IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 March 2025, accepted 13 June 2025, date of publication 16 June 2025, date of current version 23 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3580198

RESEARCH ARTICLE

Norm-Based Outlier Filtering and Consensus Aggregation for Robust Federated Learning

HASUNG YEO AND JOON-WOO LEE
Department of Computer Science and Engineering, Chung-Ang University, Seoul 06974, Republic of Korea

Corresponding author: Joon-Woo Lee (jwlee2815@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government [Ministry of Science and Information and Communications Technology (MSIT)] (Development of Privacy-Preserving Multiparty Computation Techniques for Secure Multiparty Data Integration) under Grant RS-2024-00399491; and in part by Chung-Ang University Research Scholarship Grant, in 2025.

ABSTRACT Federated Learning (FL) enables collaborative training of machine learning models across distributed clients without sharing private data, making it ideal for privacy-sensitive applications. However, FL is highly vulnerable to backdoor attacks, especially those designed to persist undetected over time. One of the most advanced of these is the Irreversible Backdoor Attack (IBA), which remains effective even after malicious clients are removed by aligning its updates with benign ones. In this paper, we propose a novel defense framework called Norm-Based Outlier Filtering and Consensus-Aware Aggregation (NOFCA) to effectively defend against such attacks. NOFCA combines norm-based outlier removal of client updates with a consensus-direction-based perturbation mechanism that disrupts the insertion of aligned adversarial updates. Unlike conventional aggregation rules or heuristic-based filtering, our method actively removes anomalous updates and performs adjusted aggregation based on the consensus direction of benign clients, thereby effectively obstructing backdoor attacks. Experimental results on MNIST and CIFAR-10 datasets demonstrate that NOFCA consistently achieves the lowest backdoor accuracy across all evaluated attack scenarios while maintaining competitive main accuracy. Furthermore, under fixed-frequency attack settings, where the attacker participates in every communication round, NOFCA shows robust and stable performance, maintaining strong defense even in highly persistent threat conditions. These findings confirm the effectiveness of NOFCA as a practical and resilient defense strategy for real-world FL systems.

INDEX TERMS Federated learning, machine learning, collaborative AI, privacy, security.

I. INTRODUCTION

Federated Learning (FL) is a decentralized machine learning paradigm that enables multiple devices or participants to collaboratively train a global model without directly sharing their private data. This approach significantly enhances privacy and data security, making it widely applicable in real-world scenarios across various domains. FL has become increasingly important in privacy-sensitive applications such as healthcare, finance, mobile services, and IoT-based applications [31], [32], [33], [34]. In healthcare, FL enables the collaborative training of disease diagnosis models without sharing sensitive patient data [31]. In finance, it allows secure fraud detection models to be developed while preserving transaction privacy [34]. Similarly, in mobile applications, FL supports personalized features such as predictive text or voice recognition while maintaining user privacy [33]. Beyond these domains, FL has also been applied in smart manufacturing, autonomous driving, and edge computing environments [32]. FL enables AI models to be trained efficiently across distributed edge devices, reducing communication overhead and improving scalability. The adoption of FL has been further accelerated by the increasing availability of distributed computing power and the demand for privacy-preserving AI solutions.

Traditional centralized learning approaches require aggregating data in a single location, which raises significant privacy concerns and computational burdens. In contrast, FL offers a highly scalable and privacy-enhancing alternative by allowing participants to train models locally and only

The associate editor coordinating the review of this manuscript and approving it for publication was Fabrizio Messina.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

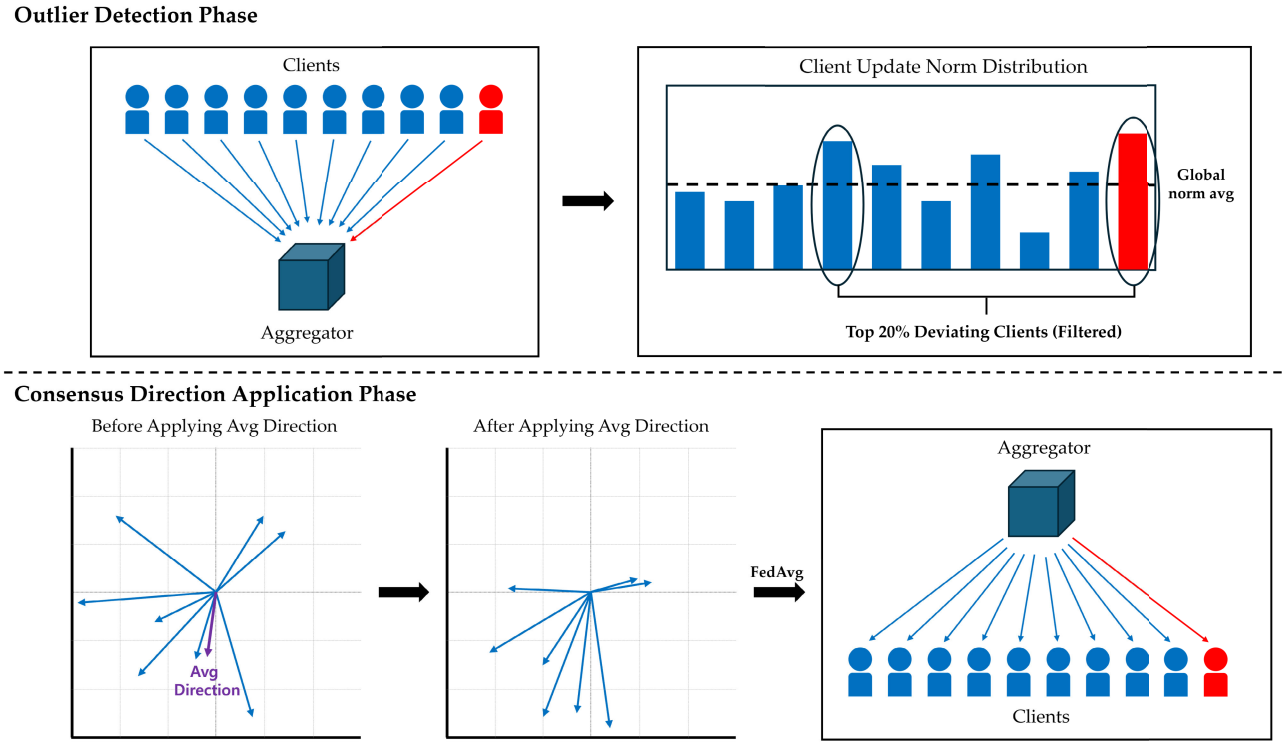


FIGURE 1. Illustration of the proposed defense framework (NOFCA). In the Outlier Detection Phase (top), clients with excessively deviating update norms are identified and filtered out to prevent extreme adversarial influence. In the Consensus Direction Application Phase (bottom), the remaining client updates are aligned with the benign consensus direction before being aggregated via FedAvg, effectively disrupting coordinated backdoor attacks.

