A Federated Learning Backdoor Attack Defense Method Based on Frequency Domain Perturbations and Model Repair

1st Jinquan Zhang, 2nd Qingmian Wei, 3rd Lina Ni*
College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao, China
zhangjinquan@sdust.edu.cn, 17860848190@163.com, nilina@sdust.edu.cn

***Abstract*—Federated Learning (FL) enables privacy-preserving distributed modeling but remains vulnerable to backdoor attacks, where adversaries embed triggers to induce targeted misclassification while keeping normal accuracy intact. Existing defenses, such as robust aggregation or cosine-similarity–based gradient filtering, cannot reliably distinguish malicious clients from benign ones and often discard suspicious updates in a coarse-grained manner, which may inadvertently remove legitimate clients and degrade model performance. To address these limitations, we propose FedFdpm, a novel defense method based on frequency-domain perturbations and model repair. FedFdpm detects suspicious client models by analyzing their frequency spectrum responses and applies flexible repairs to align their distributions with normal patterns, reducing aggregation risks without discarding clients. It employs dynamic aggregation mechanism to assign weights to clients for robust aggregation. Experimental results on three public datasets, compared to five baseline algorithms, show that FedFdpm significantly reduces the success rate of backdoor attacks (up to 96%) and maintains high accuracy (98%) in the main task, outperforming existing methods.**

***Index Terms*—Federated learning, backdoor attacks, spectrum transformation, security**

I. Introduction

Federated Learning (FL), as a novel paradigm of distributed collaborative machine learning, has demonstrated strong potential in safeguarding data privacy while enhancing model generalization capabilities [1], [2]. Unlike traditional centralized models that aggregate raw data into a single repository, FL enables participants to keep their data local and share only model parameters or gradients for joint training [3], [4]. By doing so, it effectively mitigates privacy risks during transmission and helps to break long-standing data silos, which in turn has supported the adoption of AI in privacy-sensitive areas such as medical diagnostics, financial risk control, and smart city development [5]. At the same time, the decentralized structure that underpins FL inevitably creates new security challenges, among which backdoor attacks are particularly critical [6], [7]. In these attacks, adversaries inject malicious samples with specific trigger patterns into local training and manipulate their labels toward target categories, leading the global model to misclassify inputs containing the trigger into attacker-specified outputs while still maintaining high accuracy on clean data [8], [9]. This type of attack is stealthy, difficult to detect, and has strong generalization and transferability, posing a serious threat to the reliability of FL systems, particularly in critical applications such as autonomous driving and medical-assisted diagnostics, where its potential harm cannot be ignored.

To defend against backdoor attacks in FL, researchers have proposed various strategies, which can be broadly categorized into model detection and aggregation optimization. In model detection, some studies focus on identifying anomalous behaviors in client-uploaded models by analyzing training dynamics, output stability, and local model similarity, thereby detecting potential backdoors [6], [10]. Other studies optimize global aggregation strategies by using robust statistical methods, clustering mechanisms, or weight adjustments to reduce the impact of anomalous updates on the global model [11] [12] [13]. These methods provide valuable insights and significantly mitigate the risk of backdoor attacks in FL. Despite the advantages of the aforementioned methods, there are still key shortcomings. Most detection methods rely on explicit validation sets or predefined distribution assumptions, making them difficult to deploy widely in practice, and they struggle to effectively address backdoor paths and structural perturbations hidden in the frequency domain by attackers. Additionally, some methods directly exclude suspected malicious clients once identified, preventing them from participating in aggregation. This "one-size-fits-all" approach may unintentionally remove legitimate clients with atypical data distributions. Recently, frequency-domain defenses have emerged, such as FreqFed [14], which observes that backdoored models are more sensitive to spectral perturbations and utilizes this property for detection. While effective, FreqFed still follows a detect-and-discard paradigm, directly excluding suspicious clients and potentially removing benign ones with atypical data distributions. These issues not only limit the effectiveness of existing methods but also drive us to seek a more flexible defense mechanism.

