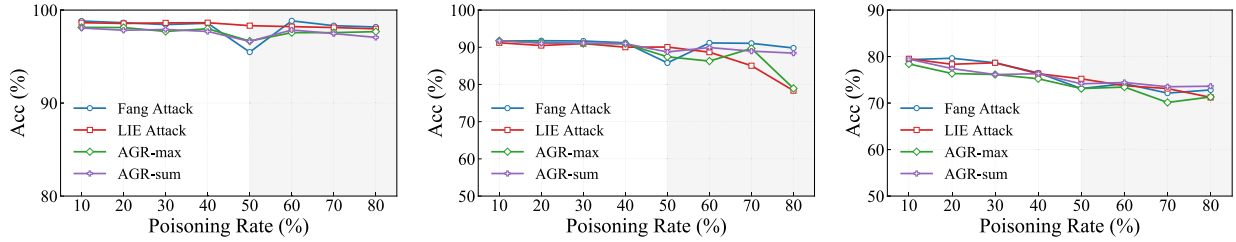
C. TPR and FPR of SpecShield

Tab. IV presents a detailed evaluation of SpecShield's detection performance in terms of FPR and TPR. Our



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Fig. 3. Norm of eigenvector for benign and malicious clients on different datasets.



(a) MNIST (b) Fashion-MNIST (c) CIFAR-10

Fig. 4. The impact of the proportion of attackers on SpecShield.

TABLE IV
COMPARISON OF FPR AND TPR FOR VARIOUS DEFENSES UNDER DIFFERENT ATTACKS

Dataset	Attack	Mkrum		FLTrust		SpecShield	
		FPR	TPR	FPR	TPR	FPR	TPR
MNIST	Fang	0.37	0.71	0.17	0.88	**0.09**	**0.91**
	LIE	0.31	0.76	**0.11**	0.84	0.13	**0.95**
	AGR-tail	0.41	0.67	0.09	0.87	**0.06**	**0.90**
	AGR-max	0.39	0.72	0.15	0.78	**0.11**	**0.92**
	AGR-sum	0.37	0.69	**0.14**	0.81	0.17	**0.89**
Fashion-MNIST	Fang	0.43	0.61	0.21	0.51	**0.14**	**0.84**
	LIE	0.41	0.51	**0.16**	**0.88**	0.19	0.86
	AGR-tail	0.41	0.51	0.23	0.82	**0.14**	**0.89**
	AGR-max	0.43	0.56	0.19	0.68	**0.06**	**0.89**
	AGR-sum	0.40	0.41	0.22	0.89	**0.06**	**0.96**
CIFAR-10	Fang	0.43	0.63	0.23	0.89	**0.21**	**0.95**
	LIE	0.41	0.66	0.22	0.88	**0.17**	**0.91**
	AGR-tail	0.37	0.53	0.27	0.81	**0.13**	**0.90**
	AGR-max	0.44	0.66	0.22	0.88	**0.26**	**0.96**
	AGR-sum	0.40	0.47	0.22	0.66	**0.19**	**0.93**

comparative analysis shows that SpecShield outperforms all baseline defense mechanisms, achieving both lower FPR and higher TPR. Specifically, SpecShield exhibits the most favorable trade-off boundary, with an average FPR of 0.13 and TPR of 0.93 across all evaluation scenarios. This significantly surpasses the performance of FLTrust, which achieves a false positive rate of 0.20 and a true positive rate of 0.81, as well as Mkrum, which yields a false positive rate of 0.40 and a true positive rate of only 0.61. These results confirm SpecShield's superiority in minimizing false alarms while maximizing detection coverage.

D. Eigenvector Divergence

To further substantiate the theoretical foundation of SpecShield, we analyze the divergence in client responses under controlled adversarial perturbations. Fig. 3 illustrates the distribution of eigenvector norms across three datasets using violin plots. On MNIST (Fig. 3a) and Fashion-MNIST (Fig. 3b), benign clients exhibit significantly higher norm values compared to malicious ones. A similar separation is observed on the more complex CIFAR-10 (Fig. 3c), where the mean eigenvector norms of benign and malicious clients are 8.1 and 3.9. These results empirically validate our theoretical insight: Benign clients' optimization objectives remain focused on overall model performance. Thus, they show more pronounced differential responses to the inducement perturbations compared to malicious clients. In all evaluation settings, SpecShield consistently reveals statistically significant distinctions between benign and adversarial populations. Importantly, this separation is achieved without relying on any assumptions about benign majority, client distribution priors, or access to trusted root data.