share processed model updates instead of raw data. This decentralized nature, however, introduces significant new challenges, particularly in terms of security, trust, and overall robustness. One of the most pressing security concerns in FL is the threat of backdoor attacks. Backdoor attacks exploit the decentralized structure by embedding malicious triggers in local models, which, when aggregated into the global model, cause it to behave unexpectedly under specific conditions. These attacks are particularly challenging to detect due to their inherent stealthiness and the lack of centralized control over local updates.

Furthermore, backdoor attacks pose a serious and persistent long-term security risk to FL-based systems by enabling adversaries to gain covert and unauthorized control over the global model without disrupting its standard performance. Since FL is widely adopted in various critical applications such as healthcare [31], financial systems [34], and autonomous driving [35], a successful backdoor attack could potentially have catastrophic and far-reaching consequences. For instance, an adversarially controlled medical diagnosis model could misclassify critical conditions while appearing to function completely normally, or an autonomous vehicle system could be maliciously manipulated to make dangerous and unpredictable driving decisions under specific conditions. These escalating risks highlight the silent yet devastating impact of backdoor attacks, ultimately making them one of the most significant and pressing threats to FL security.

Among the most concerning advancements is the Irreversible Backdoor Attack (IBA) [4], which introduces a highly practical and effective backdoor technique. Unlike conventional backdoor attacks that require continuous adversarial participation, IBA remains effective even after the removal of malicious clients. This is achieved by ensuring that adversarial updates are aligned with benign updates, making them difficult to detect and neutralize. Moreover, IBA leverages the aggregation mechanism of FL to gradually embed imperceptible backdoor triggers into the global model, allowing the attack to persist across multiple training rounds without being detected [4]. Just as FL enhances model scalability and efficiency by leveraging decentralized computation, IBA maximizes its stealth and resilience by exploiting the inherent properties of FL aggregation, making it one of the most difficult backdoor attacks to mitigate.

Various defense mechanisms have been proposed to mitigate backdoor attacks in FL. For example, Krum [12] and Bulyan-krum [13] aim to filter outliers among client contributions to reduce the impact of malicious updates. Other defenses, such as Robust Federated Aggregation (RFA) [30], Foolsgold [22], and Robust Learning Rate (RLR) [29], attempt to improve aggregation robustness by mitigating adversarial influence in FL training. However,

these defenses suffer from high computational costs, privacy concerns, or ineffective handling of sophisticated attacks like IBA, where adversarial updates mimic benign behavior. In particular, due to the alignment of malicious and benign updates in IBA, existing defenses that depend on fixed aggregation rules or simplistic anomaly detection methods often struggle to effectively detect such attacks.

Among recent approaches, RoFL [38] introduces a norm-bounding strategy where client updates are strictly constrained by L_2 or L_∞ norms, and these constraints are enforced even under secure aggregation using zero-knowledge proofs. While this method offers strong theoretical guarantees, it also introduces significant computational overhead and depends on cryptographic components, which may hinder its applicability in practical FL deployments. In contrast, our method adopts a lightweight and adaptive defense mechanism that dynamically filters out norm-based outliers based on the distribution of client update norms, without relying on explicit constraints or cryptographic assumptions. This design makes it more scalable and suitable for real-world FL systems.

To address the limitations of existing defense mechanisms, we propose a novel defense framework called Norm-Based Outlier Filtering and Consensus-Aware Aggregation (NOFCA). This approach replaces traditional defenses and is specifically designed to mitigate sophisticated attacks such as IBA, Projected Gradient Descent (PGD), and Model Replacement (MR) while maintaining the performance of the global model. The proposed defense framework consists of the following two key components:

1) NORM-BASED OUTLIER FILTERING

As illustrated in Figure 1, the Outlier Detection Phase filters out the top 20% of deviating clients based on their update norms to prevent extreme malicious updates from being incorporated into the global model. This threshold was chosen based on hyperparameter tuning, where we observed similar robustness trends for 10%, 20%, and 30%, with 20% offering the best overall balance between robustness and accuracy. A more extensive sensitivity analysis of this parameter is left as future work. By analyzing the client update norm distribution, the global model ensures that only benign updates are aggregated, mitigating adversarial manipulations.

2) CONSENSUS-AWARE AGGREGATION

The Consensus Direction Application Phase in Fig. Figure 1 demonstrates how client updates are adjusted based on the average direction of benign updates. This step aligns updates toward a common consensus direction, effectively mitigating adversarial strategies where attackers attempt to disguise malicious updates by mimicking normal client behavior.

By integrating these two mechanisms, our proposed framework effectively neutralizes advanced backdoor attacks while preserving the stability and accuracy of the FL system.

II. RELATED WORKS

FL has emerged as a critical paradigm for distributed machine learning. However, it remains highly vulnerable to adversarial attacks, including backdoor attacks, model poisoning, and data poisoning. In this section, we provide a comprehensive review of existing research in FL security, covering different types of backdoor attacks, countermeasures, and various optimization strategies for enhanced robustness.

A. BACKDOOR ATTACKS IN FEDERATED LEARNING

Backdoor attacks exploit the decentralized nature of FL to introduce hidden triggers into global models, often without immediate detection. The seminal work by Bagdasaryan et al. [1] demonstrated how model poisoning could be used to implant a persistent backdoor in FL settings. Similarly, Trojan attacks [2] and BadNets [3] introduced techniques to manipulate training data and model parameters to produce targeted misclassification. More recent backdoor strategies such as IBA [4] show that malicious updates can persist in FL models even after adversarial clients are removed. LIRA [5] and PerDoor [6] further extend this by introducing learnable and stealthy triggers that adapt to FL's non-IID data distribution. Distributed Backdoor Attacks (DBA) [7] exploit multiple colluding clients to implant backdoors more effectively than single-client approaches. Other methods, including WANET [9], Hidden Trigger Backdoors [10], and sample-specific attacks [11], refine trigger embedding techniques to evade detection-based defenses. Wang et al. [36] established a theoretical connection between backdoor robustness and adversarial robustness, arguing that detecting backdoors in FL under realistic computational assumptions is fundamentally hard. They proposed edge-case backdoors, which target inputs located on the tail of the data distribution and demonstrated their effectiveness across various tasks. Sun et al. [28] showed that many existing Byzantine-robust FL algorithms are more vulnerable to poisoning than previously thought. They proposed the Divide-and-Conquer (DnC) defense, which outperformed previous approaches and demonstrated that even simple mechanisms such as norm clipping and weak differential privacy can help mitigate such attacks in real-world non-IID settings.

