

# A Four-Pronged Defense Against Byzantine Attacks in Federated Learning

Wei Wan<sup>\*†‡§</sup>  
wanwei\_0303@hust.edu.cn  
School of Cyber Science and  
Engineering, Huazhong University of  
Science and Technology

Shengshan Hu<sup>\*†‡§</sup>  
hushengshan@hust.edu.cn  
School of Cyber Science and  
Engineering, Huazhong University of  
Science and Technology

Minghui Li  
minghuili@hust.edu.cn  
School of Software Engineering,  
Huazhong University of Science and  
Technology

Jianrong Lu<sup>\*†‡§</sup>  
lujianrong@hust.edu.cn  
School of Cyber Science and  
Engineering, Huazhong University of  
Science and Technology

Longling Zhang<sup>\*†‡§</sup>  
longlingzhang@hust.edu.cn  
School of Cyber Science and  
Engineering, Huazhong University of  
Science and Technology

Leo Yu Zhang  
leo.zhang@griffith.edu.au  
School of Information and  
Communication Technology, Griffith  
University

Hai Jin<sup>\*†¶</sup>  
hjin@hust.edu.cn  
School of Computer Science and  
Technology, Huazhong University of  
Science and Technology

## ABSTRACT

Federated learning (FL) is a nascent distributed learning paradigm to train a shared global model without violating users' privacy. FL has been shown to be vulnerable to various Byzantine attacks, where malicious participants could independently or collusively upload well-crafted updates to deteriorate the performance of the global model. However, existing defenses could only mitigate part of Byzantine attacks, without providing an all-sided shield for FL. It is difficult to simply combine them as they rely on totally contradictory assumptions.

In this paper, we propose FPD, a **four-pronged defense** against both non-colluding and colluding Byzantine attacks. Our main idea is to utilize absolute similarity to filter updates rather than relative similarity used in existing works. To this end, we first propose a reliable client selection strategy to prevent the majority of threats in the bud. Then we design a simple but effective score-based detection method to mitigate colluding attacks. Third, we construct an enhanced spectral-based outlier detector to accurately discard abnormal updates when the training data is *not independent and*

*identically distributed* (non-IID). Finally, we design update denoising to rectify the direction of the slightly noisy but harmful updates. The four sequentially combined modules can effectively reconcile the contradiction in addressing non-colluding and colluding Byzantine attacks. Extensive experiments over three benchmark image classification datasets against four state-of-the-art Byzantine attacks demonstrate that FPD drastically outperforms existing defenses in IID and non-IID scenarios (with 30% improvement on model accuracy).

## CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems.**

## KEYWORDS

Reliable Client Selection, Byzantine Attack, Robust Federated Learning

## ACM Reference Format:

Wei Wan, Shengshan Hu, Minghui Li, Jianrong Lu, Longling Zhang, Leo Yu Zhang, and Hai Jin. 2023. A Four-Pronged Defense Against Byzantine Attacks in Federated Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3612474>

## 1 INTRODUCTION

Federated learning (FL) [14, 19] emerges as a new distributed machine learning paradigm recently, where the training data and the learning process are fully controlled by the clients, thus alleviating the privacy concern. However, due to its decentralized nature, FL is found to be highly vulnerable to Byzantine attacks, where malicious participants contribute poisoned updates to damage the global model. Generally, Byzantine attacks can be categorized into

<sup>\*</sup>National Engineering Research Center for Big Data Technology and System

<sup>†</sup>Services Computing Technology and System Lab

<sup>‡</sup>Hubei Key Laboratory of Distributed System Security

<sup>§</sup>Hubei Engineering Research Center on Big Data Security

<sup>¶</sup>Cluster and Grid Computing Lab

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10...\$15.00

<https://doi.org/10.1145/3581783.3612474>

non-colluding attacks [22, 28] where attackers upload malicious updates independently, and colluding attacks [2, 5, 8, 20, 27, 32] where attackers share information (e.g., training data, and model updates) to each other and collusively design well-crafted updates. In particular, colluding attackers tend to upload similar or totally identical updates to avoid being treated as outliers [17].

To defend against these two kinds of attacks, massive defensive schemes have been proposed in recent years. For the non-colluding attacks, existing defenses, such as Krum [3], FABA [25], Median [29], FedInv [34], AFA [16], manage to remove or circumvent the outliers based on the intuition that benign updates are much similar to each other due to the same optimization objective, while the malicious ones can be considered as outliers. To resist colluding attacks, existing works like FoolsGold [9], LOF [24], and Contra [1] propose to punish the relatively similar updates by distributing smaller weights in the aggregation stage. Unfortunately, these defenses (or simply combining them) cannot mitigate non-colluding and colluding attacks simultaneously, since the intuitions behind them are almost opposite arguing whether the malicious updates are similar to each other.