E. Performance in Various FL Settings

1) Impact of the Number of Malicious Clients: We investigate the impact of varying proportions of malicious clients on the robustness of SpecShield. Fig. 4 presents the model performance as the poisoning rate increases from 10% to 80% across three datasets and four representative poisoning attacks. On MNIST (Fig. 4a), SpecShield exhibits high resilience, maintaining accuracy above 95% even with 50% of clients compromised, and only slightly decreasing to 93% at the 80% poisoning rate. A similar pattern is observed on Fashion-MNIST (Fig. 4b), where accuracy remains above 90% when the poisoning rate exceeds 40%. On the more challenging CIFAR-10 (Fig. 4c), SpecShield continues to show stable behavior, gracefully degrading as the number of malicious clients increases, and effectively avoiding the catastrophic failures commonly seen in traditional defense mechanisms.

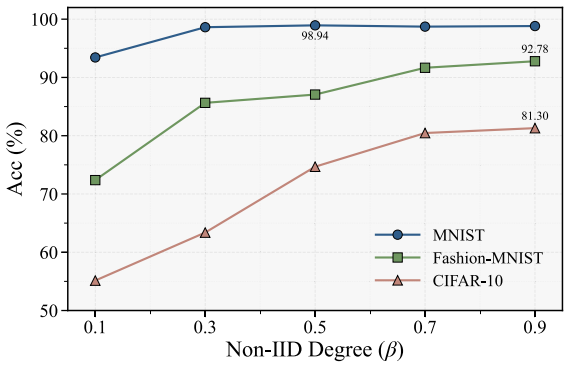
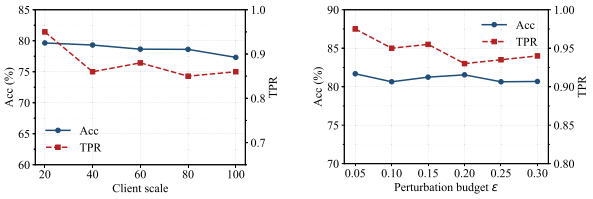


Fig. 5. SpecShield performance across Non-IID settings.



(a) Impact of client scale. (b) Impact of perturbation budget.

Fig. 6. Impact of client scale and perturbation budget.

2) Impact of Different Degrees of Non-IID Data: To evaluate the robustness of SpecShield under varying levels of data heterogeneity, we conduct experiments on three datasets where client data partitions are drawn from Dirichlet distributions. Lower values of β correspond to more extreme non-IID conditions. Fig. 5 illustrates the performance of SpecShield across a range of β values. The results show that SpecShield maintains strong performance even under severe non-IID settings ($\beta = 0.1$). Across all datasets, the most substantial performance gains occur between $\beta = 0.1$ and $\beta = 0.5$, after which performance stabilizes. This robustness can be attributed to SpecShield's distribution-adaptive clustering mechanism, which automatically adjusts to underlying data structures without requiring manual calibration or pre-defined thresholds. As a result, SpecShield generalizes naturally across different levels of data heterogeneity.

3) Impact of the Number of Clients: We vary the total number of clients from small to medium and large pools under the same CIFAR-10 FL setting. As shown in Fig. 6a, SpecShield maintains consistently strong performance as the client population grows. Although accuracy exhibits a slight downward trend at larger scales, the overall degradation remains limited, and the TPR stays at a high level throughout. This robustness comes from the way SpecShield constructs discriminative perturbation-response signatures. As the number of clients increases, benign and malicious responses remain structurally separable in the wavelet domain, which allows DBSCAN to continue identifying suspicious clients effectively.

4) Effect of the Perturbation Budget: We further study the sensitivity of SpecShield to the perturbation budget ϵ. As shown in Fig. 6b, increasing ϵ does not monotonically improve defense performance. While the test accuracy remains relatively stable, the TPR declines when the perturbation budget becomes overly large. This result indicates that stronger probing is not always beneficial, and that an excessive perturbation may weaken the discriminability between benign and malicious clients. This observation highlights the necessity of solving for a proper perturbation bound. In SpecShield, the perturbation budget ϵ must strike a balance between exposing informative response differences and preserving normal optimization dynamics.