B. DEFENSES AGAINST BACKDOOR ATTACKS

A variety of defenses have been proposed to mitigate backdoor threats in FL. Byzantine-robust aggregation methods such as Krum, Multi-Krum [12], and Bulyan [13] aim to remove adversarial client updates by selecting only those that are most representative of the majority. However, sophisticated attacks like IBA [4] can effectively bypass these defenses by aligning malicious updates with benign ones, making detection challenging. To counter this, alternative aggregation strategies have been explored. Robust Federated Aggregation (RFA) [30] replaces mean-based aggregation with the geometric median, which limits the influence of

TABLE 1. Summary of related works.

Category	Ref. No	Author(s)	Topic
Backdoor Attacks	[1] [2] [3]	Bagdasaryan et al. Liu et al. Gu et al.	Model Poisoning
	[7] [8] [9] [10]	Xie et al. Nguyen et al. Saha et al. Li et al.	Trigger-Based Attacks
	[4] [5] [6]	Nguyen et al. Doan et al. Alam et al.	Irreversible or Persistent Attacks
	[28] [36]	Sun et al. Wang et al.	Robustness-Breaking and Edge-Case Attacks
Defense Mechanisms	[11] [12] [13] [30]	Blanchard et al. Guerraoui et al. Shejwalkar et al. Pillutla et al.	Aggregation-Level Defenses
	[16] [17] [18]	Wang et al. Liu et al. Tran et al.	Detection-Based Defenses
	[19] [20] [21]	Wu et al. Huang et al. Wang et al.	Perturbation-Based Defenses
	[14] [15]	Nguyen et al. Cao et al.	Trust or Reputation-Based Defenses
	[29]	Ozdayi et al.	Learning Rate Adaptation
	[37]	Lycklama et al.	Norm-Based Constraints and Secure Aggregation
System Enhancements	[22] [23] [33]	Fung et al. McMahan et al. Nishio et al.	Client Selection and Sybil Mitigation
	[24] [25] [26]	Li et al. Rothchild et al. Vinyals et al.	Optimization and Personalization
	[27] [31] [32] [34] [35]	Wu et al. Wang et al. Wu et al. Long et al. Zeng et al.	Real-World Deployment and Applications

outliers. Similarly, Robust Learning Rate (RLR) [29] dynamically adjusts the global learning rate based on the variance in client updates, reducing the impact of adversarial modifications over multiple rounds.Beyond aggregation-based approaches, trust-based mechanisms such as FLAME [15] and FLTrust [16] verify client model updates against a trusted reference dataset to identify abnormal updates. Neural Cleanse [17] attempts to reconstruct potential backdoor triggers by analyzing neuron activations, while Fine-Pruning [18] and Spectral Signatures [19] target adversarial neurons through frequency-domain analysis. However, these defenses can be computationally expensive, making them less scalable in large FL systems. Shejwalkar and Houmansadr [14] proposed a generalized poisoning framework that substantially outperforms prior model poisoning attacks, revealing the need for robust defenses under worst-case scenarios. Recent noise-based defense methods include Federated Pruning [20], Gradient Perturbation [8], and Layerwise Adaptive Noise Injection [21], which introduce controlled noise into client updates to prevent adversarial coordination. These techniques disrupt backdoor persistence while preserving model utility. FoolsGold [22] dynamically adjusts learning rates based on update similarity, mitigating Sybil-based poisoning attacks. Another category of defenses focuses on constraining the norm of client updates to limit adversarial influence. For example, RoFL [38] enforces a strict L_2 or L_∞ bound on each client's update and verifies compliance through zero-knowledge proofs within a secure aggregation protocol. This approach enables provable robustness even in encrypted FL settings. However, the requirement for cryptographic operations and fixed norm constraints may lead to increased computational overhead and limited scalability in practice.

In contrast, our proposed method adopts a distribution-aware approach by dynamically filtering out norm-based outliers without imposing explicit constraints or relying on cryptographic verification. This design achieves

robustness against stealthy attacks such as IBA [4] while maintaining practical efficiency in real-world FL systems.

C. OPTIMIZING SECURE FEDERATED LEARNING

Beyond backdoor defenses, optimizing FL for robustness is a key area of research. Federated Averaging (FedAvg) [23] remains the most widely used aggregation method but struggles in non-IID scenarios where client data distributions differ significantly. FedProx [24] introduces a proximal term that stabilizes client updates, reducing divergence in heterogeneous environments. Communication efficiency is another major concern in FL. Compression and quantization techniques [25] effectively reduce the bandwidth required for model updates, making FL more scalable and efficient. Nishio and Yonetani [33] addressed the heterogeneity of client resources by proposing FedCS, which actively selects clients based on their computation and communication capacities to speed up FL convergence. Other research, such as Matching Networks [26] and Federated Unlearning [27], focuses on improving adaptability, robustness, and security in adversarial FL environments.

In summary, while existing defense mechanisms such as norm bounding, aggregation rule tuning, and trust-based verification have made progress in mitigating common backdoor threats, they often fall short when facing stealthy and persistent attacks like IBA. In contrast, our proposed NOFCA framework is uniquely positioned to address these challenges by dynamically filtering outliers and realigning updates based on consensus direction, which are capabilities not found in prior methods. This highlights the necessity and novelty of NOFCA, and suggests a promising research direction that emphasizes both robustness and adaptability in practical FL deployments.

III. PRELIMINARIES

FL enables decentralized model training while preserving data privacy. However, its distributed nature introduces vulnerabilities to adversarial manipulation, particularly backdoor attacks. In such attacks, adversaries embed malicious triggers into local model updates, which, when aggregated into the global model, cause misclassification for specific inputs while maintaining high accuracy on normal samples. Over time, various backdoor attack methods have evolved, ranging from simple static triggers to advanced sample-specific and dynamic backdoor techniques.