Recent studies like LFR [8], Zeno [28], FLTrust [4], DiverseFL [18] attempt to defend against both attacks simultaneously. Instead of relying on the distribution of the updates, they turn to an auxiliary dataset to validate the performance (e.g., loss or accuracy) of each update [8, 28], or construct a reliable update as a reference [4, 18]. These performance-based defenses hold that malicious updates inevitably degrade model performance in a degree. Although performing much better in non-colluding and part of colluding scenarios, these defenses fail to work when malicious updates are slightly noised but harmful (e.g., LIE attack [2]), especially when the data is *not independent and identically distributed* (non-IID). Moreover, the assumption of possessing an auxiliary dataset will violate users' privacy as they usually require that the auxiliary dataset has the same distribution as the clients' local training datasets. In summary, an effective defense providing an all-sided shield for FL is still missing yet.

To tackle these issues, we propose FPD, a **four-pronged defense** against both non-colluding and colluding Byzantine attacks. Our main observation is that the contradictory intuitions behind the existing two kinds of schemes arise because both of them rely on the relative similarity between updates due to the lack of a gold standard to evaluate each update in FL. In light of this, we propose to construct an artificial gold standard, which is an empirically determined threshold, to form absolute similarity that can be used to detect colluding attacks. Meanwhile, non-colluding attacks can still be detected based on relative similarity. In this way, the contradictory of solely exploiting relative similarity can be reconciled naturally. Specifically, we propose two defense modules relying on absolute similarity and relative similarity to defend against colluding attacks and non-colluding attacks, respectively. In addition, we design a reliable client selection strategy to prevent the majority of threats and the update denoising method to rectify the update directions, in order to further alleviate the impact of colluding attacks.

In summary, we offer the following contributions:

- We propose a new FL defense scheme FPD, which is effective in defending against non-colluding and colluding Byzantine attacks simultaneously.
- We propose two novel auxiliary defense modules (i.e., reliable client selection and update denoising) to further enhance the defense ability.
- We demonstrate the advantage of FPD via extensive experiments on three benchmark datasets against four state-of-the-art attacks. Compared with five distinguished defenses, our scheme achieves the best performance in both IID and non-IID scenarios.

## 2 BACKGROUND

### 2.1 Federated Learning

We consider a general FL system, consisting of a central server and  $K$  clients. Each client  $k \in [K]$  has a dataset  $D_k$ , the size of which is denoted as  $|D_k| = n_k$ . It is worth noting that each local dataset may be subject to a different distribution, that is, the clients' data may be distributed in a non-IID way. The clients aim to collaboratively train a shared global model  $\mathbf{w}$ . Apparently, the problem can be solved via minimizing the empirical loss, i.e.,  $\arg \min_{\mathbf{w}} f(D, \mathbf{w})$ , where  $D = \bigcup_{k=1}^K D_k$  and  $f(D, \mathbf{w})$  is a loss function (e.g., mean absolute error, cross-entropy). However, the optimization requires all the clients to share their raw data to a central server, which would result in a serious threat to client's privacy. Instead, FL obtains  $\mathbf{w}$  by optimizing  $\arg \min_{\mathbf{w}} \sum_{k=1}^K f(D_k, \mathbf{w})$ . Specifically, the FL system iteratively performs the following three steps until the global model converges:

- **Step I:** In the  $t$ -th iteration, the central server broadcasts a global model  $\mathbf{w}_t$  to the clients;
- **Step II:** After receiving  $\mathbf{w}_t$ , each client  $k$  trains a new local model  $\mathbf{w}_t^k$  over  $D_k$  by solving the optimization problem  $\arg \min_{\mathbf{w}_t^k} f(D_k, \mathbf{w}_t^k)$  and then uploads the local model update  $\mathbf{g}_t^k := \mathbf{w}_t^k - \mathbf{w}_t$  to the server;
- **Step III:** The server aggregates all the local updates according to client's proportional dataset size as follow:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \sum_{k=1}^K \frac{n_k}{n} \mathbf{g}_t^k, \text{ where } n = \sum_{k=1}^K n_k. \quad (1)$$

## 3 THREAT MODEL

### 3.1 Attack Model

Following previous studies [2, 9, 34], we consider a strong attack model where an adversary controls  $f$  out of the total  $K$  participants. The adversary can arbitrarily manipulate the data and the updates of the controlled clients. The goal of the adversary is to upload well-crafted malicious updates via the controlled clients to damage the global model accuracy. The controlled clients can collude with each other, and the adversary may possess the knowledge (e.g., the local updates) of other uncontrolled clients so as to initiate stronger attacks.

### 3.2 Defense Model

To design a practical defense, we cast away the following unrealistic assumptions that existing defenses rely on.

- **Training dataset sizes.** Recently proposed defense [34] assumes that the training dataset sizes of all the clients are known by the central server so that a fair weight distribution mechanism can be built. However, clients can arbitrarily report the sizes due to the distributed nature [21, 24].
- **Number of attackers.** Many defenses [3, 20, 25, 34] assume that the central server knows the number of attackers so as to determine how many updates should be removed. Nevertheless, the clients in FL are dynamically changing and cannot be determined in advance.
- **Auxiliary dataset.** Many defenses [4, 8, 13, 18, 28] rely on an auxiliary dataset whose distribution is the same as that of the clients, to evaluate the performance of the local updates. However, this assumption undoubtedly violates users' privacy.

