Identify Backdoored Model in Federated Learning via Individual Unlearning

Jiahao Xu Zikai Zhang Rui Hu
University of Nevada, Reno
{jiahaox,zikaiz,ruihu}@unr.edu

Abstract

Backdoor attacks present a significant threat to the robustness of Federated Learning (FL) due to their stealth and effectiveness. They maintain both the main task of the FL system and the backdoor task simultaneously, causing malicious models to appear statistically similar to benign ones, which enables them to evade detection by existing defense methods. We find that malicious parameters in backdoored models are inactive on the main task, resulting in a significantly large empirical loss during the machine unlearning process on clean inputs. Inspired by this, we propose MASA, a method that utilizes individual unlearning on local models to identify malicious models in FL. To improve the performance of MASA in challenging non-independent and identically distributed (non-IID) settings, we design pre-unlearning model fusion that integrates local models with knowledge learned from other datasets to mitigate the divergence in their unlearning behaviors caused by the non-IID data distributions of clients. Additionally, we propose a new anomaly detection metric with minimal hyperparameters to filter out malicious models efficiently. Extensive experiments on IID and non-IID datasets across six different attacks validate the effectiveness of MASA. To the best of our knowledge, this is the first work to leverage machine unlearning to identify malicious models in FL. Code is available at https://github.com/JiiahaoXU/MASA.

1. Introduction

Federated Learning (FL) [31] is an emerging paradigm for training machine learning models across multiple distributed clients while preserving their data privacy. In FL, a central server coordinates a network of clients, each owning a local dataset. During the training process, the server distributes a shared global model to each client. The clients then train this model on their local datasets and send the resulting model updates back to the server. The server aggregates these updates to refine the global model for the next round of training. FL significantly reduces privacy risks by keeping the data on the client side throughout the process. FL has been successfully applied in various fields such as financial analysis [7, 29] and remote sensing [20, 28].

However, while the distributed nature of FL enhances data security, it also introduces vulnerabilities to poisoning attacks [14, 26, 42]. For example, Byzantine attacks [4, 40] aim to disrupt the global model's convergence. Specifically, malicious clients intentionally alter their local model updates to differ significantly from those of benign clients, thereby distorting the convergence process. Yet, this substantial deviation between malicious and benign updates offers an opportunity for server-side detection. Recently, backdoor attacks [3, 4, 13, 43, 46, 53], have gained significant attention due to their stealth and practical effectiveness. Specifically, backdoor attacks aim to preserve the global model's performance on clean inputs while causing it to make incorrect predictions on inputs containing a specific pre-defined feature (i.e., trigger). Since backdoor attacks have minimal impact on the main task's accuracy, the malicious local updates closely resemble benign ones [35, 43], making anomaly detection much more challenging.

One of the most common ways to defend against backdoor attacks in FL is to employ a *robust aggregation rule* (AGR) on the server side to handle the received local model updates [52]. Existing state-of-the-art (SOTA) AGRs can generally be classified into *non-filtering-based* AGRs [11, 19, 36, 37] and *filtering-based* AGRs [5, 8, 14, 18, 21, 38]. Non-filtering-based methods aim to mitigate the harmful effects of malicious parameters in the global model. However, they often fail to fully eliminate malicious impacts during aggregation and may also degrade main task performance. In contrast, filtering-based methods focus on identifying and excluding malicious local model updates to achieve maximum robustness. They typically rely on examining statistical differences (e.g., L_1-norm [18, 21], L_2-norm [5, 14, 18, 21], and Cosine Similarity [8, 38]) between malicious and benign updates. However, due to the dual optimization objectives of malicious clients, these statistical differences are often minimal, a phenomenon known as the poison-coupling effect [19]. Furthermore, as the global model approaches convergence, the statistical differences

between updates shrink further, reducing the effectiveness of filtering-based AGRs in detecting malicious updates.

Through a detailed observation of the poison-coupling effect in malicious local models, we find that backdoor parameters contribute negligibly when fed with clean inputs. This observation suggests that benign and malicious models can exhibit different behaviors during machine unlearning [6], a process aimed at removing learned information. Specifically, when unlearning the information associated with clean data, benign and malicious models show distinct behaviors in terms of convergence speed and unlearning loss, which can serve as effective metrics for anomaly detection. Motivated by this, we propose a novel AGR called **MASA**, which leverages **M**achine unle**A**rning on local model**S** individu**A**lly with pre-unlearning model fusion to identify malicious models. In MASA, the server first reconstructs the local models using the local model updates it received. Next, the server performs machine unlearning on each reconstructed local model and tracks its training losses during the unlearning to capture its unlearning behavior. Given that local models can exhibit high divergence in non-IID settings, which poses significant challenges for detecting backdoored models, MASA integrates a pre-unlearning model fusion process. This allows each local model to incorporate parameters learned from other local datasets before unlearning, reducing inconsistencies in unlearning behavior caused by non-IID data and effectively exposing backdoored models during the unlearning process. Finally, MASA filters out model updates with unusually large unlearning losses using a novel hyperparameter-efficient anomaly detection metric.

In summary, our main contribution is of four folds:

- We find that to preserve the performance of the main task, malicious parameters in backdoored models are less active than benign parameters when evaluated on clean inputs. Consequently, these less active parameters lead to significantly different unlearning behavior compared to benign models. This finding offers a new perspective for designing backdoor detection methods in FL.

- We design a new AGR called MASA, which leverages the distinct machine unlearning dynamics between backdoored and benign local models to identify backdoored models in FL. ***To the best of our knowledge, this is the first work to leverage machine unlearning for identifying backdoored models in FL***.

- MASA incorporates a pre-unlearning model fusion process, which significantly reduces the divergence in local models' unlearning behavior caused by non-IID data distributions among clients, helping to expose backdoored models in non-IID settings. Moreover, MASA is equipped with a hyperparameter-efficient anomaly detection metric to identify those local models with unusual unlearning loss.