Among these, IBA [4] has emerged as a highly effective and stealthy attack that remains persistent even after malicious clients are removed. Unlike traditional backdoor attacks that often rely on fixed trigger patterns, IBA employs imperceptible modifications to local model updates, making it difficult to detect and mitigate. The core mechanism behind IBA involves embedding malicious perturbations within the local training process by modifying inputs using the transformation function:

$$T_\xi(x) = x + G_\xi(x), \quad ||G_\xi(x)||_\infty \leq \epsilon, \forall x \tag{1}$$

where $G_\xi(x)$ represents a carefully crafted perturbation designed to remain undetectable while preserving the overall effectiveness and persistence of the backdoor. The optimization process behind IBA aims to maximize both clean and poisoned loss functions:

$$\min_w \sum_{x \in D} \alpha L_{\text{clean}}(f_{w_k}(x), y_x) + \beta L_{\text{poison}}(f_{w_k}(T_\xi(x)), y_T) \tag{2}$$

where L_{clean} and L_{poison} correspond to the loss functions for clean and poisoned samples, respectively, with α and β controlling their relative importance.

To further enhance stealthiness and durability, IBA incorporates two critical techniques: PGD and MR.

1) PROJECTED GRADIENT DESCENT

PGD is utilized in IBA to perform partial model poisoning. Specifically, it constrains poisoned local updates to remain close to the global model by projecting the gradient onto a norm-bounded ball centered around the previous global model. This helps IBA evade norm-based defense mechanisms and ensures the poisoned updates are not distinguishable from benign updates.

2) MODEL REPLACEMENT

IBA can also be combined with MR [1]. In MR, the adversary amplifies the local update before submitting it to the server to cancel out the updates from benign clients. This is achieved by scaling the poisoned model w_k such that the poisoned update can override the average contribution of benign clients during aggregation:

$$\hat{w}_k \approx w_k + \sum_{i \in C \setminus k} \frac{n_i}{n_{C \setminus k}} (w_k - w^*) \tag{3}$$

where w^* is the previous global model and n_i represents the number of samples held by client i. This enables the backdoor to be implanted effectively even if the attacker controls only a small fraction of clients.

By integrating PGD and MR into IBA, the attack becomes highly robust, achieving near-perfect backdoor success rates while bypassing state-of-the-art defenses across different FL settings.

IV. METHODOLOGY

FL enables decentralized training while preserving data privacy. However, backdoor attacks remain a major security threat. To effectively mitigate these attacks while maintaining model performance, we propose a novel defense mechanism called **NOFCA**, which integrates norm-based outlier filtering and consensus-aware aggregation. Unlike Byzantine-robust aggregation techniques, which primarily focus on filtering extreme updates, our method removes outlier clients and aligns the update directions of normal clients to a common consensus direction, effectively disrupting adversarial attempts.

Algorithm 1 Norm-Based Outlier Filtering

Notations:
ΔW_i: Local update from client i
n_i: Norm of client update
$\bar{n}$: Average norm
D_i: Norm deviation
q: Filtering percentile
S_{normal}: Clients not in top q% deviations
Input: Client updates $\{\Delta W_i\}_{i=1}^{N}$, q
Output: Filtered client set S_{normal}
foreach *client i* **do**
  $n_i \leftarrow \|\Delta W_i\|$
$\bar{n} \leftarrow \frac{1}{N}\sum_i n_i$;
foreach *client i* **do**
  $D_i \leftarrow |n_i - \bar{n}|$
Identify top q% clients with highest D_i;
$S_{\text{normal}} \leftarrow$ clients not in top q% deviations;

Algorithm 2 Consensus-Aware Aggregation

Notations:
S_{normal}: Filtered (non-outlier) client set
ΔW_i: Local update from client $i \in S_{\text{normal}}$
ΔW_{cons}: Average update of filtered clients
d_{cons}: Normalized consensus direction
c: Clipping threshold
W_{global}: Global model update
Input: Filtered updates $\{\Delta W_i\}_{i\in S_{\text{normal}}}$, c
Output: Global model update W_{global}
$\Delta W_{\text{cons}} \leftarrow \frac{1}{|S_{\text{normal}}|}\sum_{i\in S_{\text{normal}}} \Delta W_i$;
$d_{\text{cons}} \leftarrow \frac{\Delta W_{\text{cons}}}{\|\Delta W_{\text{cons}}\|}$;
foreach *client* $i \in S_{normal}$ **do**
  $\Delta W_i \leftarrow \text{ClipNorm}(\Delta W_i, c) + d_{\text{cons}} \circ |\Delta W_i|$
$W_{\text{global}} \leftarrow \text{FedAvg}(\{\Delta W_i\})$;

A. OVERVIEW OF NOFCA DEFENSE PIPELINE

The NOFCA framework consists of two major components: (1) norm-based outlier filtering and (2) consensus-aware aggregation. In the first phase, client updates with the largest deviation from the collective norm distribution are removed. In the second phase, remaining updates are adjusted to align with a consensus direction derived from the benign majority. This dual-stage pipeline is designed to (i) prevent extreme adversarial updates from being aggregated and (ii) neutralize the effect of malicious updates that imitate benign behavior.

B. NORM-BASED OUTLIER FILTERING

We first calculate the norm of each client's model update and compare it to the global average norm. Clients with excessively large deviations are considered outliers and excluded from aggregation. This prevents extreme malicious updates from contaminating the global model. Let the model update of client i be represented as $\Delta \mathbf{w}_i$. The global model update norm is defined as:

$$\|\Delta \mathbf{w}_{\text{global}}\| = \frac{1}{N}\sum_{i=1}^{N} \|\Delta \mathbf{w}_i\| \tag{4}$$

where N denotes the total number of participating clients. Next, we compute the deviation of each client's update from the global norm:

$$D_i = \left| \|\Delta \mathbf{w}_i\| - \|\Delta \mathbf{w}_{\text{global}}\| \right| \tag{5}$$

Clients whose deviation values are in the highest q% are considered outliers and removed from the global model aggregation:

$$S_{\text{outlier}} = \{i \mid D_i \geq \tau_q\} \tag{6}$$

where τ_q represents the q-th percentile of all deviation values, with $q = 20$ in our implementation. These clients are excluded from the aggregation process to ensure robustness against extreme updates. The key design rationale behind using deviation-based filtering, rather than enforcing absolute norm bounds as in prior work like RoFL [38], is to accommodate natural heterogeneity in benign updates. In FL, clients may operate on highly non-IID data, resulting in honest updates with widely varying magnitudes. Applying fixed constraints may inadvertently penalize such benign clients. Instead, NOFCA identifies updates that statistically deviate from the group, preserving flexibility while still filtering suspicious behaviors. The filtering threshold q can be tuned based on the desired trade-off between robustness and inclusiveness. A higher q increases tolerance to potentially malicious clients but risks excluding more benign ones.