On the contrary, our defense makes minimum assumptions. The only information the central server holds is the local updates uploaded by the clients. The goal of our defense is to achieve the competitive model accuracy in both non-colluding and colluding scenarios.

## 4 FPD: A FOUR-PRONGED DEFENSE AGAINST BYZANTINE ATTACKS

### 4.1 Motivation and Overview of FPD

After reviewing state-of-the-art defenses, we find that none of them can fully protect FL. Specifically, the colluding oriented defenses cannot defend against non-colluding attacks, and vice versa. Simply combining these two kinds of defenses seems promising, but they rely on totally contradictory assumptions. The former assumes that malicious updates are relatively similar, while the latter considers benign updates are more compact. Since all of these defenses employ relative similarity as a metric to filter out outliers, a combination of them inevitably leads to the rejection of benign updates in either colluding scenario or non-colluding scenario. Although the performance based defenses try to handle both of these two attacks, they are unable to detect malicious updates which are slightly perturbed but maintain toxicity.

To reconcile such a dilemma, we propose using absolute similarity to filter out extremely similar updates, and then employ an outlier detector based on relative similarity to discard abnormal updates. Furthermore, we propose two auxiliary defense modules (*i.e.*, the client selection and the update denoising) to further restrain the attack space of the poisoned updates, thus making it easier to filter out colluding and non-colluding poisoned updates. As shown in Fig. 1, our proposed FPD consists of the following four steps.

- **Step I: Reliable Client Selection.** Instead of randomly selecting a subset of clients to participate in each iteration, the central server selects the reliable clients who are more likely to contribute high quality updates according to the historical performance of each participant.
- **Step II: Mitigating Colluding Attacks.** The central server detects and rejects the updates that are excessively similar

in the direction space once receiving all the local updates from the currently selected clients.

- **Step III: Mitigating Non-Colluding Attacks.** The central server detects and rejects the outliers via a spectral-based outlier detector.
- **Step IV: Update Denoising.** The central server applies an autoencoder to reconstruct the malicious updates that are too similar to benign ones to detect.

REMARK 1. *Step I ensures that most of the malicious clients cannot participant in FL at all, in other words, only a limited number of compromised clients have a chance to poison the global model. Step II prohibits the adversary from designing excessively similar malicious updates, enhancing the difficulty of launching a covert attack. Step III guarantees that any update far from the overall distribution would be discarded. Step IV is designed to rectify the direction of the slightly noised but harmful updates. Note that Steps I and IV are directly dependent on the detection capacity of the Steps II and III, which inform the server whether an update is benign or malicious.*

### 4.2 Reliable Client Selection

Client selection is widely studied in the FL community, through which the researchers aim to reduce the communication overhead [6], solve the data heterogeneity challenge [33], and deal with the resource constrained FL scenarios [31]. However, it is rarely considered in the Byzantine-robust FL field. To the best of our knowledge, there are only two related defenses. In AFA [16], the authors propose a blocking mechanism to forbid the clients to participate in the subsequent iterations once they have shared sufficient bad updates. Recently, Wan *et al.* [23] proposed MAB-RFL, which applies a Beta distribution to estimate the probability of each client providing a benign update in the current iteration. However, both the defenses only focus on the overall performance of each client without taking their recent behaviors into account. Therefore, the attackers, in the early stages, can pretend to be benign clients by uploading well-trained updates to earn trust from the central server, and thus they will be constantly selected even though their latest updates are malicious.

Based on the above observation, we propose a new client selection strategy which considers both the overall and the recent performance of each client such that:

- The client who has uploaded too many malicious updates is selected with a low probability even though it has performed well in the recent iterations;
- The client who has contributed substantial benign updates while performing badly in the recent iterations is also selected with a low probability;
- Only the client who persistently shares benign updates is selected with a high probability.

Formally, in the client selection stage, the central server selects each client  $k$  with the probability:

$$p_k \sim \begin{cases} \text{Beta}(\alpha + B_k^O, \beta + M_k^O), & \text{if } \frac{B_k^O}{B_k^O + M_k^O} < \frac{B_k^R}{B_k^R + M_k^R} \\ \text{Beta}(\alpha + B_k^R, \beta + M_k^R), & \text{else} \end{cases} \quad (2)$$