- We conduct extensive empirical evaluations of MASA, testing its performance on IID, extreme non-IID, and extremely high attack ratio scenarios under various SOTA backdoor attacks. Results demonstrate that MASA consistently achieves superior backdoor robustness compared to SOTA defense methods.

2. Related Works

Existing backdoor attacks in FL. Empirical evidence has shown that FL is susceptible to backdoor attacks due to its lack of control over the local training data of clients [3]. The first backdoor attack, known as Badnet [13], adheres to the standard backdoor attack strategy by embedding the backdoor trigger in the local training data of malicious clients. Following this, numerous studies have been conducted to refine and strengthen backdoor attacks in FL [2, 3, 10, 13, 34, 43, 46, 51, 53], such as Scaling attack [3], Projected Gradient Descent (PGD) attack [43], Distributed Backdoor Attack (DBA) [46], Neurotoxin attack [53], Little is Enough (Lie) attack [4], etc. Recently, trigger-optimization backdoor attacks [2, 10, 30, 34, 51] (as known as adaptive attacks in some literature) have been studied to compromise the model with optimized triggers.

Defend against backdoor attacks in FL. Existing defense methods generally fall into two categories: filtering-based and non-filtering-based. Filtering-based methods aim to identify and exclude malicious local models from the aggregation process [5,8,14,18,21,23,38], while non-filtering-based methods seek to mitigate the impact of backdoor models on the global model [1, 11, 19, 36, 37, 39, 45, 47].

(1) Non-filtering-based methods: For instance, *Lockdown* [19] operates under the assumption that malicious parameters of backdoored models, which are used to recognize backdoor triggers, are considered unimportant by benign clients. Consequently, it applies sparsification to prune these unimportant parameters from all the local models. *FedSKU* [47] aims to recover the trigger on the server side first and then removes the knowledge of the identified triggers via machine unlearning while selectively transferring the useful knowledge into a surrogate clean model using distillation. Nevertheless, malicious parameters in backdoored models are often coupled with benign parameters, meaning they also contribute to the main task. Consequently, modifying or removing these parameters can lead to a significant loss in main task performance, rendering such methods ineffective.

(2) Filtering-based methods: On the other hand, filtering-based methods strive for the highest backdoor robustness by identifying and filtering out malicious local models/updates. For example, *Multi-Metrics* [18] computes

Manhattan distance, Euclidean distance, and Cosine Similarity for each local model update with the latest global model. Afterward, it projects the value of these metrics to its corresponding principal axis and calculates a score using the covariance matrix of these values. Those model updates with a low score will be dropped. *MESAS* [21] assesses various metrics for each local model update including L_1-norm, L_2-norm, variance, maximum value, minimum value, and the count of weights that have increased relative to the latest global model. Using these metrics, MESAS iteratively detects and eliminates malicious updates through statistical tests and clustering. However, due to the poison-coupling effect, many of these statistical metrics (e.g., L_1-norm, L_2-norm, Cosine Similarity) are often similar for both malicious and benign models, reducing the effectiveness of existing filtering methods.

In contrast to existing defense methods, our method, MASA enjoys the following advantages: **(I)** As a filtering-based approach, MASA aims to achieve the highest backdoor robustness compared to non-filtering-based methods. **(II)** Unlike existing methods that use machine unlearning to make the global model forget the trigger through a trigger reversal process, MASA eliminates the need to recover the backdoor trigger. **(III)** Instead of relying on statistical metrics from local models, MASA leverages the intrinsic nature of malicious parameters, which exhibit reduced activity when presented with clean input and conducts individual machine unlearning on each local model to accurately and robustly expose malicious local model updates. **(IV)** Existing defense methods struggle in non-IID settings, where local models vary due to diverse local datasets. In contrast, MASA addresses this challenge using a pre-unlearning model fusion. This approach allows local models to integrate global knowledge learned from data, thereby reducing divergence in their unlearning behaviors within non-IID environments. **(V)** MASA utilizes a hyperparameter-efficient anomaly detection metric, the median deviation score, to effectively identify local model updates with abnormal unlearning behaviors.

3. Key Motivation

3.1. Dual objectives of malicious clients

In a typical FL system, a set of n clients aim to collaboratively train a shared global model $\theta \in \mathbb{R}^d$ in an iterative manner under the coordination of a central server. Generally, the FL problem in a benign environment can be formulated as $\min_\theta (1/n) \sum_{i=1}^{n} F_i(\theta; D_i)$, where $F_i(\cdot)$ represents the local learning objective of client i. For example, for a benign client i performing a multi-class classification task, its local objective can be formulated as:

$$F_i(\theta; D_i) \coloneqq \mathbb{E}_{(z,y)\in D_i} \mathcal{L}(\theta; z, y), \tag{1}$$

where $\mathcal{L}(\cdot)$ is the cross-entropy loss function, and (z, y) is a datapoint sampled from benign dataset D_i. The classic method to solve the FL problem iteratively is known as FedAvg [32]. Specifically, at each training round t, client $i \in [n]$ downloads the latest global model θ^{t-1} from the server, refines the model towards optimizing its local objective to obtain an updated local model θ_i^t and then sends its local model updates $\Delta_i^t \coloneqq \theta_i^t - \theta^{t-1}$ back to the server. The server refines the global model by averaging the local updates as follows: $\theta^t \coloneqq \theta^{t-1} + (1/n) \sum_{i=1}^{n} \Delta_i^t$. This process repeats until the global model converges.