To address this, we propose FedFdpm, a FL backdoor defense method based on frequency-domain perturbations and model repair. FedFdpm characterizes model stability and potential anomalies from a frequency perspective and builds an integrated defense mechanism combining detection, repair, and aggregation optimization. In the model detection process, FedFdpm employs a frequency-domain consistency evaluation index to assess client models' responses to specific frequency perturbations, thereby identifying suspicious models with unstable spectral features. For these models, the method applies flexible structural repairs based on a global spectral energy template, adjusting their frequency distribution to align

with normal client features, thereby reducing security risks without fully excluding them. Moreover, FedFdpm leverages frequency response to optimize aggregation, assigning differentiated weights to client models and guiding parameter integration toward a more secure outcome. The method does not require access to raw data, ensuring privacy compatibility and generalization, offering a new approach to enhance the robustness of FL systems in complex threat scenarios. The main contributions of this paper are summarized as follows:

- We propose FedFdpm, a novel backdoor defense framework for federated learning that integrates frequency-domain perturbation consistency detection, spectral-domain model repair, and dynamic aggregation, enabling effective mitigation of backdoor attacks without directly discarding clients.
- We introduce a perturbation-induced frequency consistency metric to characterize model vulnerability in the spectral domain, which provides a fine-grained and data-free mechanism to identify suspicious client updates.
- Extensive experiments on three public datasets under multiple attack scenarios, including recent and strong attacks, demonstrate that FedFdpm significantly reduces attack success rates while maintaining high accuracy.

II. PROPOSED METHOD

A. FedFdpm Model

The overall process of FedFdpm is shown in Fig. 1. FedFdpm consists of a cloud server $\mathcal{S}$ and client set $\mathcal{C} = \{c_1, c_2, ...c_C\}$. The operation flow of FedFdpm is as follows.

1) Initialization of the global model. The server $\mathcal{S}$ initializes the model parameters ω^0, and then distributes them to all client $c \in \mathcal{C}$.

2) Local training on the client. The client c conducts local training on the global model ω^0 based on its own data to generate the local model ω_c. For benign clients $c \in \mathcal{C}^g$, the model is trained by minimizing the local dataset loss $L_c(\omega)$ over the dataset D_c, with the objective function defined as:

$$\omega_c = \arg\min_{\omega} L_c(\omega) = \frac{1}{|D_c|} \sum_{(x,y)\in D_c} f(x, y, \omega) \tag{1}$$

where $f(x, y, \omega)$ represents the loss function on the sample (x, y).

For malicious clients, the core objective of a backdoor attack is to maintain the global model's performance on normal inputs while allowing the attacker to predict errors on specific inputs by injecting a trigger, thus achieving covert control over the target class. The attacker injects malicious samples into the local model during training, forming a set of malicious clients $\mathcal{C}^b \subset \mathcal{C}$, where $\mathcal{C}^b = \mathcal{C} \setminus \mathcal{C}^g$. Each client aims to generate a backdoored model ω_c by training on its local dataset D_c, so that when encountering inputs with a trigger, the model outputs a target label y_b. The objective function of the malicious client $c \in \mathcal{C}^b$local model is:

$$\omega_c = \arg\min_{\omega}(1-\beta)L_{benign}(\omega; D_c) + \beta L_{backdoor}(\omega; D_c^b), \tag{2}$$

where $\beta \in [0, 1]$ is the attack strength adjustment factor, L_{benign} is the loss for normal training, and $L_{backdoor}$ is the loss for the backdoor attack. The dataset D_c^b represents the set of poisoned samples with backdoor triggers that are uploaded by malicious client c.

3) Upload local model. After completing the local training, the client uploads the updated model parameters ω_c to the $\mathcal{S}$ for global aggregation.

4) Model aggregation and detection: After receiving all local models uploaded by the clients, the server executes the frequency-domain perturbation consistency detection mechanism to identify and distinguish suspicious clients. For suspicious clients, FedFdpm performs a spectral energy repair mechanism, adjusting their models in the frequency domain to achieve consistency and assigning appropriate weights γ_c based on evaluation. For benign clients, their weights are kept unchanged. Finally, all client models and their weights are combined through a dynamic aggregation mechanism to generate the global model ω^{new}:

$$\omega^{new} = \sum_{c\in\mathcal{C}} \omega_c \times \gamma_c. \tag{3}$$

The above steps iterate for T rounds or until the global model meets the desired criteria, then terminate.

B. Frequency-Domain Perturbation Consistency Detection Mechanism

Backdoor attacks increase the model's "vulnerability" in specific frequency regions; this vulnerability often leads to more significant fluctuations in the model's response to disturbances in the frequency domain [14]. Unlike FreqFed, which performs static spectral analysis on model updates, we characterize such vulnerability through perturbation-induced frequency response consistency, enabling a finer-grained and model-agnostic detection mechanism. We quantify the model's output stability to identify potential backdoor updates. We add a small random perturbation ϵ to the local model ω_i^t uploaded by client i in the t-th round, generating the perturbed model $\tilde{\omega}_i^t = \omega_i^t + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then, we select a set of reference samples $\mathcal{X}_{\text{ref}}$, and calculate the following two outputs:

$$f_i(x) = f(x; \omega_i^t), \quad \tilde{f}_i(x) = f(x; \tilde{\omega}_i^t). \tag{4}$$