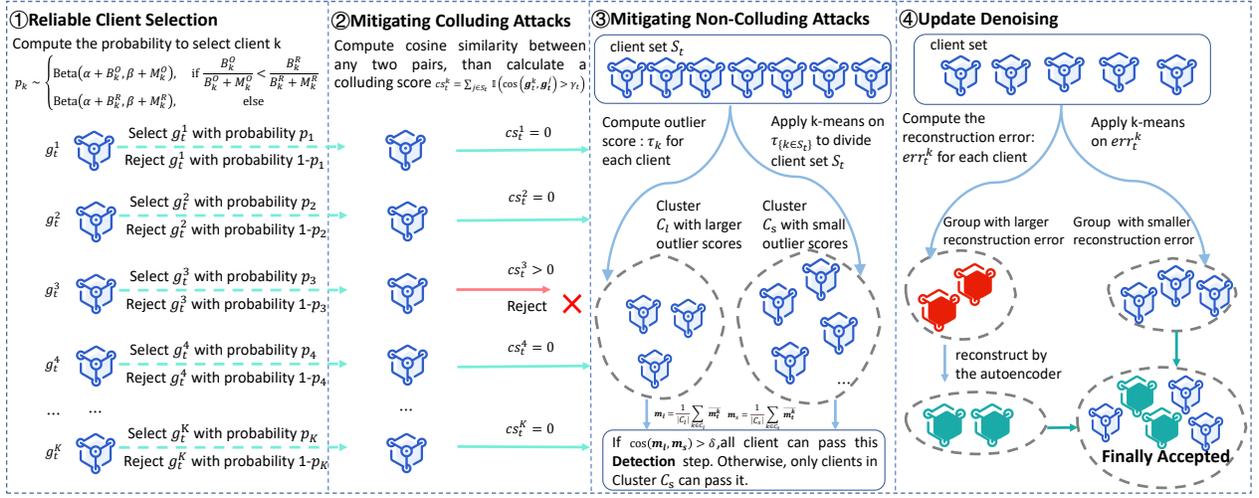


Figure 1: The workflow of our proposed FPD

where  $\alpha$  and  $\beta$  are prior parameters.  $B_k^O$  and  $M_k^O$  denote the overall frequencies that the local updates from client  $k$  are identified as benign and malicious respectively. Analogously,  $B_k^R$  and  $M_k^R$  indicate the recent frequencies. In this paper, we define the "recent" as the latest 10 iterations a specific client is selected.

Note that the central server possesses limited information about a client's identity (*i.e.*, benign or malicious) in the early iterations, thus it nearly makes a random choice, which deteriorates the convergence rate. As a remedy to this concern, we propose a bootstrap trick, by allowing all the clients to participate in the training in the first 10 iterations so as to fully understand their identities.

### 4.3 Mitigating Colluding Attacks

Recently, colluding attacks have aroused extensive attention for its effectiveness in designing covert but powerful Byzantine attacks. For example, LIE attack [2] adds well-crafted noise, which is tiny enough to circumvent the defense while huge enough to degrade the global model accuracy, to a benign update. IPM attack [27] reverses the direction of a benign update in order to maximize the attack effect. Wan *et al.* [24] proposed free-riding attack, where attackers train local models on small amounts of data but declare large training set sizes so as to dominate the global model. Fang *et al.* [8], and Shejwalkar *et al.* [20] proposed optimization based attacks respectively. Albeit different in implementation, all the attacks are based on a core idea, that is, the attackers should collude with each other to make the malicious updates as similar as possible or even totally identical. Colluding attack indeed poses a great threat to existing defenses as verified by our experiments. The difficulty in defending against colluding attack lies in the following facts:

- (i) Benign updates are inevitably got punished [1, 9, 24];
- (ii) It is hard to reconcile colluding attack and non-colluding attack.

To address these two challenges, we first propose a simple yet effective solution to mitigate colluding attacks by constructing absolute similarity. Specifically, we calculate a colluding score for

each selected client  $k$  as follow:

$$cs_t^k = \sum_{j \in S_t} \mathbb{I}(\cos(\mathbf{g}_t^k, \mathbf{g}_t^j) > \gamma_t), \quad (3)$$

where  $\mathbb{I}(\cdot)$  is the indicator function,  $\cos(\cdot, \cdot)$  indicates the cosine similarity,  $S_t$  is the selected client set in iteration  $t$ ,  $\gamma_t \in [-1, 1]$  denotes the tolerable cosine similarity threshold. As demonstrated in [1, 9, 23], the benign updates will not be extremely similar to each other in the direction space even in an IID scenario, thus it is easy to set the threshold  $\gamma_t$  (in our experiments we set  $\gamma_t = 0.8$ ) to filter out colluding attackers without affecting benign clients. Specifically, any client  $k$  with a positive colluding score  $cs_t^k$  will be regarded as malicious and rejected in this stage.

### 4.4 Mitigating Non-Colluding Attacks

In the scenario of non-colluding attacks, where malicious updates are quite different from each other in direction as well as magnitude, attackers can easily circumvent the detection of colluding attacks, which motivates the need of an additional abnormal detection step based on relative similarity. To this end, we borrow the idea from [7], where a spectral-based outlier detector is proposed. At a high level, the algorithm first computes the top singular vector of a matrix composed of all the involved vectors. Then any vector whose projection onto the singular vector (*i.e.*, the outlier score) is too large will be removed (by assuming the number of outliers is known in advance). Despite its good performance on several datasets with theoretical guarantee, it does not readily apply to our case due to the following challenges:

- **Challenge I.** As demonstrated in the original paper, the method performs badly in non-IID scenario, which is the most representative feature in FL.
- **Challenge II.** The method is highly sensitive to the magnitudes of the involved vectors even in the IID scenario.
- **Challenge III.** The method requires the number of outliers. Unfortunately, FL is a dynamic distributed network where

honest and malicious clients can join in and drop out arbitrarily.