For FL under backdoor attacks, the adversary compromises or injects multiple malicious clients into the system. These malicious clients poison a portion of their local datasets to enable the injection of backdoor triggers into their local models during training. Specifically, the local dataset D_j of a malicious client j is divided into two subsets: $D_{j,M}$ and $D_{j,B}$. $D_{j,M}$ consists of benign data used for training on the main task with the same objective in Equation (1). $D_{j,B}$ contains the poisoned data, generated by stamping a trigger on each data sample and modifying the ground truth label to a target label specified by the adversary. With $D_{j,B}$, malicious clients can achieve that any input containing the trigger will be misclassified as the target label instead of its correct ground truth label. Formally, the local learning objective of a malicious client j can be formulated as follows.

$$F_j(\theta; D_j) \coloneqq \underbrace{\mathbb{E}_{(z,y)\in D_{j,M}} \mathcal{L}(\theta; z, y)}_{\text{main task loss}} + \underbrace{\mathbb{E}_{(\widetilde{z},\widetilde{y})\in D_{j,B}} \mathcal{L}(\theta; \widetilde{z}, \widetilde{y})}_{\text{backdoor task loss}},$$

where $(\widetilde{z}, \widetilde{y})$ is a datapoint sampled from the poisoned data $D_{j,B}$. These dual objectives allow malicious clients to effectively inject backdoors into their local models while preserving performance on the main task. Once the server aggregates the local model updates from all clients, these backdoors can be transferred to the global model.

3.2. Parameter coupling in backdoored models

Given that the malicious clients optimize for both the main task and the backdoor task simultaneously, their local model updates are statistically similar to those of benign ones (also known as the poison-coupling effect [19]), rendering existing filtering-based methods ineffective. More precisely, the parameters of a backdoored model θ^* can be decomposed as $\theta^* = \theta_M \cup \theta_B$, where θ_M represents the benign parameters related to the main task, and θ_B represents the backdoor parameters associated with the backdoor task. This decomposition holds because of the high-level independence between the main task and the backdoor task [24, 25]; by design, backdoor attacks should not impact the model's performance on clean inputs [13]. Additionally, recent observations suggest that in backdoored models, the backdoor parameters have a negligible impact

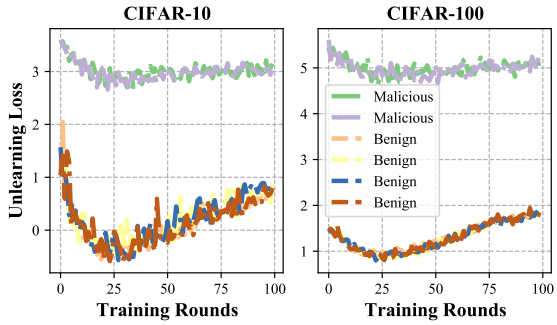

Figure 1. The unlearning loss of local models under Badnet attack on CIFAR-10 (left) and CIFAR-100 (right) w.r.t. training rounds.

during forward propagation with clean data, preserving the performance of the main task [44]. Note that there can be overlap between θ_M and θ_B, i.e., $\theta_M \cap \theta_B \neq \emptyset$, which means that some parameters may influence both tasks. Consequently, the coupling of parameters in malicious models poses significant challenges for mitigating/detecting backdoor attacks. For example, for non-filtering-based AGRs, if parameters critical to the main task also influence the backdoor task in the global model, removing or adjusting these parameters can significantly degrade the main task performance. For existing filtering-based AGRs, malicious models often display statistical patterns similar to benign ones, rendering traditional statistical metrics ineffective. Therefore, an effective method is required that not only identifies malicious models before aggregation but also does not rely on ineffective statistical metrics to distinguish malicious and benign models. In the following subsection, we demonstrate that machine unlearning is a desired solution that can effectively induce distinct unlearning behavior differences between malicious and benign models.

3.3. Exposing backdoored models via unlearning

Since the malicious parameters θ_B are less active for the main task, one approach to identify malicious and benign models is to perform *machine unlearning* [6] on all local models with respect to the main task (i.e., maximize the main task loss). Here, machine unlearning refers to a process that selectively removes specific learned information from the model. The key intuition is that machine unlearning can be regarded as a reverse process of machine learning, hence during unlearning, primarily the benign parameters θ_M are involved. As θ_B is less active in the main task, consequently, during unlearning, this reduced activity results in a *faster* and *more pronounced* unlearning compared to benign models. Here, the faster and more pronounced unlearning, resulting from the main engagement of θ_M in the unlearning leads to two unlearning dynamics: 1) *a larger empirical loss for backdoored models on the unlearning task*; and 2) *a quicker reversion of the backdoored model to random guessing*.

To verify this intuition, we study the loss of both malicious and benign local models on unlearning the main task, and present results in Figure 1. Specifically, we experiment with a simple FL system where 2 of a total of 6 clients are malicious on CIFAR-10 (with ResNet18 [16]) and CIFAR-100 (with VGG16 [41]) datasets under Badnet attack [13]. Upon receiving the local model updates, the server reconstructs each local model and performs individual unlearning for 5 epochs. We report the changes of averaged log-transformed unlearning loss with respect to training rounds for each client. We can see that during unlearning, malicious models exhibit a significantly larger empirical unlearning loss compared to benign ones for both datasets. Additionally, benign models share similar loss patterns, as do malicious models among themselves. The distinct difference in loss patterns between benign and malicious models provides room for malicious model identification. Furthermore, the consistent loss patterns within groups of benign clients and malicious clients respectively simplify the model identification process in practice. These observations motivate the design of a novel backdoor detection method that leverages the significant differences in unlearning dynamics between malicious and benign local models.

4. Our Solution: MASA

MASA is a server-side AGR that can be seamlessly integrated into existing FL systems. Specifically, upon receiving local model updates, MASA first reconstructs each local model and examines them using a well-designed individual unlearning process. Model updates exhibiting abnormal unlearning behavior are excluded from the aggregation, thereby achieving backdoor robustness of the FL system. The overall algorithm of MASA is given in Algorithm 1.