TABLE V
IMPACT OF UNTARGETED ADAPTIVE POISONING ATTACKS

Dataset	Acc (%)	TPR	FPR
MNIST	95.22	0.54	0.11
Fashion-MNIST	92.42	0.56	0.28
CIFAR-10	83.36	0.85	0.09
Tiny-ImageNet	44.18	0.92	0.03

F. Defenses Under Strong Adaptive Adversaries

1) Untargeted Adaptive Poisoning Attacks: In a more challenging scenario, an attacker may adaptively modify its attack strategy to evade the defense once the defense mechanism is known. Under this stronger setting, we assume that the attacker has full knowledge of the defense pipeline of SpecShield. The adaptive attack follows the strategy described below. To bypass SpecShield, the attacker crafts its local update so that its probed response signature resembles that of benign clients. Meanwhile, the update is also required to preserve the attacker's original poisoning capability. The attack objective is formulated as:

$$\min_{\theta_i^{\text{mal}}} \mathcal{L}_{\text{poison}}^{(i)} + \lambda \mathcal{L}_{\text{clean}}^{(i)} + \rho \left\| D_i - \bar{D}_{\text{ben}} \right\| \tag{46}$$

where $\mathcal{L}_{\text{poison}}^{(i)}$ denotes the poisoning objective of malicious client i, and $\mathcal{L}_{\text{clean}}^{(i)}$ denotes the loss on clean local data used to preserve benign utility and improve stealthiness. The term $\left\| D_i - \bar{D}_{\text{ben}} \right\|$ measures the discrepancy between the malicious client's differential feature vector and the average benign differential feature vector, where D_i is constructed from the pre-perturbation and post-perturbation model parameters (θ_i, θ_i'), and $\bar{D}_{\text{ben}}$ denotes the mean differential feature vector of benign clients.

Tab. V reports the robustness of SpecShield under a strong adaptive attack. The results show that adaptive evasion can indeed reduce the detection rate, especially on simpler datasets, but SpecShield still maintains high model accuracy and low false-positive rates across all benchmarks. Moreover, on more complex datasets such as CIFAR-10 and Tiny-ImageNet, the defense continues to achieve relatively strong TPR, indicating that its perturbation-response signatures remain informative even against attack-aware adversaries. This behavior is due to the intrinsic tradeoff faced by the adaptive attacker. To evade SpecShield, the attacker must simultaneously preserve poisoning strength, maintain benign-task utility, and imitate benign perturbation responses. These goals are only partially compatible, so reducing one detection signal often comes at the expense of attack effectiveness or

TABLE VI
IMPACT OF TARGETED ADAPTIVE POISONING ATTACKS

PDR	No Attack		No Defense		SpecShield	
	BA	MA	BA	MA	BA	MA
10	0.0	87.0	91.0	86.2	0.0	77.2
15	0.0	87.3	96.9	85.1	0.0	76.7
50	0.0	87.1	97.9	85.4	0.0	75.9
100	0.0	87.4	98.3	85.1	0.0	75.3

TABLE VII
ABLATION OF FREQUENCY-DOMAIN FEATURE EXTRACTORS

Variant	Acc (%)	TPR	FPR
Raw parameter	77.86	0.75	0.25
DCT Variant	73.32	0.50	0.50
SpecShield	81.36	1.00	0.00

optimization stability. Consequently, the attacker can weaken but not fully erase the response discrepancy exploited by SpecShield.

2) Targeted Adaptive Poisoning Attacks: Consistent with FreqFed, we also evaluate the effectiveness of SpecShield against the data poisoning attack proposed by Wang et al. [34]. It injects trigger signals into a specific frequency band of the frequency-domain representation of the data. Tab. VI reports the results in terms of both BA and MA, where BA denotes the Backdoor Accuracy, and MA denotes the Main-task Accuracy. The results show that SpecShield consistently reduces the targeted attack success rate while preserving high clean accuracy. The results indicate that, although the attacker improves stealthiness through adaptive optimization, it still cannot fully eliminate the abnormal perturbation-response discrepancy captured by SpecShield.