Since the input is consistent, the output difference is entirely due to the change in the model's structure. We perform a fast fourier transform (FFT) on $f_i(x)$ and $\tilde{f}_i(x)$:

$$F_i(x) = \mathcal{F}(f_i(x)), \quad \tilde{F}_i(x) = \mathcal{F}(\tilde{f}_i(x)). \tag{5}$$

We then calculate the frequency-domain consistency evaluation index s_i^{freq} based on $F_i(x)$ and $\tilde{F}_i(x)$:

$$s_i^{\text{freq}} = \frac{1}{|\mathcal{X}_{\text{ref}}|} \sum_{x\in\mathcal{X}_{\text{ref}}} \left\| F_i(x) - \tilde{F}_i(x) \right\|_2 \tag{6}$$

where $|\cdot|_2$ represents the L2 norm. A larger value indicates that the model is more sensitive to perturbations and may be a backdoor model. Based on s_i^{freq}, we set an adaptive threshold:

$$\tau_{freq} = \mu_s + 2 \cdot \sigma_s \tag{7}$$

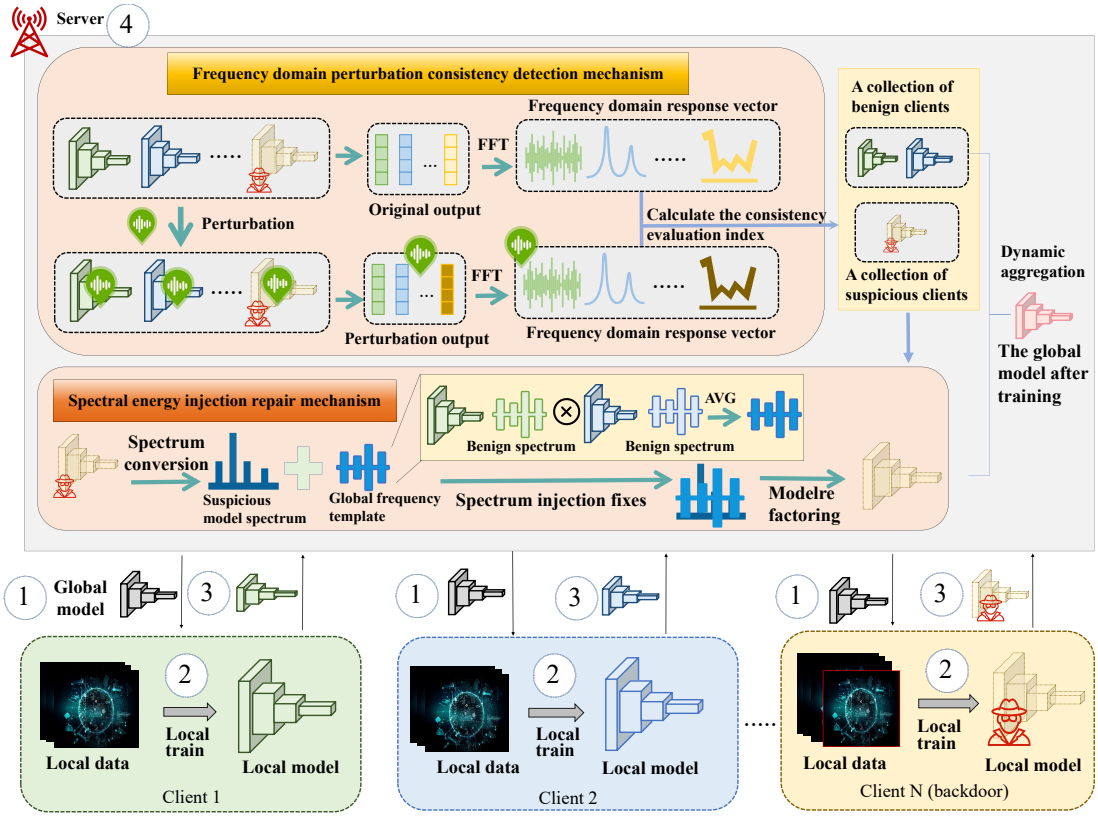

Fig. 1. Schematic of FedFdpm.

where μ_s and σ_s represent the mean and standard deviation of the client frequency-domain consistency evaluation index If the client i satisfies: $s_i^{\text{freq}} > \tau_{\text{freq}}$, then it will be classified as a suspicious client and added to the suspicious client set: $i \in \mathcal{C}_{\text{suspect}}^t$.