Individual unlearning on each local model. The core of MASA is the *individual unlearning* which unlearns each local model on the main task. To perform individual unlearning, the server must first reconstruct the local models after receiving the local model updates. Specifically, at round t, the server recovers the local model θ_i^t of client i by $\theta_i^t = \theta^{t-1} + \Delta_i^t$, where Δ_i^t is the local model update of client i at round t, and θ^{t-1} is the global model at previous round. With the recovered local model, the server then conducts unlearning on a proxy clean dataset D_p by solving the following minimization problem.

$$\min_{\theta_i^t} -\mathbb{E}_{(z,y)\in D_p}\mathcal{L}(\theta_i^t; z, y). \tag{2}$$

Note that the unlearning loss given in Equation (2) is with a negative sign "$-$" compared with the main task loss given in Equation (1). The server performs the unlearning process for all local models individually (line 10 in Algorithm 1) and accumulates the training loss generated during unlearning (line 8- 9) in order to capture the detailed unlearning

Algorithm 1 MASA

Require: Local updates $\{\Delta_i^t\}_{i=1}^n$ at round t, the latest global model θ^{t-1}, proxy dataset D_p, fusion degree λ, filter radius δ, unlearning epoch E, unlearning rate η_u.

1: Initialize selection set $\mathcal{S} \leftarrow \emptyset$
2: Initialize loss accumulator $\mathrm{A} \leftarrow \{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n\}$
3: $\bar{\Delta}^t \leftarrow (1/n) \sum_{i=1}^n \Delta_i^t$
4: **for** client $i \in [n]$ **do**
5: $\quad \mathrm{A}_i \leftarrow 0$
6: $\quad \widetilde{\theta}_i^t \leftarrow \theta^t + \lambda \Delta_i^t + (1-\lambda)\bar{\Delta}^t$ ◁ model fusion
7: $\quad$ **for** $e = 0$ to E **do**
8: $\quad\quad$ loss $\leftarrow \mathcal{L}(\widetilde{\theta}_i^t; x)$ with each mini-batch x sampled from D_p
9: $\quad\quad \mathrm{A}_i \leftarrow \mathrm{A}_i +$ loss ◁ loss accumulation
10: $\quad\quad \widetilde{\theta}_i^t \leftarrow \widetilde{\theta}_i^t - \left(-\eta_u \nabla \mathcal{L}(\widetilde{\theta}_i^t; x)\right)$ ◁ unlearning
11: $\quad$ **end for**
12: **end for**
13: **for** client $i \in [n]$ **do**
14: $\quad c_i \leftarrow \mathrm{calculate\_mds}(\mathrm{A}_i, \mathrm{A})$ ◁ by Equation (4)
15: $\quad$ **if** $c_i < \delta$ **then**
16: $\quad\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{i\}$
17: $\quad$ **end if**
18: **end for**
19: $\widetilde{\Delta} \leftarrow (1/|\mathcal{S}|) \sum_{i \in \mathcal{S}} \Delta_i^t$ ◁ benign update aggregation
20: **return** $\widetilde{\Delta}$

behavior of a model. Note that it is crucial for the proxy dataset to be carefully collected to ensure a certain degree of overlap with the training dataset. If this condition is not met, the unlearning process may be performed on a task unrelated to the main one, making it ineffective and potentially useless. In practice, the proxy dataset can be easily obtained by prompting cutting-edge generative models (e.g., generative adversarial networks [12] and diffusion models [17]).

Pre-unlearning model fusion. In practice, clients in FL often have non-IID data, where each client only possesses a subset of the classes from the overall dataset. This leads to significant divergence among local models during training. As a result, the unlearning behavior of each model also diverges, as each model can only unlearn the information learned from samples in its local dataset. This variation, caused by the non-IID data, allows backdoored models to blend in with benign models, making their detection more challenging. To this end, we propose a method called *pre-unlearning model fusion* to ensure the exposure of backdoored models through the unlearning process in non-IID settings. Specifically, before performing individual unlearning, the server constructs an aggregated model update $\bar{\Delta}$ by averaging all the received local updates (line 3). When reconstructing each local model, the server combines the local model update with the aggregated model update (line 6), allowing the reconstructed local model to incorporate global information. Formally, for client i, its local model reconstruction with pre-unlearning model fusion can be formulated as follows.

$$\widetilde{\theta}_i^t = \theta^{t-1} + \lambda \Delta_i^t + (1-\lambda)\bar{\Delta}, \tag{3}$$

where $\widetilde{\theta}_i^t$ is the reconstructed local model for client i, $\bar{\Delta} := (1/n) \sum_{i=1}^n \Delta_i^t$, n is the number of received local model updates, and $\lambda \in (0.5, 1]$ is a coefficient called *fusion degree*, which controls the fusion level. In an extreme case when $\lambda = 1.0$, pre-unlearning model fusion is disabled. The value of λ must be greater than 0.5 to ensure that the local model update predominates over the aggregated model update after fusion. With pre-unlearning model fusion, $\widetilde{\theta}_i^t$ incorporates local model updates from other clients, which contain information learned from their local datasets. This enhances the generalization ability of the reconstructed local models in non-IID settings, allowing them to perform unlearning on the main task more comprehensively and consistently, thereby reducing the divergence in unlearning behaviors across local models.

Efficient outlier filtering. With the accumulated training loss values obtained in individual unlearning, MASA filters out local model updates that exhibit significantly high loss values. One approach to achieve this is by using existing clustering methods (e.g., KMeans [15] and MeanShift [9]) to group malicious and benign models based on their unlearning loss. However, these clustering methods often require multiple hyperparameters, necessitating time-consuming hyperparameter tuning to achieve satisfactory performance. To perform outlier filtering efficiently and accurately, we propose a new metric called *Median Deviation Score* (MDS), based on the classical *Z-score*. Specifically, given a set X which contains n elements $\{x_1, x_2, \ldots, x_n\}$, the Z-score for $x_i \in X$ is calculated as $(x_i - \bar{X})/\sigma$ where $\bar{X}$ and σ are the mean and the standard deviation of X, respectively. As for the MDS c_i of x_i, it is calculated as

$$c_i = (x_i - \hat{X})/\sigma, \tag{4}$$

where $\hat{X}$ is the median element of X (line 14). Note that median is widely used in FL as a robust metric to defend against poisoning attacks [48–50]. By substituting the mean with the median in the Z-score calculation, MDS becomes a more robust metric for highlighting extremely large values in a given set. Besides, we observed that the accumulated unlearning loss of malicious models is *significantly* and *consistently* ***larger*** than that of benign ones. This allows us to utilize a single threshold parameter, called the *filter radius* δ, to distinguish between benign and malicious local models based on their MDS value. More precisely, local models with $c < \delta$ are considered benign and added to the selection set $\mathcal{S}$ (line 15-16). The server then aggregates

the local updates in $\mathcal{S}$ to construct the global model update $\widetilde{\Delta}$ (line 19), which will be used to refine the global model.