To address Challenge I, we introduce momentum, which is shown to be effective to reduce the variance between updates [10, 15]. In this way, an IID-like distribution can be built. Formally, we compute the momentum vector as:

$$\mathbf{m}_t^k = \mathbf{g}_t^k + \lambda^{t-t_k} \mathbf{m}_{t_k}^k, \quad (4)$$

where  $t_k$  is the latest selected iteration for client  $k$ ,  $\lambda \in (0, 1)$  indicates the importance of historical information. Initially, we set  $\mathbf{m}_{t_k}^k = \mathbf{0}$ . Note that the iteration interval for a client being selected twice may be quite large, making the historical information that lies in the momentum vector  $\mathbf{m}_{t_k}^k$  obsolete. Thus we multiply it by a smaller discount factor  $\lambda^{t-t_k}$ , rather than using  $\lambda$  as existing works did.

To address Challenge II, we further normalize the momentum vector into an unit one:

$$\overline{\mathbf{m}}_t^k = \frac{\mathbf{m}_t^k}{\|\mathbf{m}_t^k\|}. \quad (5)$$

In this way, the outlier-detector will focus on the direction only, without being affected by the magnitude. Moreover, Eq. (5) also ensures that a single malicious update has a limited impact on the aggregation result, and a benign update with a small magnitude can contribute more information.

To address Challenge III, we apply the  $k$ -means algorithm to divide the normalized momentum vectors into two groups based on the outlier scores obtained by the outlier-detector due to its simpleness and effectiveness. Instead of simply treating the group with smaller outlier scores as being benign, we take the similarity between the two groups into consideration. Specifically, if the two groups are much similar (*i.e.*, the cosine similarity exceeds a threshold  $\delta$ ), it is very likely that all the updates are benign. In such a case, both groups will be kept for aggregation; otherwise, the group with larger outlier scores will be removed.

A detailed description for detecting non-colluding attack is summarized in Algorithm 1.

## 4.5 Update Denoising

Recent studies [2, 8, 20] show that attackers can upload well-crafted updates (by adding tiny noises to a benign update) that are extremely similar to benign ones to circumvent the defenses as well as maintain the attack effect. Distinguishing them from benign updates is really challenging. Therefore, instead of detecting and removing them, we denoise and utilize the slightly disturbed updates to facilitate the convergence. Specifically, we turn to an autoencoder to denoise the normalized momentum vectors that successfully get through the preceding detection steps, then the ones with large reconstruction errors will be reconstructed while the remaining vectors keep unchanged. Formally, the reconstruction error of client  $k$  in iteration  $t$  is given by:

$$\text{err}_t^k = \|\overline{\mathbf{m}}_t^k - \text{ae}(\overline{\mathbf{m}}_t^k)\|^2, \quad (6)$$

where  $\text{ae}(\cdot)$  represents the autoencoder. Then, we utilize the  $k$ -means algorithm to divide the normalized momentum vectors into two groups based on the reconstruction errors. The group with

---

### Algorithm 1 Mitigating Non-Colluding Attacks

---

**Input:** Current iteration  $t$ , left clients  $S_t$ , latest selected iterations  $\{t_k, k \in S_t\}$ , local updates  $\{\mathbf{g}_t^k, k \in S_t\}$ , momentum vectors  $\{\mathbf{m}_{t_k}^k, k \in S_t\}$ , acceptable difference between clusters  $\delta$ , importance of historical information  $\lambda$

**Output:** Set of removed clients  $R$

- 1: Compute the normalized momentum vectors  $\{\overline{\mathbf{m}}_t^k, k \in S_t\}$  through Eq. (4) and Eq. (5).
  - 2: Let  $\boldsymbol{\mu} = \frac{1}{|S_t|} \sum_{k \in S_t} \overline{\mathbf{m}}_t^k$ .
  - 3: Let  $\mathbf{G} = [\overline{\mathbf{m}}_t^k - \boldsymbol{\mu}]_{k \in S_t}$  be the matrix of centered vectors.
  - 4: Let  $\mathbf{v}$  be the top right singular vector of  $\mathbf{G}$ .
  - 5: Compute outlier scores defined as  $\tau_k = ((\overline{\mathbf{m}}_t^k - \boldsymbol{\mu}) \cdot \mathbf{v})^2$ .
  - 6: Apply  $k$ -means on  $\tau_{\{k \in S_t\}}$  to divide  $S_t$  into two clusters  $C_l$  with larger outlier scores and  $C_s$  with smaller outlier scores.
  - 7: Compute the mean vector of each cluster:
 
$$\mathbf{m}_l = \frac{1}{|C_l|} \sum_{k \in C_l} \overline{\mathbf{m}}_t^k;$$

$$\mathbf{m}_s = \frac{1}{|C_s|} \sum_{k \in C_s} \overline{\mathbf{m}}_t^k.$$
  - 8: **if**  $\cos(\mathbf{m}_l, \mathbf{m}_s) > \delta$  **then**
  - 9:   Let the removed set  $R = \emptyset$ .
  - 10: **else**
  - 11:   Let the removed set  $R = C_l$ .
  - 12: **end if**
  - 13: **return**  $R$
- 

larger reconstruction errors will be denoised by the autoencoder, and the other group remains unchanged.