C. Spectral Energy Injection Repair Mechanism

Directly removing suspicious clients may enhance security but weaken the model's generalization ability; therefore, we propose repairing suspicious models. For each client $i \in C_{suspect}^t$ identified as suspicious in the t-th communication round, we apply a spectral transformation to its uploaded model parameters w_i^t, obtaining the frequency-domain representation W_i^t:

$$W_i^t = T(w_i^t), \tag{8}$$

where $T(\cdot)$ denotes the Discrete Cosine Transform (DCT), which maps the model parameters from the spatial domain to the frequency domain. The detection stage requires FFT for fine-grained full-spectrum information, while the repair stage relies more on the energy compaction and real-valued advantages of DCT. Next, we extract the frequency structures from the set of benign clients C_{benign}^t identified in the same round to construct the global frequency template $\overline{W}^t$:

$$\overline{W}^t = \frac{1}{|C_{benign}^t|} \sum_{j \in C_{benign}^t} T(w_j^t), \tag{9}$$

where $\overline{W}^t$ represents the average frequency structure, reflecting the typical normal frequency distribution of models in the current communication round.

For each suspicious client $i \in C_{suspect}^t$, we perform a linearly fusion of its frequency representation W_i^t with the global template $\overline{W}^t$:

$$\widetilde{W}_i^t = \alpha \cdot W_i^t + (1-\alpha) \cdot \overline{W}^t, \tag{10}$$

where $\alpha \in [0,1]$ controls the proportion between the original frequency and the global template.

Subsequently, we apply the inverse DCT to map the rectified frequency representation back to the parameter space, obtaining the repaired client model:

$$\hat{w}_i^t = T^{-1}(\widetilde{W}_i^t), \tag{11}$$

where $T^{-1}(\cdot)$ denotes the inverse DCT operation.

D. Dynamic Aggregation Mechanism

The repaired suspicious model may still have residual anomalies, so it is necessary to suppress the influence of backdoors by assigning differentiated aggregation weights. For each suspicious client $i \in C_{suspect}^t$, we calculate its spectral response score as:

$$s_i^r = \left\| T(w_i^t) - T(\hat{w}_i^t) \right\|_2, \tag{12}$$

where the score quantifies the difference between the original and repaired frequency representations. Subsequently, we allocate weights to all clients:

$$\gamma_i^t = \begin{cases} 1, & i \in \mathcal{C}_{\text{benign}}^t \\ \exp\left(-\hat{s}_i^r \cdot s_i^{\text{freq}}\right), & i \in \mathcal{C}_{\text{suspect}}^t \end{cases} \tag{13}$$

Then, normalize it to obtain the final aggregated weight:

$$\gamma_i^t = \frac{\gamma_i^t}{\sum_{j \in \mathcal{C}} \gamma_j^t}. \tag{14}$$

Overall, directly discarding all suspicious clients is undesirable in federated learning, especially under non-IID settings, where even backdoor clients are partially trained on genuine and potentially rare data. Blind removal may therefore eliminate valuable information and degrade the model's generalization ability [15]. FedFdpm instead repairs suspicious updates by aligning their spectral energy with a benign template and then down-weights them during aggregation, limiting any residual backdoor influence while retaining useful information.

Algorithm 1 FedFdpm: End-to-End Federated Training with Frequency-Domain Detection and Repair

Input: Client set $\mathcal{C}$; total rounds T; initial model ω^0; reference samples $\mathcal{X}_{ref}$; perturbation magnitude σ; fusion coefficient α.
Output: Final global model ω^T.

Server initializes global model ω^0;
for $t = 0$ to $T-1$ **do**
  // Broadcast
  Server sends ω^t to all clients $c \in \mathcal{C}$;
  // Local training and upload
  for each client $c \in \mathcal{C}$ **do**
    Client trains locally on D_c and obtains ω_c^t using Eq.(1)–Eq.(2);
    Client uploads ω_c^t to the server;
  end for
  // Detection
  Compute frequency consistency scores $\{s_c^{freq}\}$ on $\mathcal{X}_{ref}$ using Eq.(4)–Eq.(7);
  Split clients into $\mathcal{C}_{suspect}^t$ and $\mathcal{C}_{benign}^t$ by τ_{freq};
  // Repair and reweight
  Construct benign spectral template from $\mathcal{C}_{benign}^t$ using Eq.(8)–Eq.(9);
  Repair models in $\mathcal{C}_{suspect}^t$ and compute residual scores using Eq.(10)–Eq.(12);
  Assign and normalize aggregation weights $\{\gamma_c^t\}$ using Eq.(13)–Eq.(14);
  // Aggregation
  Update global model $\omega^{t+1} \leftarrow \sum_{c \in \mathcal{C}} \gamma_c^t \omega_c^t$ using Eq.(3);
end for
return ω^T;

E. The Overall Process of FedFdpm and Its Time Complexity

Algorithm 1 presents the end-to-end training procedure of FedFdpm. In each round, the server broadcasts the current global model, clients perform local training and upload updated models, and the server detects suspicious updates via perturbation-induced frequency consistency. Suspicious models are then repaired in the frequency domain using a benign spectral template, and dynamically down-weighted during aggregation to limit any residual backdoor influence.