5. Experimental Settings

Datasets and models. In our experiments, we primarily use two benchmark datasets CIFAR-10 [22] and CIFAR-100 [22] and simulate both IID and non-IID data settings. For IID settings, we evenly split the dataset over the clients. We use *Dirichlet distribution* [33] $Dir(\alpha)$ to simulate the non-IID settings with a default non-IID degree $\alpha = 0.5$. We use ResNet18 [16] and VGG16 [41] as the model for CIFAR-10 and CIFAR-100, respectively.

Training settings. We present the detailed training settings in Appendix C. For all datasets, we simulate a cross-silo FL system with 20 clients. For the individual unlearning process of MASA, we use SGD with a learning rate of 0.001 and momentum of 0.9 for 5 epochs of training. To construct the proxy dataset on the server, we sample 1% of the training data from the original dataset. Note that in practice the proxy dataset can be generated by cutting-edge generative models, we demonstrate that MASA with a generated proxy dataset achieves very similar performance to MASA in Appendix E.

Evaluation metrics. We evaluate the performance of defense methods using three key metrics: *main task accuracy (MA)*, which reflects the percentage of clean test samples that are correctly classified to their ground-truth labels by the global model; *backdoor task accuracy (BA)*, which indicates the percentage of triggered samples that are misclassified to the target label by the global model; and *robustness accuracy (RA)*, which measures the percentage of triggered samples that, despite the presence of the trigger, are accurately classified to their ground-truth labels by the global model. An effective AGR against backdoor attacks should achieve high MA and RA while maintaining a low BA.

Evaluated attack methods. We mainly consider six different attacks and they can be categorized by the attacker's capability to manipulate model updates into three levels: *limited*, *intermediate*, and *advanced*. At the limited level, malicious clients are restricted to manipulating their own local datasets to generate malicious updates, which are then sent to the server for aggregation (i.e., Badnet [13] and DBA [46]). At the intermediate level, attackers extend their capability by modifying the training algorithm itself in addition to dataset manipulation, resulting in more sophisticated malicious updates (i.e., Scaling [3] and PGD [43]). These two levels of attacker capabilities are commonly assumed in existing literature, where attackers control malicious devices but lack access to additional information from servers or benign clients. At the advanced level, however, attackers can access and exploit global information from the server, such as aggregated model updates, to enhance their attacks (i.e., Neurotoxin [53] and Lie [4]). Note that in our work, the defense method employed by the server remains confidential to the attacker (i.e., the server is always trustable). To simulate effective backdoor attacks (achieving a BA over 60% [21]), the malicious client will poison $r = 50\%$ of its local data by default, where r represents the *data poisoning ratio*. For all the attacks, the *attack ratio* is set to $f/n = 20\%$ by default where f is the number of malicious clients, which means 20% of the clients in the system are malicious. Note that achieving robustness against model poisoning attacks using filtering-based methods is generally impossible when $f \geq n/2$ [27]. Similar to previous work on defending against backdoor attacks in FL [18, 35], in our work, the number of malicious clients, f, is assumed to be $f < n/2$. More detailed settings of attacks are given in Appendix A.

Evaluated defense methods. We compare MASA with the non-robust baseline *FedAvg* [32] and six existing SOTA defense methods, including *RLR* [36], *RFA* [37], *Multi-Krum (MKrum)* [5], *Foolsgold* [11], *Multi-Metric (MM)* [18], and *Lockdown* [19]. Additionally, we compare our approach to an ideal filtering-based robust aggregation, which perfectly identifies and removes all malicious updates while averaging the benign ones to update the global model. We refer to this as the most robust baseline, *FedAvg**. Note that FedAvg* achieves the highest level of robustness theoretically.

6. Experimental Results

Performance of MASA in IID settings. We first evaluate MASA on CIFAR-10 and CIFAR-100 datasets under three attack scenarios with $r = 0.3$ and $r = 0.5$, respectively, and present the results for clean MA (without attacks), BA, and RA in Table 1. Overall, MASA consistently achieves the lowest average BA and the highest average RA across both datasets, demonstrating superior backdoor robustness compared to its counterparts.

Specifically, on the CIFAR-10 dataset, RLR achieves an average BA of only 18.03% and an RA of 65.63%, which are 17.15% and 23.07% lower than those of MASA, respectively. This performance gap arises because RLR reverses the global learning direction when inconsistencies in local model parameter signs are detected. While this approach reduces some malicious influences, it degrades the global model's performance on clean inputs and fails to fully neutralize the backdoor attack. For CIFAR-100 dataset, RFA and MKrum show slightly higher BA (0.11% and 0.17% higher, respectively) but considerably lower RA (11.66% and 2.36% lower, respectively) than MASA. RFA selects the geometric median of local model updates as the global model update, effectively preventing the global model from being compromised. However, the geometric median is suboptimal as an optimization direction compared to aggregated local model updates, leading to a lower RA due to per-

Table 1. The clean MA, BA, and RA of different methods on IID CIFAR-10 and CIFAR-100 under three types of attacks.