Note that training such an autoencoder does not require any raw data shared by participants, thus users' privacy is well protected. Instead, we use the historical reliable normalized momentum vectors (derived from local updates) as the training samples. Moreover, the dimension of the momentum vector  $\mathbf{m}_t^k$  (the same with that of the model weights) is generally quite large, making it time-consuming to train the autoencoder. Hence we only consider the weights between the last two layers, which are decisive for the classification results [30].

## 5 EXPERIMENTS

### 5.1 Experimental Setup

**Datasets, models, and codes.** Our experiments are conducted on three benchmark image classification datasets: MNIST [12], Fashion-MNIST [26], and CIFAR-10 [11], as most of existing works did [1, 15, 23]. The model structures are consistent with those in [23]. Our codes are available at <https://github.com/CGCL-codes/FPD>.

**Data distribution.** We follow existing works [4, 34] to simulate non-IID data distribution. Roughly, the non-IID degree  $q \in [0, 1]$  is related to the proportion of the training data with a single specific label  $l \in [L]$  ( $L$  is the total kinds of the labels). A larger  $q$  indicates a higher non-IID degree, and  $q = \frac{1}{L}$  corresponds to the IID case. In our experiments, we set  $q = 0.5$  by default, which is the highest non-IID setting existing works considered. Moreover, the training set sizes vary among clients. For MNIST and Fashion-MNIST, they

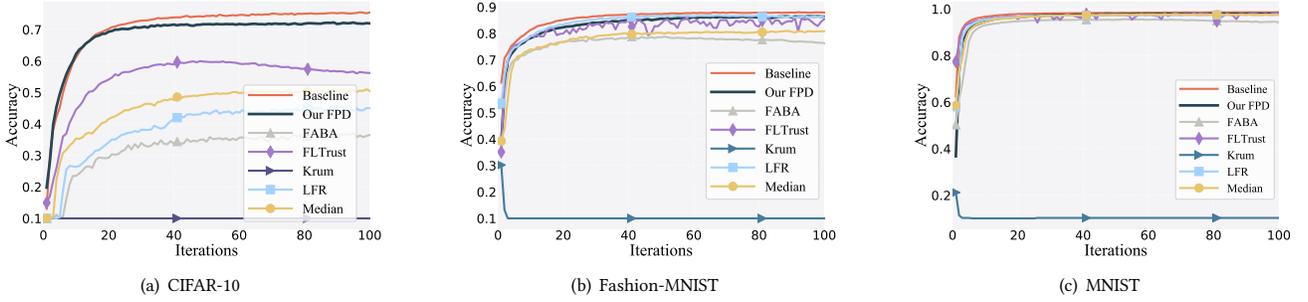


Figure 2: Model accuracy under LIE attack

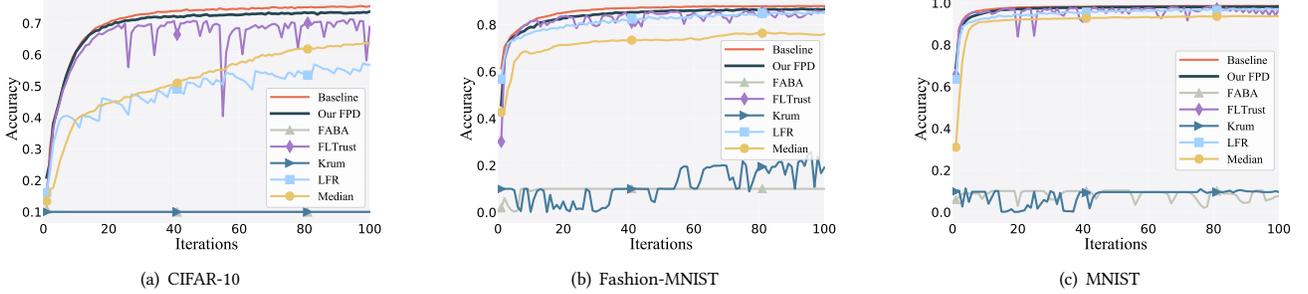


Figure 3: Model accuracy under IPM attack

are evenly sampled from  $[10, 500]$ . For CIFAR-10, they are randomly chosen from  $[1000, 1500]$ .

**Evaluated attacks.** We consider two colluding attacks, *i.e.*, *little is enough* (LIE) attack [2], and *inner product manipulation* (IPM) attack [27], as well as two non-colluding attacks, *i.e.*, *label flipping* (LF) attack [22], and *sign flipping* (SF) attack [24]. Note that our defense is not limited to these attacks.