Let $C = |\mathcal{C}|$ be the number of participating clients per round, T the total number of communication rounds, P the number of model parameters, and $R = |\mathcal{X}_{ref}|$ the size of the reference set. Let E denote the number of local epochs, B the local batch size, and K the cost of one forward/backward pass on the model. **Client-side cost.** In each round, each client performs standard local training, whose complexity is dominated by backpropagation and is approximately $O(E \cdot \frac{|D_c|}{B} \cdot K)$ for client c. Summed over all clients, the per-round client-side computation is $O\Big(\sum_{c \in \mathcal{C}} E \cdot \frac{|D_c|}{B} \cdot K\Big)$, and the total client-side cost across T rounds is $O\Big(T \cdot \sum_{c \in \mathcal{C}} E \cdot \frac{|D_c|}{B} \cdot K\Big)$. **Server-side cost.** In each round, (i) detection requires evaluating each client model on $\mathcal{X}_{ref}$ twice (before and after perturbation), resulting in $O(C \cdot R \cdot K_f)$ where K_f is the cost of one forward pass, plus a lightweight FFT-based scoring overhead; (ii) repair constructs a benign spectral template and performs DCT/IDCT-based repair, which is upper bounded by $O(C \cdot P \log P)$ in the worst case (when all clients are suspicious), while in practice it scales with $O(|\mathcal{C}_{suspect}^t| \cdot P \log P)$; and (iii) reweighting and weighted aggregation add $O(C) + O(C \cdot P)$. Therefore, the per-round server overhead is dominated by $O(C \cdot R \cdot K_f + C \cdot P \log P)$, and the total server-side cost over T rounds is $O(T \cdot (C \cdot R \cdot K_f + C \cdot P \log P))$. Combining both sides, the overall computational complexity of Algorithm 1 is: $O\Big(T \cdot \sum_{c \in \mathcal{C}} E \cdot \frac{|D_c|}{B} \cdot K \; + \; T \cdot (C \cdot R \cdot K_f + C \cdot P \log P)\Big)$, where the first term corresponds to standard FL local training and the second term represents the additional server-side overhead introduced by FedFdpm.

In practice, the additional overhead introduced by FedFdpm is mainly affected by the reference set size R, the model dimension P, and the number of suspicious clients $|\mathcal{C}_{suspect}^t|$. Since R is typically small and only suspicious updates undergo frequency-domain repair, the extra cost remains moderate compared to standard FL training. Moreover, the server-side overhead can be further reduced by limiting the size of $\mathcal{X}_{ref}$, performing detection on a subset of clients, or applying repair only every few rounds, making FedFdpm efficient and scalable in real-world deployments.

III. EXPERIMENTS

A. Experiment Setup

Experiments are implemented in Python 3.9 with PyTorch, simulating a typical FL setting with one server and 100 clients. Malicious clients (default 30%) perform three types of attacks:

TABLE I
COMPARISON OF ACC (%) AND ASR (%) UNDER DIFFERENT BASELINES ACROSS THREE DATASETS WITH τ FROM 0% TO 90%

Dataset	Baseline	τ=0%		τ=10%		τ=20%		τ=30%		τ=40%		τ=50%		τ=60%		τ=70%		τ=80%		τ=90%	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
EDGE_IIOT	RFA	97.01	2.11	96.99	34.78	95.82	73.47	95.65	93.55	95.47	95.87	96.23	98.89	94.91	99.78	94.63	99.56	95.62	99.49	95.31	99.78
	RLR	96.21	2.02	96.62	3.23	96.07	5.58	96.78	8.72	95.98	16.58	97.01	36.28	97.22	42.58	97.31	50.36	96.88	66.39	96.53	78.87
	Freqfed	96.78	1.92	97.90	0.78	**98.21**	1.03	97.92	0.68	97.54	1.69	97.73	2.60	98.07	1.97	97.42	3.95	96.86	8.12	97.32	10.89
	FedID	97.92	1.71	97.61	2.35	97.38	2.47	97.32	1.98	96.92	1.35	97.21	1.89	97.10	2.39	96.88	2.98	97.08	2.42	95.93	2.93
	BADFL	98.23	1.53	98.13	0.05	97.99	0.08	**98.61**	**0.01**	98.22	**0.15**	**98.65**	0.66	**98.32**	0.98	**98.33**	**1.31**	**98.10**	1.39	98.17	**1.36**
	FedFdpm	**98.33**	**1.34**	**98.15**	**0.03**	98.00	**0.05**	98.01	0.12	**98.23**	0.18	97.85	**0.22**	97.69	**0.85**	98.20	1.55	97.66	**1.15**	**98.20**	1.66
FEMNIST	RFA	**81.07**	3.68	78.99	99.72	79.78	99.98	78.43	100	77.56	100	78.86	100	77.45	100	78.36	100	78.68	100	77.69	100
	RLR	74.98	5.68	75.65	5.98	77.53	25.87	71.34	36.72	73.62	55.65	74.58	63.98	74.97	75.25	73.89	79.82	71.25	83.64	72.68	88.52
	Freqfed	80.22	5.38	80.12	0.83	**80.68**	0.95	79.85	0.83	79.32	1.05	79.59	1.33	**79.84**	1.51	79.13	1.62	78.88	5.12	78.97	15.67
	FedID	76.72	1.89	74.87	2.45	79.21	2.19	78.58	2.55	77.37	2.13	76.87	1.87	76.43	2.01	76.18	2.78	75.34	2.41	74.13	3.55
	BADFL	80.32	**0.52**	80.46	0.35	80.19	0.12	79.58	**0.06**	**80.36**	0.75	**80.78**	0.68	79.13	**1.02**	**80.12**	1.65	80.25	1.87	80.01	2.01
	FedFdpm	80.16	1.34	**80.55**	**0.15**	79.98	**0.11**	**80.33**	0.35	80.25	**0.52**	80.37	**0.38**	79.46	1.45	79.39	**1.49**	**80.32**	**1.33**	**80.11**	**1.92**
CIFAR-10	RFA	**82.32**	**1.62**	81.12	97.47	79.78	99.66	80.73	99.28	80.21	99.67	79.58	99.69	**81.36**	99.82	**83.17**	99.87	**81.34**	99.87	**80.91**	99.88
	RLR	75.36	3.85	73.21	0.27	71.31	1.22	75.68	13.65	76.34	38.55	78.63	62.68	75.12	75.78	76.13	89.66	77.41	96.31	73.35	95.33
	Freqfed	72.10	3.50	71.30	1.68	72.31	1.01	70.64	1.66	71.33	1.98	70.43	1.32	72.32	1.98	70.32	2.23	70.55	4.66	71.38	8.65
	FedID	78.89	5.65	**81.67**	2.21	**80.99**	1.77	**81.89**	1.55	**80.93**	1.89	79.78	1.68	76.48	2.65	75.86	3.35	74.97	4.15	74.79	4.27
	BADFL	78.32	2.58	75.16	**0.03**	76.35	0.12	77.35	**0.08**	78.31	0.56	75.68	0.77	77.68	0.89	76.32	1.02	75.13	1.32	73.56	1.99
	FedFdpm	78.65	4.65	79.92	0.15	79.93	**0.11**	80.69	1.13	79.65	**0.38**	**79.93**	**0.62**	80.32	**0.76**	76.32	**0.68**	77.42	**1.22**	74.85	**0.97**

centralized backdoor attack (CBA) [16], distributed backdoor attack (DBA) [10], or the Bad-MFL attack [17], with 50% poisoned local data. Each malicious client randomly selects one of these three attack types. Client data is partitioned non-iid using a Dirichlet distribution (Coefficient is 0.8), and training lasts 100 communication rounds. Three datasets are used: EDGE_IIOT [18], FEMNIST [15], and CIFAR-10 [15]. Models are LSTM, a two-layer CNN, and ResNet-18, respectively. All clients use SGD (lr=0.01, batch size=10), $\alpha = 0.5$. We evaluate FedFdpm with four metrics: accuracy on clean data (ACC), attack success rate (ASR), true positive rate (TPR), and true negative rate (TNR). Baselines include RFA [12], RLR [19], Freqfed [14], FedID [20], and BADFL [13]. For all baselines, we follow the original papers and additionally perform grid search over their key hyperparameters within the recommended ranges to ensure a fair and competitive comparison. For FedFdpm, the perturbation noise magnitude σ is set to 10^{-3} and the spectral fusion coefficient α is set to 0.5 by default.