Dataset (Model)	Method	Clean MA↑	Badnet				DBA				Scaling				Avg. BA↓	Avg. RA↑
			BA↓		RA↑		BA↓		RA↑		BA↓		RA↑			
			r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5		
CIFAR-10 (ResNet18)	FedAvg	91.43	99.98	99.99	0.02	0.01	99.93	99.99	0.07	0.01	99.98	99.99	0.02	0.01	99.98	0.02
	FedAvg*	91.43	0.56	0.56	89.03	89.03	0.56	0.56	89.03	89.03	0.56	0.56	89.03	89.03	0.56	89.03
	RLR	80.15	3.53	99.99	75.54	0.00	0.94	**0.72**	78.46	77.12	1.90	3.00	80.57	80.10	18.03	65.63
	RFA	89.73	0.99	0.90	87.21	87.43	1.10	0.99	87.36	87.78	1.04	1.06	87.34	86.17	1.01	87.22
	MKrum	90.71	1.01	1.12	88.08	88.08	1.13	1.14	87.64	87.57	**0.80**	1.01	87.81	87.99	1.03	87.86
	Foolsgold	91.10	99.97	99.99	0.03	0.01	99.92	99.98	0.07	0.02	99.99	99.99	0.01	0.01	99.97	0.03
	MM	89.59	99.99	99.99	0.01	0.01	99.99	99.99	0.01	0.00	99.98	99.99	0.02	0.01	99.99	0.01
	Lockdown	**91.25**	72.71	46.99	26.33	49.03	95.59	29.80	4.34	57.70	75.22	10.12	24.13	74.03	55.74	39.26
	MASA	91.24	**0.90**	**0.87**	**88.30**	**88.91**	**0.81**	0.87	**88.81**	**88.63**	0.96	**0.87**	**88.66**	**88.91**	**0.88**	**88.70**
CIFAR-100 (VGG16)	FedAvg	67.25	99.65	99.68	0.26	0.24	99.64	99.65	0.32	0.32	99.83	99.91	0.11	0.07	99.73	0.22
	FedAvg*	67.25	0.83	0.83	57.09	57.09	0.83	0.83	57.09	57.09	0.83	0.83	57.09	57.09	0.83	57.09
	RLR	34.83	97.47	98.86	0.42	0.39	0.75	0.75	29.35	27.89	0.58	1.06	35.52	35.95	33.25	21.59
	RFA	60.51	1.04	0.37	42.25	44.61	0.57	**0.28**	43.26	44.38	0.46	0.39	47.21	48.95	0.52	45.11
	MKrum	62.81	0.65	0.60	53.27	53.95	0.70	0.64	55.65	55.46	0.44	0.44	54.05	54.05	0.58	54.41
	Foolsgold	**67.36**	99.59	99.74	0.33	0.24	99.45	99.83	0.46	0.15	99.79	99.91	0.16	0.07	99.72	0.24
	MM	65.81	99.92	99.95	0.07	0.05	99.96	99.99	0.03	0.00	99.79	99.91	0.17	0.08	99.92	0.07
	Lockdown	65.67	70.88	17.49	14.96	34.51	73.69	49.96	14.02	20.62	74.55	23.34	15.73	34.13	51.65	22.33
	MASA	67.04	**0.51**	**0.35**	**56.39**	**56.56**	**0.40**	0.48	**56.62**	**57.93**	**0.35**	**0.35**	**56.56**	**56.56**	**0.41**	**56.77**

Table 2. The MA and RA results of MASA on non-IID CIFAR-100 datasets under Badnet, Lie, and Neurotoxin attacks.

Method	Badnet		Neurotoxin		Lie		Avg. MA↑	Avg. RA↑
	MA↑	RA↑	MA↑	RA↑	MA↑	RA↑		
RLR	3.66	3.94	2.47	2.53	1.27	0.94	2.47	2.47
RFA	15.11	13.87	16.03	14.61	15.22	12.83	15.45	14.57
MKrum	49.74	0.42	49.00	1.20	47.27	0.08	48.67	16.30
Lockdown	55.71	6.84	29.42	28.06	55.78	6.06	46.97	30.23
MASA	**56.10**	**50.00**	**56.44**	**50.22**	**56.14**	**50.54**	**56.23**	**50.25**

formance loss in the main task. On the other hand, MKrum is effective in identifying malicious local model updates on IID datasets, resulting in low BA. However, MKrum's inability to select all benign local model updates (since the number of selected updates per round is typically less than the number of benign updates) leads to performance degradation in the main task, thus yielding a lower RA.

In contrast, MASA leverages the different behaviors between backdoor and benign parameters and employs individual unlearning to accurately identify and expose malicious local model updates. Moreover, MASA retains most of the benign local model updates in the aggregation process, thereby preventing performance loss in the main task.

Performance of MASA in non-IID settings. Here, we consider a non-IID case with $\alpha = 0.1$ to simulate an extremely high data heterogeneity among clients. Such an extreme non-IID case complicates the defense against backdoor attacks in FL, as the variations in local data distributions can severely impact the consistency of local models. We report the MA and BA of MASA, RLR, MKrum, RFA, and Lockdown on CIFAR-100 datasets in Table 2. We observed that despite the challenge of extreme non-

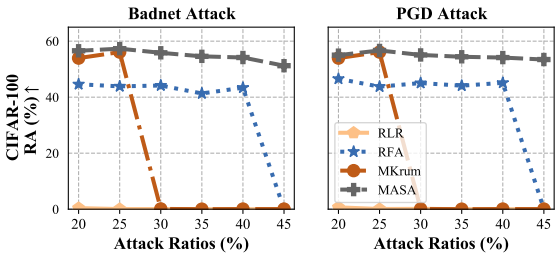

Figure 2. The RA of MASA, compared to RLR, RFA, and MKrum, under Badnet (left) and PGD (right) attacks with attack ratios ranging from 20% to 45%.

IIDness, MASA consistently outperforms the other methods across all attack scenarios, achieving the highest MA and RA. For example, under Badnet attack, MASA attains an MA of 56.10% and an RA of 50.00%, outperforming other methods like Lockdown, which achieves an MA of 55.71% and an RA of 6.84%. Similarly, under Neurotoxin attack, MASA achieves an MA of 56.44% and an RA of 50.22%, significantly higher than MKrum's MA of 49.00% and RA of 1.20%. These results demonstrate MASA's effectiveness in maintaining both high accuracy and robustness, even under severe non-IID cases across different attacks. MASA incorporates a pre-unlearning model fusion, which effectively mitigates the challenges posed by diverse local model updates, enabling the detection of backdoored models in such intensely non-IID scenarios.