The pseudocode of the proposed norm-based outlier filtering is presented in Algorithm 1.

C. CONSENSUS-AWARE AGGREGATION

After filtering out outlier clients, we apply Consensus-Aware Aggregation to ensure that the global model update aligns with the majority of benign clients while preventing adversarial updates from being effectively inserted. First, we compute the average update of the non-outlier clients, which will serve as the Consensus Direction:

$$\Delta \mathbf{w}_{\text{normal}} = \frac{1}{|S^c_{\text{outlier}}|}\sum_{i\in S^c_{\text{outlier}}} \Delta \mathbf{w}_i \tag{7}$$

where S^c_{outlier} denotes the set of non-outlier clients. Then, we normalize this average update to define the consensus direction:

$$d_{\text{consensus}} = \frac{\Delta \mathbf{w}_{\text{normal}}}{\|\Delta \mathbf{w}_{\text{normal}}\|} \tag{8}$$

This direction ensures that the global model update follows the benign majority while disrupting adversarial consistency. Notably, stealthy attacks like IBA [4] often craft malicious updates that subtly mimic benign gradients. By enforcing alignment with a consensus direction, our method neutralizes such strategies by re-centering the aggregated gradient space.

TABLE 2. Robustness comparison of various defenses against IBA, IBA+PGD, and IBA+PGD+MR on MNIST and CIFAR-10 datasets. MA: Main Accuracy, BA: Backdoor Accuracy.

Dataset	Defense	IBA (Standalone)		IBA + PGD		IBA + PGD + MR	
		MA(↑)	BA(↓)	MA(↑)	BA(↓)	MA(↑)	BA(↓)
CIFAR-10	RFA	**86.08%**	89.69%	85.89%	87.99%	86.06%	88.97%
	Krum	78.96%	26.46%	79.63%	57.88%	79.88%	44.37%
	Multi-krum	85.65%	30.69%	85.88%	77.64%	85.95%	62.91%
	RLR	83.65%	32.35%	83.61%	32.27%	83.69%	39.40%
	FoolsGold	85.25%	68.82%	84.64%	65.42%	85.49%	76.78%
	No-defense	85.82%	89.55%	**86.22%**	89.04%	**86.35%**	92.97%
	Proposed*	84.88%	**24.03%**	84.33%	**27.00%**	85.70%	**24.30%**
MNIST	RFA	99.04%	99.82%	99.02%	99.77%	99.03%	99.86%
	Krum	97.64%	73.22%	97.66%	70.30%	97.39%	78.99%
	Multi-krum	98.98%	99.39%	99.07%	98.69%	99.03%	97.41%
	RLR	90.43%	20.62%	90.40%	19.27%	90.56%	20.90%
	FoolsGold	99.03%	96.64%	99.04%	96.99%	**99.04%**	99.63%
	No-defense	**99.10%**	99.71%	**99.12%**	99.78%	99.00%	99.97%
	Proposed*	98.76%	**11.57%**	98.76%	**11.77%**	98.85%	**11.35%**

To further hinder adversarial update alignment, we modify each client update using the consensus direction:

$$\Delta \mathbf{w}_i' = \text{ClipNorm}(\Delta \mathbf{w}_i, c) + d_{\text{consensus}} \circ |\Delta \mathbf{w}_i| \quad (9)$$

where c is the clipping threshold to suppress overly large updates, $d_{\text{consensus}}$ is the normalized direction of benign updates, and the element-wise multiplication $\circ|\Delta \mathbf{w}_i|$ ensures that noise is proportionally injected based on the magnitude of each coordinate. This term $d_{\text{consensus}} \circ |\Delta \mathbf{w}_i|$ allows the adjustment to reflect the original update's strength on a per-coordinate basis, thereby preserving the semantic importance of each dimension from the client's perspective. In other words, the perturbation magnitude follows the original shape of the client's update while enforcing alignment with the benign consensus direction. The use of clipping in conjunction with directional adjustment provides dual protection. Clipping ensures bounded influence per client, while direction-based noise disturbs coordination among colluding adversaries. The adjusted client updates are finally aggregated into the global model using FedAvg.

The pseudocode of the proposed consensus-aware aggregation is presented in Algorithm 2.

V. EXPERIMENTS

We organize this section as follows. We first describe the experimental setup, including datasets, model configurations, and attack strategies. Then, we present results under various backdoor attack scenarios and evaluate the effectiveness of the proposed defense.