G. Ablation Study

To directly examine the importance of the feature extractor, we perform an ablation study under the same proactive perturbation mechanism and clustering method. As shown in Tab. VII, SpecShield achieves the best overall performance, significantly outperforming both the raw-parameter variant and the DCT-based variant in terms of Acc, TPR, and FPR. In particular, replacing DWT with DCT leads to a clear performance drop, while directly using raw parameter differences also yields inferior discrimination. These results show that the advantage of SpecShield does not come merely from proactive probing itself, but critically depends on the DWT-based feature representation.

We attribute this gain to the fact that the perturbation-induced response signal is inherently non-stationary and often localized across parameter blocks. Raw parameter differences are easily affected by noise and lack a compact discriminative structure, while DCT mainly captures global spectral patterns. In contrast, DWT simultaneously retains multi-scale and localized information, making the induced benign–malicious discrepancy more distinguishable in the transformed space.

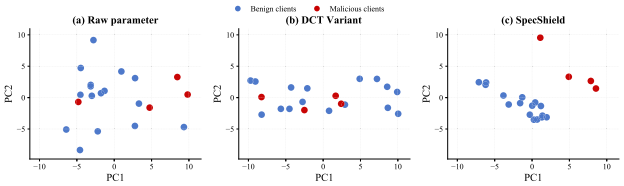


Fig. 7. Ablation of frequency-domain feature extractors.

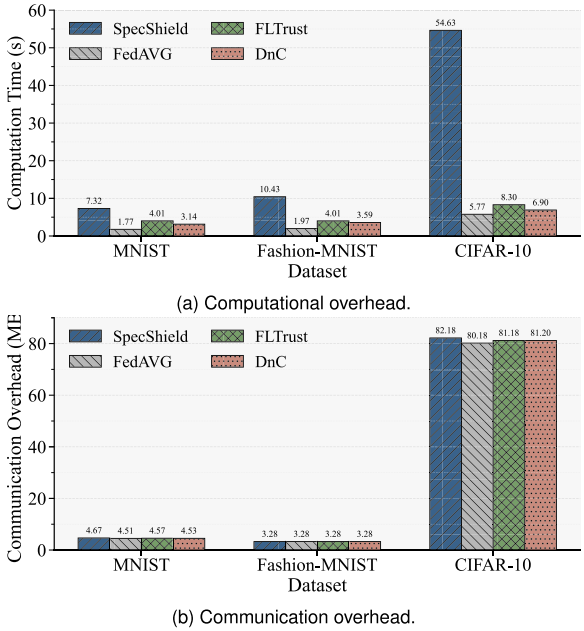


Fig. 8. Computational and communication overheads per round on different datasets.

This effect is also confirmed by the visualization in Fig. 7. Compared with the raw-parameter and DCT-based variants, the SpecShield produces the clearest separation between benign and malicious clients.

H. Computational Cost and Communication Overhead

To assess the practical deployability of SpecShield, we evaluate its computational and communication overhead in comparison with the baseline FL algorithm (FedAvg) and two widely adopted defense mechanisms, FLTrust and DnC. Fig. 8 summarizes the per-round training cost across three benchmark datasets. The additional computational cost introduced by SpecShield primarily arises from the wavelet transform operations and the clustering-based filtering mechanism. Among the evaluated datasets, this overhead is most prominent on CIFAR-10, where SpecShield requires 54.63s per training round, whereas FedAvg completes the same round in only 5.77s, as shown in Fig. 8a. This discrepancy reflects the increased computational demands associated with processing high-dimensional model parameters. Despite the higher computation time, the cost remains within practical bounds for real-world deployment scenarios, especially considering the substantial security enhancements provided by SpecShield. It is important to note that both the frequency-domain analysis and the clustering operations are performed exclusively on the

server side, where computational resources are typically sufficient. Furthermore, these operations are executed only once per communication round, which ensures that the amortized computational cost remains manageable.

The system is designed based on an asymmetric overhead model, where the computational burden is shifted to the server in order to preserve client-side efficiency and limit communication costs. Our analysis indicates that SpecShield accepts a moderate increase in server-side computational cost while incurring no additional communication overhead (Fig. 8b), making it particularly well-suited for deployment in resource-constrained environments where uplink bandwidth is limited but server-side processing capabilities are adequate.