Performance of MASA under various attack ratios. We conduct experiments to validate MASA's robustness in scenarios with high attack ratios. Specifically, we vary the attack ratio from 20% to an extreme value of 45% for both

Table 3. Performance of different methods in cross-device FL settings on CIFAR-10 dataset under DBA and Lie attacks.

Method	DBA			Lie			Avg. BA↓	Avg. RA↑
	MA↑	BA↓	RA↑	MA↑	BA↓	RA↑		
FedAvg	84.91	99.89	0.09	84.73	99.99	0.01	99.94	0.05
FedAvg*	85.23	1.43	81.71	85.23	1.43	81.71	1.43	81.71
RFA	83.86	1.71	81.53	84.76	1.52	81.66	1.62	81.60
Mkrum	83.98	1.83	80.93	84.12	1.33	82.29	1.58	81.61
Lockdown	84.14	99.99	0.00	83.49	99.99	0.01	99.99	0.01
MASA	**85.11**	**1.31**	**82.91**	**84.80**	**1.10**	**83.03**	**1.21**	**82.97**

Badnet and PGD attacks on CIFAR-100 dataset. The results, summarized in Figure 2, demonstrate that MASA consistently outperforms other methods, including RLR, RFA, and MKrum, across all attack ratios. As the attack ratio increases, MASA consistently maintains the highest robustness, though it experiences a slight drop in RA. This decrease occurs because MASA discards more local model updates in response to the growing number of malicious clients in the system. In contrast, MKrum exhibits sensitivity to attack ratios, losing its effectiveness when the attack ratio reaches 30% or higher, and RLR fails to defend against both attacks across all attack ratios. While RFA shows some degree of robustness, it loses its effectiveness under both attacks when the attack ratio reaches 45%.

Effectiveness of MASA in cross-device FL with client sampling. The majority of our experiments are conducted within the cross-silo FL setting. It is also important to consider the cross-device FL scenario, where a large number of clients participate in the system. To this end, we simulate 100 clients, with the server randomly selecting 20 clients per round to perform training. This random sampling of clients enables 1) dynamically selected local updates to contribute to refining the global model, and 2) the possibility of selecting a majority of malicious clients, creating a more dynamic and challenging environment for AGR operations. We present the MA, BA, and RA of different baselines on CIFAR-10 under DBA and Lie attacks in Table 3. Generally, MASA consistently achieves the best performance in terms of all evaluation metrics among the evaluated methods. Specifically, MASA achieves an average RA of 82.97% and an average BA of 1.21%, outperforming Mkrum with a gap of +0.37% and +1.36%, respectively. While Lockdown achieves competitive MA under DBA, however, it fails to provide any robustness in both cases. MASA's ability to balance high accuracy and robustness in a challenging cross-device FL environment with client sampling underscores its effectiveness and adaptability in real-world scenarios.

Hyperparameter study. In this section, we discuss the impact of two critical hyperparameters in MASA: the fusion degree λ and the filter radius δ. The parameter λ controls the intensity of pre-unlearning model fusion, with smaller values indicating more intense fusion, while δ determines

Table 4. The TPR and FPR of MASA with different λ and δ on non-IID CIFAR-10 datasets under Badnet, DBA, and PGD attacks.

Setting	Badnet		DBA		PGD		Avg. TPR↑	Avg. FPR↓
	TPR↑	FPR↓	TPR↑	FPR↓	TPR↑	FPR↓		
$\lambda = 0.3$	81.56	2.50	80.00	7.00	81.44	10.25	81.00	6.58
$\lambda = 0.5$	95.13	1.75	94.75	6.25	92.81	9.25	94.23	5.75
$\lambda = 1.0$	98.94	7.00	98.12	15.50	97.44	15.00	98.17	12.50
$\delta = 0.5$	89.62	**0.25**	88.88	**5.75**	87.75	**7.50**	88.75	**4.50**
$\delta = 1.5$	97.38	78.00	**99.94**	55.50	97.19	81.50	98.17	71.67
$\delta = 2.0$	**99.88**	94.00	**99.94**	93.75	**99.19**	98.25	**99.67**	95.33
MASA	99.25	2.75	98.81	7.75	97.56	10.00	98.54	6.83

detection sensitivity, with smaller values reflecting stricter filtering. We report the average True Positive Rate (TPR) and False Positive Rate (FPR) over training for different settings on the non-IID CIFAR-10 dataset under three different attacks in Table 4. It is observed that when $\lambda = 0.3$, the TPR is relatively lower compared to cases with higher λ values. This suggests that the aggregated model update predominates in the local models, compromising their unique features excessively. Therefore, we recommend selecting the value of λ between just above 0.5 and 1.0. It is important to note that when $\lambda = 1.0$, where pre-unlearning model fusion is disabled, the FPR increases across all cases, highlighting the effectiveness of model fusion. On the other hand, δ affects the strictness of the outlier filtering process. Lower values (e.g., $\delta = 0.5$) result in significantly lower FPRs (an average of 4.50%) but also lead to a notable reduction in TPRs (an average of 88.75%). In contrast, higher values (e.g., $\delta = 2.0$) nearly maximize TPRs across all attacks but cause a substantial increase in FPR. The parameter δ provides flexibility in balancing TPR (main task performance) and FPR (backdoor robustness) in MASA, allowing it to be tailored to specific security requirements. In comparison, MASA with a λ of 0.7 and a δ of 1.0 strikes an optimal balance between TPR and FPR in this setting. We additionally conduct experiments on various proxy data sizes and present the results in Appendix D.