B. Experimental Results

Performance comparison: We assume the proportion of malicious clients to the total number of clients as τ. As shown in Table I, FedFdpm consistently achieves optimal or near-optimal performance across most attack ratios and task scenarios, attaining higher ACC and the lowest or near-lowest ASR, thereby demonstrating stability and robustness. In contrast, methods such as RFA, RLR, and Freqfed suffer sharp accuracy drops under high attack ratios, while BADFL and FedID show good performance in certain cases but lack overall stability compared with FedFdpm. In summary, FedFdpm effectively mitigates rapid degradation under high attack ratios and provides stronger general defense capability. FedFdpm can still defend against over 50% of backdoor attacks due to its robust frequency-domain repair and aggregation mechanism. Although the global spectral template may be contaminated when malicious clients are included in the benign set, the weighted aggregation minimizes their influence by down-weighting their contribution to the global model. The repair process aligns suspicious models with the global template, and the weighting mechanism ensures that legitimate updates dominate, effectively suppressing the impact of malicious updates. This enables FedFdpm to maintain robustness even when more than 50% of the clients are compromised.

Ability to resist backdoor attacks: We conducted a comparative analysis of TPR and TNR on the CIFAR-10 dataset under four poisoning ratios (0.2, 0.4, 0.6, and 0.8) in the local data of backdoor clients, evaluating four representative defense methods (Freqfed, FedID, BADFL, and FedFdpm), with the results shown in Fig. 2. In most poisoning ratios, FedFdpm consistently achieves the highest or near-highest TPR and TNR, demonstrating significant superiority. In contrast, BADFL and Freqfed exhibit a clear decline in recognition capability under high poisoning ratios, while FedID shows considerable fluctuations, further highlighting the robustness and general defense capability of FedFdpm in complex attack scenarios.

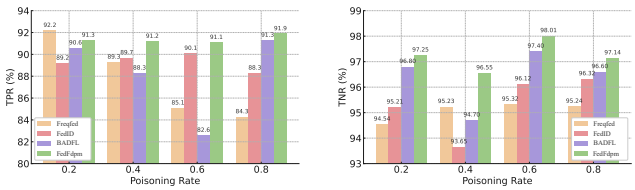

Fig. 2. Comparison of TPR(%) and TNR (%) under different baselines with client data poisoning rates.

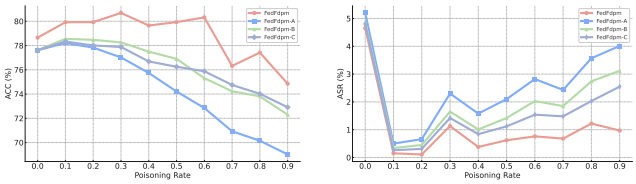

Fig. 3. Comparative performance of FedFdpm and its variants.

Ablation study: We designed three ablation variants (FedFdpm-A, FedFdpm-B, FedFdpm-C), which respectively remove or simplify the detection, recovery, and aggregation mechanisms, and conducted comparisons with the complete model on the CIFAR-10 dataset. As shown in Fig. 3, the

complete FedFdpm model demonstrates superior stability and robustness against backdoor attacks of varying intensities, indicating that each sub-module plays a critical role in enhancing its performance.

TABLE II
SENSITIVITY TO PERTURBATION NOISE MAGNITUDE σ ON CIFAR-10.

σ	ACC (%)	ASR (%)
1×10^{-4}	79.6	2.10
5×10^{-4}	79.8	1.28
1×10^{-3}	79.7	0.62
5×10^{-3}	79.4	0.74
1×10^{-2}	78.9	1.36

TABLE III
SENSITIVITY TO SPECTRAL FUSION COEFFICIENT α ON CIFAR-10.

α	ACC (%)	ASR (%)
0.1	78.8	0.18
0.3	79.3	0.26
0.5	79.7	0.62
0.7	79.9	1.05
0.9	79.0	1.92

Hyperparameter sensitivity: We analyze two key hyperparameters in FedFdpm: the perturbation noise magnitude σ used in frequency-domain detection, and the spectral fusion coefficient α used in model repair. Tables II and III illustrate the sensitivity of FedFdpm to the two key hyperparameters. As shown in Table II, a very small σ leads to weak discriminability and higher ASR, while an overly large σ slightly degrades ACC and increases ASR. Moderate values achieve a clear reduction in ASR with negligible impact on ACC, indicating a good balance between stability and sensitivity. Table III shows that smaller α values suppress backdoor behaviors more aggressively but slightly reduce ACC, whereas larger values preserve more client-specific information at the cost of higher ASR. FedFdpm maintains stable performance within a moderate range of α, demonstrating that the method is not overly sensitive to hyperparameter choices.