It is noteworthy that all the parameter settings strictly follow the recommendations stated in the original papers, as it ensures the optimal attack effectiveness.

**Evaluated defenses.** We compare FPD with five state-of-the-art defenses, *i.e.*, Krum [3], FABA [25], Median [29], FLTrust [4], and LFR [8]. Besides, we also implement FedAvg [14] in non-adversarial case as a comparison (*i.e.*, Baseline).

It is worth noting that these defenses rely on additional assumptions, which enhance their defense effectiveness. For example, Krum, FABA, and LFR require prior knowledge of the number of attackers to determine the number of updates to be discarded, while FLTrust and LFR depend on a clean dataset to assess the trustworthiness of updates. In contrast, our proposed FPD does not introduce any unrealistic assumptions, making it a more desirable defense for deployment in realistic scenarios with limited knowledge (*e.g.*, just local updates).

**Performance metric and parameter settings.** We use *accuracy* (*i.e.*, the ratio of correctly predicted samples over all the testing samples) to evaluate the performance of each defense. For a fair comparison, all the experimental results are based on the mean of three repeated experiments. We set the number of total clients  $K = 50$ . The number of compromised clients  $f = 15$  by default. Each

client performs  $E = 3$  epochs of local training for faster convergence. The prior parameters  $\alpha = \beta = 1$ . The total iterations  $T = 100$ . The tolerable cosine similarity  $\gamma_t = 0.8$ . The importance of historical information  $\lambda = 0.1$ . For MNIST and Fashion-MNIST, the acceptable difference between clusters  $\delta = -0.1$ . For CIFAR-10,  $\delta = 0$ .

## 5.2 Experimental Results

**Defense against LIE attack.** In Fig. 2, we give the accuracy curves of the defenses under LIE attack on three different datasets. It is clear that the results vary across datasets. Specifically, on MNIST, Krum fail to defense. FPD, FLTrust, LFR, and Median achieve the similar accuracy with the Baseline. FABA performs slightly worse than the four defenses, with the accuracy gap of about 4%. On Fashion-MNIST, FPD and LFR perform best and are slightly superior to FLTrust, FABA, and Median. Krum still provides no protection. On the more complicated dataset CIFAR-10, the only defense that can effectively resist LIE attack is FPD. The other five defenses perform significantly worse than the Baseline with an accuracy gap of 20% ~ 65%.

**Defense against IPM attack.** As shown in Fig. 3, under IPM attack, FPD outperforms all the competitors on the three datasets with a minor gap to Baseline. Specifically, FABA and Krum are uncompetitive, because their accuracies hover at 10% in all scenarios. FLTrust and LFR, which perform as well as FPD on Fashion-MNIST and MNIST, cannot defend against IPM attack on CIFAR-10. To be specific, FLTrust fluctuates sharply, and LFR converges slowly. Although Median performs much better than FABA and Krum, its

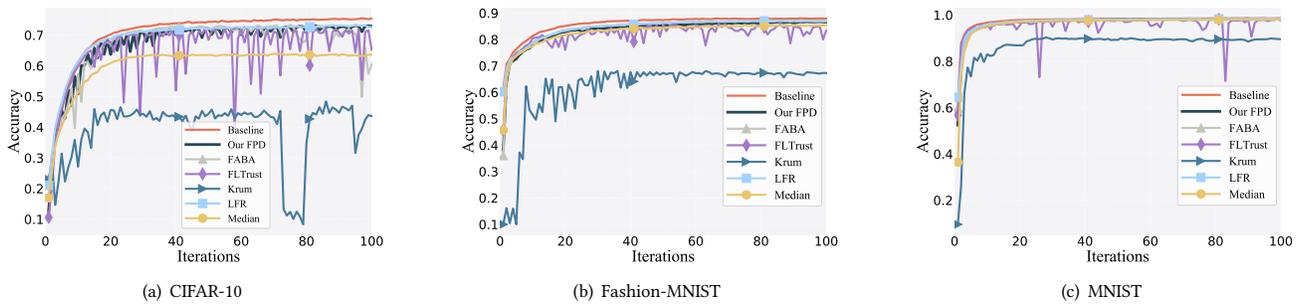


Figure 4: Model accuracy under LF attack

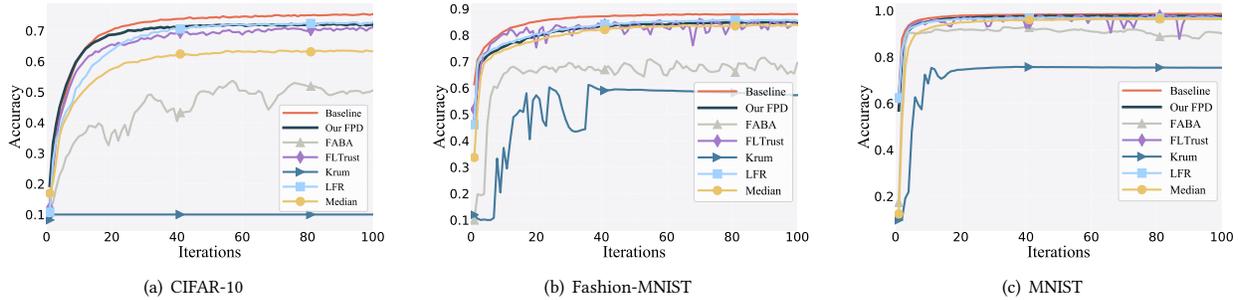


Figure 5: Model accuracy under SF attack

accuracy is not satisfactory, especially on CIFAR-10 and Fashion-MNIST.