7. Conclusion

We propose MASA to defend against stealthy backdoor attacks in FL. MASA identifies malicious parameters by leveraging individual unlearning on the main task for each local model, based on the observation that malicious parameters are less active on the main task. To address non-IID challenges, MASA incorporates a pre-unlearning model fusion mechanism, incorporating global information with local models to improve the consistency in unlearning behaviors across local models. It then filters malicious updates using a newly introduced metric, the median deviation score, based on the accumulated unlearning loss values of local models. We extensively evaluate MASA's superior performance across various scenarios and provide a detailed analysis of its hyperparameters.

References

[1] Manaar Alam, Hithem Lamri, and Michail Maniatakos. Get rid of your trail: Remotely erasing backdoors in federated learning. *arXiv preprint arXiv:2304.10638*, 2023. 2

[2] Manaar Alam, Esha Sarkar, and Michail Maniatakos. Perdoor: Persistent non-uniform backdoors in federated learning using adversarial perturbations. *arXiv preprint arXiv:2205.13523*, 2022. 2

[3] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pages 2938–2948. PMLR, 2020. 1, 2, 6

[4] Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 6

[5] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017. 1, 2, 6

[6] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021. 2, 4

[7] David Byrd and Antigoni Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–9, 2020. 1

[8] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020. 1, 2

[9] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002. 5

[10] Pei Fang and Jinghui Chen. On the vulnerability of backdoor defenses for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11800–11808, 2023. 2

[11] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, pages 301–316, 2020. 1, 2, 6

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[13] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 2, 3, 4, 6

[14] Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pages 3521–3530. PMLR, 2018. 1, 2

[15] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[18] Siquan Huang, Yijiang Li, Chong Chen, Leyu Shi, and Ying Gao. Multi-metrics adaptively identifies backdoors in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4652–4662, 2023. 1, 2, 6

[19] Tiansheng Huang, Sihao Hu, Ka-Ho Chow, Fatih Ilhan, Selim Tekin, and Ling Liu. Lockdown: Backdoor defense for federated learning with isolated subspace training. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 6

[20] Ji Chu Jiang, Burak Kantarci, Sema Oktug, and Tolga Soyata. Federated learning in smart city sensing: Challenges and opportunities. *Sensors*, 20(21):6230, 2020. 1

[21] Torsten Krauß and Alexandra Dmitrienko. Mesas: Poisoning defense for federated learning resilient against adaptive attackers. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1526–1540, 2023. 1, 2, 3, 6

[22] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009. 6

[23] Songze Li and Yanbo Dai. Backdoorindicator: Leveraging ood data for proactive backdoor detection in federated learning. *arXiv preprint arXiv:2405.20862*, 2024. 2

[24] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021. 3

[25] Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *International Conference on Machine Learning*, pages 19837–19854. PMLR, 2023. 3

[26] Han Liu, Zhiyuan Yu, Mingming Zha, XiaoFeng Wang, William Yeoh, Yevgeniy Vorobeychik, and Ning Zhang. When evil calls: Targeted adversarial voice over ip network. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2009–2023, 2022. 1

[27] Shuo Liu, Nirupam Gupta, and Nitin H Vaidya. Approximate byzantine fault-tolerance in distributed optimization. In *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, pages 379–389, 2021. 6

[28] Yi Liu, Jiangtian Nie, Xuandi Li, Syed Hassan Ahmed, Wei Yang Bryan Lim, and Chunyan Miao. Federated learning in the sky: Aerial-ground air quality sensing framework with

uav swarms. *IEEE Internet of Things Journal*, 8(12):9827–9837, 2020. 1

[29] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*, pages 240–254. Springer, 2020. 1

[30] Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Jiqiang Liu, and Xiangliang Zhang. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9020–9028, 2023. 2

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 3, 6

[33] Thomas Minka. Estimating a dirichlet distribution, 2000. 6

[34] Thuy Dung Nguyen, Tuan A Nguyen, Anh Tran, Khoa D Doan, and Kok-Seng Wong. Iba: Towards irreversible backdoor attacks in federated learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[35] Thien Duc Nguyen, Phillip Rieger, Roberta De Viti, Huili Chen, Björn B Brandenburg, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, et al. {FLAME}: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1415–1432, 2022. 1, 6

[36] 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*, volume 35, pages 9268–9276, 2021. 1, 2, 6

[37] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. 1, 2, 6

[38] Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, and Ahmad-Reza Sadeghi. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. *arXiv preprint arXiv:2201.00763*, 2022. 1, 2

[39] Devansh Shah, Parijat Dube, Supriyo Chakraborty, and Ashish Verma. Adversarial training in communication constrained federated learning. *arXiv preprint arXiv:2103.01319*, 2021. 2

[40] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021. 1

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4, 6

[42] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022. 1

[43] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in Neural Information Processing Systems*, 33:16070–16084, 2020. 1, 2, 6

[44] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 4

[45] Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. Crfl: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, pages 11372–11382. PMLR, 2021. 2

[46] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*, 2019. 1, 2, 6

[47] Guofu Xie, Bingyan Liu, and Yu Zhou. FedSKU: Defending backdoors in federated learning through selective knowledge unlearning, 2024. 2

[48] Jian Xu, Shao-Lun Huang, Linqi Song, and Tian Lan. Signguard: Byzantine-robust federated learning through collaborative malicious gradient filtering. *arXiv preprint arXiv:2109.05872*, 2021. 5

[49] Jiahao Xu, Zikai Zhang, and Rui Hu. Achieving byzantine-resilient federated learning via layer-adaptive sparsified model aggregation. *arXiv preprint arXiv:2409.01435*, 2024. 5

[50] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. Pmlr, 2018. 5

[51] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3fl: Adversarially adaptive backdoor attacks to federated learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[52] Junpeng Zhang, Hui Zhu, Fengwei Wang, Jiaqi Zhao, Qi Xu, and Hui Li. Security and privacy threats to federated learning: Issues, methods, and challenges. *Security and Communication Networks*, 2022(1):2886795, 2022. 1

[53] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning*, pages 26429–26446. PMLR, 2022. 1, 2, 6