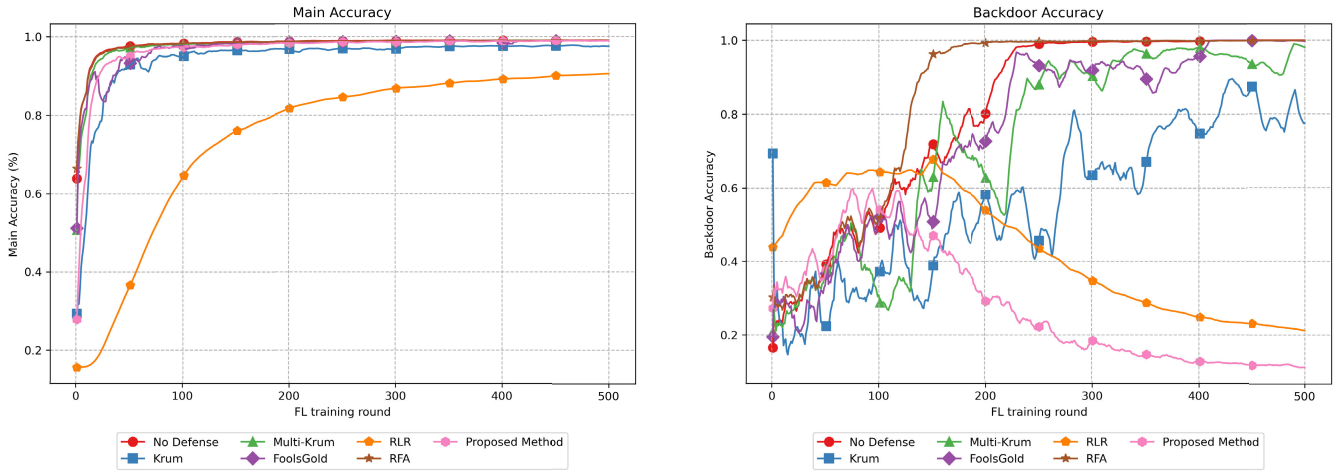
A. EXPERIMENTAL SETUP

We follow the established protocol outlined in previous works [4], [7]. Specifically, each client performs training using stochastic gradient descent (SGD) with $E = 2$ local epochs per communication round. The local learning rate is set to 0.01, and a decay factor of 0.998 is applied during training. The batch size is fixed at 32 for training and 256 for testing. The FL environment consists of 200 clients, with 10 clients randomly selected to participate in each round. The global model is updated using the standard FedAvg algorithm [23]. To simulate non-IID data distributions among clients, we partition the dataset using a Dirichlet distribution with concentration parameter $\alpha = 0.5$. The backdoor trigger generator is implemented using a U-Net architecture [37]. The perturbation strength ϵ is initially set to 0.3 and decayed gradually to a minimum value of $\hat{\epsilon} = 0.05$ over training. The trigger is generated every 10 communication rounds, and the learning rate for training the trigger generator is set to $\gamma_A = 0.0001$. We adopt the fixed-frequency attack scenario [36] to define the participating patterns of adversarial clients, where the attacker periodically joins every n-th communication round, representing a persistent and realistic threat model.

B. EXPERIMENTAL RESULTS

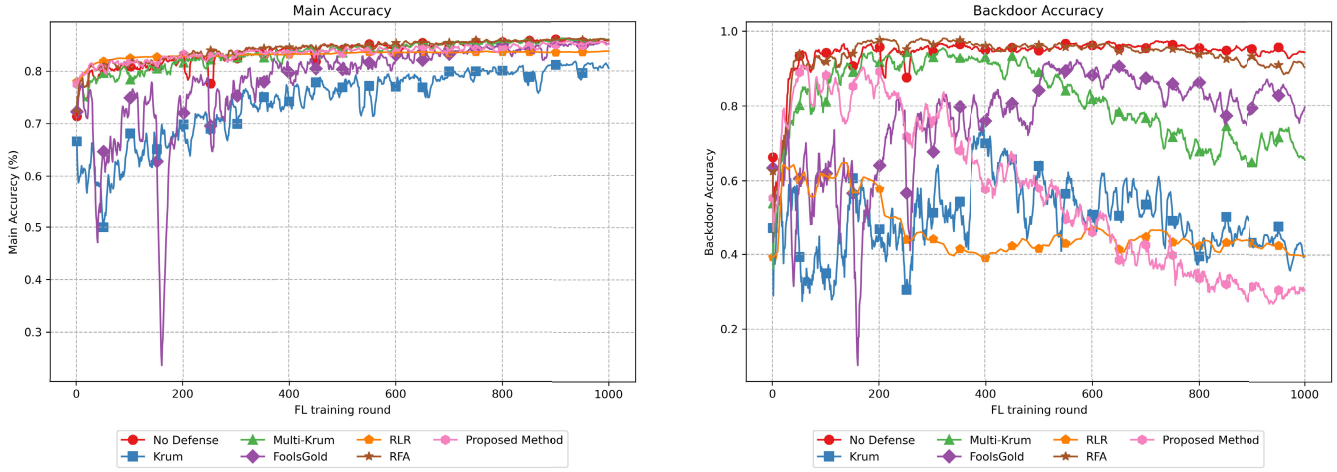
1) EVALUATION UNDER VARIOUS BACKDOOR ATTACK STRATEGIES

To assess the robustness of NOFCA under diverse threat models, we evaluate it against three backdoor attack scenarios



(a) Main Accuracy (b) Backdoor Accuracy

FIGURE 2. Evaluation of various defense mechanisms against the IBA + PGD + Model Replacement attack on the MNIST dataset. The attack is launched every 10 communication rounds (fixed-frequency) and lasts for 500 total FL rounds. The proposed method consistently achieves low backdoor accuracy while maintaining high main accuracy across the training process.



(a) Main Accuracy (b) Backdoor Accuracy

FIGURE 3. Evaluation of various defense mechanisms against the IBA + PGD + Model Replacement attack on the CIFAR-10 dataset. The attack is launched every 10 communication rounds (fixed-frequency) and lasts for 1000 total FL rounds.

of increasing sophistication: (1) IBA (Standalone), (2) IBA + PGD, and (3) IBA + PGD + Model Replacement. As shown in Table 2, we report both main accuracy (MA) and backdoor accuracy (BA) on CIFAR-10 and MNIST datasets. In the standalone IBA attack, the proposed defense achieves the lowest BA on both CIFAR-10 (24.03%) and MNIST (11.57%), while maintaining competitive MA of 84.88% and 98.76%, respectively. Under the IBA + PGD scenario, NOFCA continues to outperform other methods, reducing BA to 27.00% on CIFAR-10 and 11.77% on MNIST. Even under the strongest composite attack (IBA + PGD + MR), the proposed defense maintains the lowest BA at 24.30% (CIFAR-10) and 11.35% (MNIST), while preserving MA comparable to other SOTA defenses.

This trend is attributed to NOFCA's ability to eliminate malicious outliers early through norm-based filtering and reinforce benign update directions via consensus alignment, making it resilient against both simple and composite backdoor strategies.