VIII. Conclusion

This paper proposes an anomaly detection mechanism, SpecShield. It establishes a new class of proactive defenses in FL by repurposing adversarial perturbations for active client probing, enabling asymmetry-aware resilience beyond the honest-majority paradigm. Our design integrates three core components: (1) a personalized perturbation mechanism that elicits adversarial deviations, (2) a wavelet-based frequency-domain framework that captures client-specific spectral anomalies, and (3) a clustering-based aggregation scheme that removes the need for statistical priors on client behavior. These are supported by formal theoretical bounds. Extensive experiments demonstrate the robustness of SpecShield under diverse attack vectors and challenging conditions. Future work will explore extending SpecShield to asynchronous FL settings, enabling robust aggregation without synchronization assumptions.

References

[1] K. Bonawitz et al., "Towards federated learning at scale: System design," in *Proc. Conf. Mach. Learn. Syst. (MLSys)*, 2019, pp. 374–388.

[2] H. Xing et al., "Achieving flexible fairness metrics in federated medical imaging," *Nature Commun.*, vol. 16, no. 1, pp. 1–12, Apr. 2025.

[3] F. Zhang et al., "Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare," *Nature Commun.*, vol. 16, no. 1, pp. 1–14, Mar. 2025.

[4] T.-T. Kuo, R. A. Gabriel, J. Koola, R. T. Schooley, and L. Ohno-Machado, "Distributed cross-learning for equitable federated models–privacy-preserving prediction on data from five California hospitals," *Nature Commun.*, vol. 16, no. 1, pp. 1–17, Feb. 2025.

[5] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 1605–1622.

[6] V. Shejwalkar and A. Houmansadr, "Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2021, pp. 1–18.

[7] C. Xie, K. Huang, P. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–18.

[8] W. Shen, W. Huang, G. Wan, and M. Ye, "Label-free backdoor attacks in vertical federated learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, vol. 39, no. 19, pp. 20389–20397.

[9] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2018, pp. 5650–5659.

[10] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in Byzantium," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3521–3530.

[11] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2021, pp. 1–18.

[12] H. Fereidooni, A. Pegoraro, P. Rieger, A. Dmitrienko, and A.-R. Sadeghi, "FreqFed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2024, pp. 1–16.

[13] N. M. Jebreel and J. Domingo-Ferrer, "FL-defender: Combating targeted attacks in federated learning," *Knowledge-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110178.

[14] Y. Liu, C. Chen, L. Lyu, Y. Jin, and G. Chen, "Exploit gradient skewness to circumvent Byzantine defenses for federated learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2025, vol. 39, no. 18, pp. 19024–19032.

[15] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8635–8645.

[16] A. Yazdinejad, A. Dehghantanha, H. Karimipour, G. Srivastava, and R. M. Parizi, "A robust privacy-preserving federated learning model against model poisoning attacks," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6693–6708, 2024.

[17] Z. Ma, J. Ma, Y. Miao, Y. Li, and R. H. Deng, "ShieldFL: Mitigating model poisoning attacks in privacy-preserving federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1639–1654, 2022.

[18] A. Lad, R. Bhale, and S. Belgamwar, "Fast gradient sign method (FGSM) variants in white box settings: A comparative study," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Apr. 2024, pp. 382–386.

[19] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, no. 4, pp. 628–666, Dec. 1989.

[20] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.

[21] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. NIPS*, 2017, pp. 118–128.

[22] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proc. ACM Meas. Anal. Comput. Syst.*, pp. 1–25, 2017.

[23] D. N. Yaldiz, T. Zhang, and S. Avestimehr, "Secure federated learning against model poisoning attacks via client filtering," in *Proc. ICLR*, 2023, pp. 1–10.

[24] H. Yang, W. Xi, Y. Shen, C. Wu, and J. Zhao, "RoseAgg: Robust defense against targeted collusion attacks in federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2951–2966, 2024.

[25] G. Yan, H. Wang, X. Yuan, and J. Li, "FedRoLA: Robust federated learning against model poisoning via layer-based aggregation," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2024, pp. 3667–3678.

[26] E. Kabir, Z. Song, M. R. Ur Rashid, and S. Mehnaz, "FLShield: A validation based federated learning framework to defend against poisoning attacks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2024, pp. 2572–2590.

[27] X. Mu et al., "FedDMC: Efficient and robust federated learning via detecting malicious clients," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 6, pp. 5259–5274, Nov. 2024.