IV. CONCLUSION

In this paper, we propose FedFdpm, which achieves effective identification and repair of backdoor clients through spectrum detection, spectrum repair, and dynamic aggregation, thereby significantly enhancing model robustness. Experimental results demonstrate that FedFdpm consistently outperforms existing methods under varying attack intensities and heterogeneous scenarios, while maintaining low attack success rates and high accuracy even under extreme conditions. Future work will explore extending high-dimensional spectral feature discrimination capabilities to support the construction of trustworthy federated intelligent systems.

REFERENCES

[1] Nouhaila Innan, Muhammad Al-Zafar Khan, Alberto Marchisio, Muhammad Shafique, and Mohamed Bennai, "Fedqnn: Federated learning using quantum neural networks," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–9.

[2] Amin Aminifar, Jonathan Dan, and David Atienza, "Federated learning with patient-annotated data in epileptic seizure detection," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–10.

[3] Jiaming Pei, Wenxuan Liu, Jinhai Li, Lukun Wang, and Chao Liu, "A review of federated learning methods in heterogeneous scenarios," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5983–5999, 2024.

[4] Yuncan Tang and Yongquan Liang, "Credit card fraud detection based on federated graph learning," *Expert Systems with Applications*, vol. 256, pp. 124979, 2024.

[5] Rongqing Zhang, Jingxin Mao, Hanqiu Wang, Bing Li, Xiang Cheng, and Liuqing Yang, "A survey on federated learning in intelligent transportation systems," *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proceedings of the 35th annual computer security applications conference*, 2019, pp. 113–125.

[7] Yuncan Tang, Lina Ni, Jufeng Li, Jinquan Zhang, and Yongquan Liang, "Federated learning based on dynamic hierarchical game incentives in industrial internet of things," *Advanced Engineering Informatics*, vol. 65, pp. 103214, 2025.

[8] Yinbin Miao, Rongpeng Xie, Xinghua Li, Zhiquan Liu, Kim-Kwang Raymond Choo, and Robert H Deng, "Efficient and secure federated learning against backdoor attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4619–4636, 2024.

[9] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang, "Vertical federated learning: Concepts, advances, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3615–3634, 2024.

[10] Yongkang Wang, Di-Hua Zhai, and Yuanqing Xia, "Rope: Defending against backdoor attacks in federated learning systems," *Knowledge-Based Systems*, vol. 293, pp. 111660, 2024.

[11] Francesco Colosimo and Floriano De Rango, "Median-krum: A joint distance-statistical based byzantine-robust algorithm in federated learning," in *Proceedings of the Int'l ACM Symposium on Mobility Management and Wireless Access*, 2023, pp. 61–68.

[12] Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.

[13] Haiyan Zhang, Xinghua Li, Mengfan Xu, Ximeng Liu, Tong Wu, Jian Weng, and Robert H. Deng, "Badfl: Backdoor attack defense in federated learning from local model perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5661–5674, 2024.

[14] Hossein Fereidooni, Alessandro Pegoraro, Phillip Rieger, Alexandra Dmitrienko, and Ahmad-Reza Sadeghi, "Freqfed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning," *arXiv preprint arXiv:2312.04432*, 2023.

[15] Yuncan Tang, Yongquan Liang, Yang Liu, Jinquan Zhang, Lina Ni, and Liang Qi, "Reliable federated learning based on dual-reputation reverse auction mechanism in internet of things," *Future Generation Computer Systems*, vol. 156, pp. 269–284, 2024.

[16] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li, "Dba: Distributed backdoor attacks against federated learning," in *International Conference on Learning Representations*, 2020.

[17] Yuefeng Lai, Lizhao Wu, Hui Lin, and Jianmin Liu, "Bad-mfl: A cross-modality bi-trigger backdoor attack against multimodal federated learning," *IEEE Internet of Things Journal*, vol. 12, no. 19, pp. 39326–39337, 2025.

[18] Mohamed Amine Ferrag, Othmane Friha, Djallel Hamouda, Leandros Maglaras, and Helge Janicke, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022.

[19] Mustafa Safa Ozdayi, Murat Kantarcioglu, and Yulia R Gel, "Defending against backdoors in federated learning with robust learning rate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 9268–9276.

[20] Siquan Huang, Yijiang Li, Chong Chen, Ying Gao, and Xiping Hu, "Fedid: Enhancing federated learning security through dynamic identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2025.