2) COMPARATIVE ANALYSIS AGAINST EXISTING SOTA DEFENSES

To rigorously assess the effectiveness of NOFCA, we conduct a comprehensive comparison against widely used defense mechanisms, including Krum, Multi-Krum [12], RFA [30], Foolsgold [22], and RLR [29]. As shown in Figure 2 and Figure 3, our method is evaluated under the IBA + PGD +

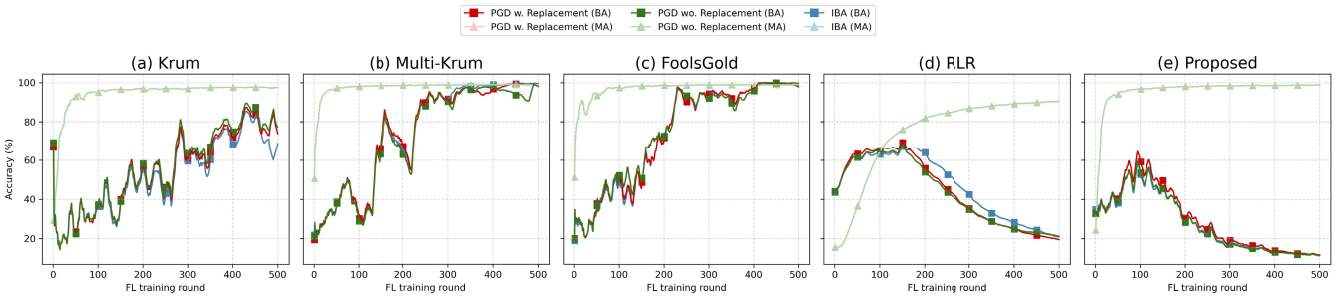


FIGURE 4. Comparison of different defense mechanisms under IBA + PGD + Model Replacement, IBA + PGD, and IBA standalone attacks on the MNIST dataset. Each subfigure shows the evolution of MA and BA over 500 FL rounds for (a) Krum, (b) Multi-Krum, (c) FoolsGold, (d) RLR, and (e) the proposed method.

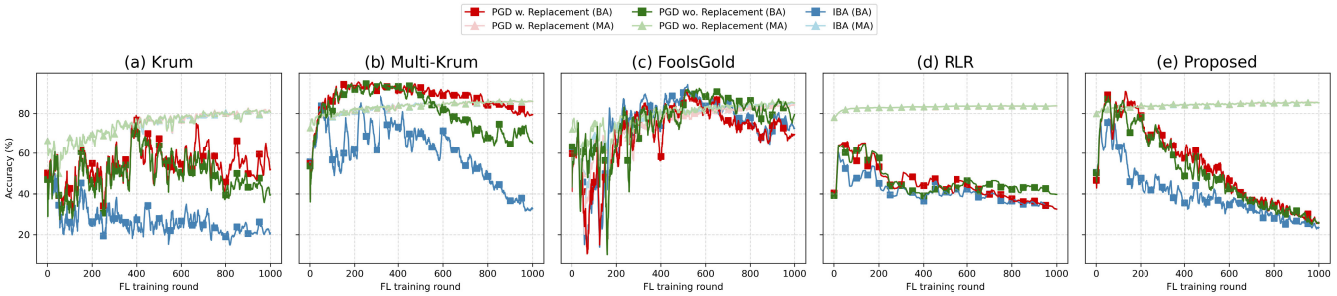


FIGURE 5. Comparison of different defense mechanisms under IBA + PGD + Model Replacement, IBA + PGD, and IBA standalone attacks on the CIFAR-10 dataset. Each subfigure shows the evolution of MA and BA over 1000 FL rounds for (a) Krum, (b) Multi-Krum, (c) FoolsGold, (d) RLR, and (e) the proposed method.

TABLE 3. Performance of the proposed defense method under varying attack frequencies on the MNIST dataset.

Freq	IBA		IBA + PGD		IBA + PGD + MR	
	MA(↑)	BA(↓)	MA(↑)	BA(↓)	MA(↑)	BA(↓)
1	0.9892	0.1112	0.9892	0.1101	0.9897	0.1093
2	0.9885	0.1046	0.9888	0.1054	0.9887	0.1049
3	0.9887	0.1068	0.9892	0.1107	0.9891	0.1066
4	0.9883	0.1152	0.9883	0.1082	0.9882	0.1094
5	0.9874	0.1152	0.9874	0.1154	0.9877	0.1123
6	0.9872	0.1149	0.9872	0.1151	0.9876	0.1201
7	0.9867	0.1181	0.9867	0.1142	0.9869	0.1192
8	0.9875	0.1139	0.9875	0.1152	0.9884	0.1132
9	0.9880	0.1094	0.9880	0.1071	0.9882	0.1102
10	0.9876	0.1157	0.9876	0.1177	0.9885	0.1135

Model Replacement setting across both MNIST and CIFAR-10 datasets. The results clearly demonstrate that our proposed method consistently outperforms the baselines by achieving a lower BA while maintaining competitive or superior MA. In particular, while methods like RLR are partially effective in mitigating backdoor effects, they exhibit slower convergence and slightly lower MA compared to NOFCA. Furthermore, our approach achieves a lower BA than RLR throughout the training process. This performance gap becomes more apparent as training progresses. After the initial rounds, our method shows stable and robust defense performance, while some baseline methods suffer from instability or partial backdoor persistence. Notably, our method exhibits a slightly elevated BA during the early training rounds. However, as training proceeds, it effectively and stably suppresses backdoor activation, eventually outperforming all other methods in terms of both stability and robustness. As illustrated in Figures 4 and 5, NOFCA achieves fast convergence without degrading the main task performance, and effectively suppresses backdoor activation across a variety of complex attack scenarios.

The superior performance of NOFCA can be explained by its dual mechanism: filtering out extreme deviations at the aggregation stage and adjusting gradients toward a benign consensus, which enhances convergence speed and suppresses backdoor signals over time.

3) PERFORMANCE CONSISTENCY UNDER FIXED-FREQUENCY ATTACKS

We further evaluate the consistency of the proposed NOFCA method under varying attack frequencies in the fixed-frequency attack scenario. In this setting, the adversarial client participates periodically in every n-th FL round (e.g., every 10 rounds), which is considered a strong and persistent threat model [36]. As shown in Table 3, the proposed method consistently achieves the lowest BA across different attack frequencies while maintaining high MA. Even as the attack frequency varies from 1 to 10, the MA of NOFCA remains

above 98.6%, and the BA stays stably suppressed with minimal fluctuation.

The consistent suppression of BA across all frequencies suggests that NOFCA is able to detect and nullify repeated malicious behaviors, preventing their accumulation in the global model over time.