[28] S. Shen, S. Tople, and P. Saxena, "AUROR: Defending against poisoning attacks in collaborative deep learning systems," in *Proc. 32nd Annu. Conf. Comput. Secur. Appl.*, 2016, pp. 508–519.

[29] T. D. Nguyen et al., "FLAME: Taming backdoors in federated learning," in *Proc. 31st USENIX Security Symp. (USENIX Secur.)*, 2022, pp. 1415–1432.

[30] V. Tolpegin, S. Truex, M. E. Gürsoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Proc. Eur. Symp. Res. Comput. Secur.*, 2020, pp. 480–501.

[31] H. Guo et al., "Siren+: Robust federated learning with proactive alarming and differential privacy," *IEEE Trans. Dependable Secure Comput. (TDSC)*, vol. 21, no. 5, pp. 4843–4860, May 2024.

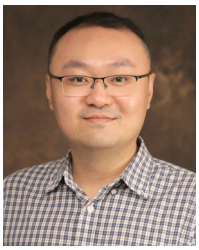
[32] Q. Chen et al., "RECESS vaccine for federated learning: Proactive defense against model poisoning attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 8702–8713.

[33] K. Naveen Kumar, C. Krishna Mohan, and L. Reddy Cenkeramaddi, "Federated learning minimal model replacement attack using optimal transport: An attacker perspective," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 478–487, 2025.

[34] T. Wang, Y. Yao, F. Xu, S. An, H. Tong, and T. Wang, "An invisible black-box backdoor attack through frequency domain," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 396–413.

Fangjie Hu received the B.E. degree in electronic information engineering from Anhui Normal University, Wuhu, China, in 2023, where he is currently pursuing the M.S. degree in electronic science and technology with the School of Physical and Electronic Information Engineering. His research interests include applied cryptography and federated learning.

Aiqing Zhang (Member, IEEE) received the M.S. degree in circuits and systems from Xiamen University, China, in 2006, and the Ph.D. degree in signal and information processing from Nanjing University of Posts and Telecommunications, China, in 2016. She was a Visiting Scholar with the University of Ontario Institute of Technology, Canada, in 2016. She is currently a Professor with Anhui Normal University, China. She has authored more than 40 technical articles, including top journals, such as IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING. Her research interests include privacy preservation, federated learning, and wireless network security.



Meng Li (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from the School of Computer Science and Technology, Beijing Institute of Technology (BIT), China, in 2019. He is currently a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT), China. He was a Post-Doctoral Researcher with the Department of Mathematics and the HIT Center, University of Padua, Italy, where he is with the Security and PRIvacy Through Zeal (SPRITZ) Research Group led by Prof. Mauro Conti (IEEE Fellow). He was sponsored by the ERCIM "Alain Bensoussan" Fellowship Programme to conduct a Post-Doctoral Research supervised by Prof. Fabio Martinelli at CNR, Italy, from October 2020 to March 2021. He was sponsored by China Scholarship Council (CSC) as a Joint Ph.D. Student supervised by Prof. Xiaodong Lin (IEEE Fellow) with the Broadband Communications Research (BBCR) Laboratory, University of Waterloo, and Wilfrid Laurier University, Canada, from September 2017 to August 2018. He is supported by CSC as a Visiting Scholar collaborating with Prof. Mauro Conti (IEEE Fellow) at the HIT Center, University of Padua, Italy, from March 2025 to June 2025. In December 2025, he was promoted to a Professor. His research interests include security, privacy, applied cryptography, blockchain, TEE, and the Internet of Vehicles. In this area, he has published 150 papers in top most journals and conferences, including IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *ACM Transactions on Database Systems*, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, *ACM Transactions on Social Computing*, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE S&P, USENIX Security, MobiCom, INFOCOM, and ISSTA. He is a Senior Member of CIE, CIC, and CCF. He is an Associate Editor for IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT.

Chen Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Automation, Wuhan University, China, in 2008 and 2013, respectively. From 2013 to 2017, he was a Post-Doctoral Research Fellow with the Networked and Communication Systems Research Laboratory, Huazhong University of Science and Technology, China. Thereafter, he joined the Huazhong University of Science and Technology, where he is currently an Associate Professor. His research interests include wireless networking, the Internet of Things, and mobile computing, with a recent focus on privacy issues in intelligent systems. He is a Senior Member of ACM.