VI. CONCLUSION

In this paper, we introduced **NOFCA**, a robust defense framework for FL that effectively mitigates sophisticated backdoor attacks while preserving model performance. Our method is specifically designed to defend against IBA and its advanced variants that incorporate PGD and MR, which are known to bypass many state-of-the-art defenses. NOFCA combines two complementary strategies. First, *Norm-Based Outlier Filtering* identifies and excludes clients whose updates significantly deviate from the global norm, preventing extreme adversarial influence. Second, *Consensus-Aware Aggregation* adjusts the remaining updates to align with the direction of benign client updates, thereby disrupting adversarial coordination without sacrificing utility. Comprehensive experiments on MNIST and CIFAR-10 datasets under various IBA attack scenarios, including IBA+PGD, and IBA+PGD+MR, demonstrate that NOFCA consistently achieves the lowest BA while maintaining high MA. Furthermore, under the fixed-frequency attack scenario, NOFCA maintains robust and stable defense performance across various attack intervals, including the most frequent case where the adversary participates in every single round. This demonstrates NOFCA's capability to consistently suppress backdoor threats even in persistent and high-frequency attack settings. These findings establish NOFCA as a highly effective and practical solution for securing FL systems against emerging backdoor threats, while preserving both robustness and utility across a wide range of adversarial conditions. We plan to further evaluate the computational overhead and deployment feasibility of NOFCA in practical FL environments such as mobile edge scenarios.

REFERENCES

[1] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. Int. Conf. Artif. Intell. Stat.*, Jan. 2018, pp. 2938–2948.

[2] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 304–318.

[3] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47230–47244, 2019.

[4] T. D. Nguyen, T. A. Nguyen, A. Tran, K. D. Doan, and K.-S. Wong, "IBA: Towards irreversible backdoor attacks in federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 66364–66376.

[5] K. Doan, Y. Lao, W. Zhao, and P. Li, "LIRA: Learnable, imperceptible and robust backdoor attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11946–11956.

[6] M. Alam, E. Sarkar, and M. Maniatakos, "PerDoor: Persistent non-uniform backdoors in federated learning using adversarial perturbations," 2022, *arXiv:2205.13523*.

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

[8] Z. Huang, Z. Wei, and J. Wang, "FL-AGN: A privacy-enhanced federated learning method based on adaptive Gaussian noise for resisting gradient inference attacks," *IEEE Access*, vol. 12, pp. 101366–101373, 2024.

[9] A. Nguyen and A. Tran, "WaNet-imperceptible warping-based backdoor attack," 2021, *arXiv:2102.10369*.

[10] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 11957–11965.

[11] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16443–16452.

[12] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Dec. 2017, pp. 118–128.

[13] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in byzantium," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jan. 2018, pp. 3521–3530.

[14] 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. 1278–1295.

[15] T. D. Nguyen, P. Rieger, R. De Viti, H. Chen, B. B. Brandenburg, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, F. Koushanfar, A.-R. Sadeghi, and T. Schneider, "FLAME: Taming backdoors in federated learning," in *Proc. 31st USENIX Secur. Symp. (USENIX Secur.)*, 2022, pp. 1415–1432.

[16] X. Cao, M. Fang, J. Liu, and N. Zhenqiang Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," 2020, *arXiv:2012.13995*.

[17] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723.

[18] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. Int. Symp. Res. Attacks, Intrusions, Defenses (RAID)*, Jan. 2018, pp. 273–294.

[19] B. Tran, J. Li, and A. Mądry, "Spectral signatures in backdoor attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2018, pp. 8000–8010.

[20] C. Wu, X. Yang, S. Zhu, and P. Mitra, "Mitigating backdoor attacks in federated learning," 2020, *arXiv:2011.01767*.

[21] S. Wang, Z. Fang, and K. Gai, "LRPAFL: Layer-wise relevance propagation-based adaptive federated learning," in *Proc. IEEE 11th Int. Conf. Cyber Secur. Cloud Comput. (CSCloud)*, Jun. 2024, pp. 174–179.

[22] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating Sybils in federated learning poisoning," 2018, *arXiv:1808.04866*.

[23] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Jan. 2016, pp. 1273–1282.

[24] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, Jan. 2018, pp. 429–450.

[25] D. Rothchild, A. Panda, E. Ullah, N. Ivkin, I. Stoica, V. Braverman, J. E. Gonzalez, and R. Arora, "FetchSGD: Communication-efficient federated learning with sketching," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jan. 2020, pp. 8253–8265.

[26] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Dec. 2016, pp. 3637–3645.

[27] C. Wu, S. Zhu, and P. Mitra, "Federated unlearning with knowledge distillation," 2022, *arXiv:2201.09441*.

[28] Z. Sun, P. Kairouz, A. Theertha Suresh, and H. Brendan McMahan, "Can you really backdoor federated learning?" 2019, *arXiv:1911.07963*.

[29] M. S. Özdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 10, pp. 9268–9276.

[30] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 1142–1154, 2022.

[31] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *Int. J. Med. Informat.*, vol. 112, pp. 59–67, Apr. 2018.
[32] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022.
[33] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.
[34] G. Long, Y. Tan, J. Jiang, and C. Zhang, "Federated learning for open banking," in *Federated Learning: Privacy and Incentive*, Q. Yang, L. Fan, and H. Yu, Eds., Cham, Switzerland: Springer, 2020, pp. 240–254.
[35] T. Zeng, O. Semiari, M. Chen, W. Saad, and M. Bennis, "Federated learning on the road autonomous controller design for connected and autonomous vehicles," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10407–10423, Dec. 2022.
[36] H. Wang, K. K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-Y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 16070–16084.
[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, Jan. 2015, pp. 234–241.
[38] H. Lycklama, L. Burkhalter, A. Viand, N. Küchler, and A. Hithnawi, "RoFL: Robustness of secure federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 453–476.



HASUNG YEO received the B.S. degree in computer engineering from the Lifelong Education Institute, in 2023. He is currently pursuing the M.S. degree in computer engineering with Chung-Ang University, Seoul, South Korea. His current research interests include robust federated learning and side-channel attack.



JOON-WOO LEE received the B.S. and Ph.D. degrees in electrical and computer engineering with Seoul National University, Seoul, South Korea, in 2016 and 2022, respectively. He is currently an Assistant Professor with the Engineering Department, School of Computer Science, Chung-Ang University, Seoul. His research interests include privacy-preserving machine learning, homomorphic encryption, and post-quantum cryptography.