**Defense against LF attack.** Fig. 4 presents the impact of LF attack on the defenses. In general, this attack is not as strong as the foregoing attacks (*i.e.*, LIE attack and IPM attack). Specifically, FPD and LFR can perfectly shield against the attack. FLTrust and FABA can also achieve similar performance in terms of accuracy, however, they are not steady. For example, the accuracy curves of FLTrust fluctuate on all datasets (noticeably on CIFAR-10 and MNIST), and FABA suffers a drop in accuracy on CIFAR-10. Krum provides quite limited protection with lowest accuracy.

**Defense against SF attack.** Fig. 5 shows the accuracy of the defenses under SF attack. We observe that FPD and LFR achieve the same global model accuracy, which comes near to Baseline. FLTrust is slightly inferior to the above two and incurs some fluctuation in accuracy. Median performs well on Fashion-MNIST and MNIST, however, its accuracy is about 10% lower than that of FPD and LFR on CIFAR-10. FABA can partially defend against SF attack on the most fundamental MNIST dataset, nevertheless, it performs badly on CIFAR-10 and Fashion-MNIST. Krum performs worst all the time. Worse still, its accuracy on CIFAR-10 is 10%, which means that Krum is dispensable.

**Impact of the percentage of compromised clients.** Table 1 shows the impact of the percentage of compromised clients under LIE attack on CIFAR-10 with the non-IID degree  $q = 0.5$ . We observe that as the percentage of attackers increases, the accuracy of all the defenses decreases. However, the degree of decreased accuracy varies from different defenses. Krum performs the worst. When there are 10% attackers, its accuracy is only 43.13%, which is 32.38%

Table 1: Impact of the percentage of compromised clients

Attackers	Accuracy (%)						
	Krum	FABA	Median	FLTrust	LFR	FPD	Baseline
10%	43.13	70.43	67.78	71.97	72.69	<b>74.81</b>	75.51
20%	10.00	63.44	59.68	68.07	57.60	<b>74.54</b>	
30%	10.00	36.50	50.47	56.20	45.01	<b>73.43</b>	
40%	10.00	10.00	10.00	48.73	10.00	<b>72.51</b>	
44%	10.00	10.00	10.00	48.06	10.00	<b>72.02</b>	
48%	10.00	10.00	10.00	46.42	10.00	<b>71.61</b>	

lower than the Baseline. When attackers account for 20% or more, Krum fails to converge (with the accuracy of 10%). FABA, Median, and LFR perform better than Krum, the accuracy gap between them and the Baseline is no more than 8% in the case of 10% attackers. However, the gap widens significantly as the number of attackers increases to 30%. When attackers account for more than 30%, the three defenses fail to converge. FLTrust outperforms the above four defenses. When there are no more than 20% attackers, FLTrust is not heavily affected, with the accuracy of about 8% lower than the Baseline. We also notice that FLTrust possesses the accuracy of 46.42% even in the case of 48% attackers, which is drastically higher (*i.e.*, 36.42%) than that of the above four defenses. However, it is about 30% lower than that of the Baseline, which means that FLTrust fails to offer a satisfactory global model in high-percentage attackers scenarios. In contrast, the proposed FPD achieves the best performance all the time. More importantly, it is highly stable. Specifically, its accuracy drops from 74.81% to 71.61% as the fraction of attackers increases from 10% to 48%.

**Table 2: Impact of the non-IID degree**

Non-IID Degree	Accuracy (%)						
	Krum	FABA	Median	FLTrust	LFR	FPD	Baseline
0.1	10.00	48.48	56.93	71.88	75.52	75.15	76.91
0.3	10.00	47.29	55.45	71.33	75.51	75.00	75.89
0.5	10.00	36.50	50.47	56.20	45.01	73.43	75.51
0.7	10.00	10.00	10.00	47.49	33.54	71.55	71.85
0.9	10.00	10.00	10.00	28.31	10.00	60.31	61.79
0.95	10.00	10.00	10.00	23.58	10.00	53.61	54.41

**Table 3: Ablation study on CIFAR-10 with 30% attackers**

Combination	Accuracy (%)			
	LIE	IPM	LF	SF
A+B+C+D	73.43	73.96	74.26	73.42
B+C+D	72.86	72.52	72.38	70.05
A+C+D	68.58	71.71	73.89	73.34
A+B+D	72.30	72.43	63.76	67.96
A+B+C	71.42	72.57	71.89	71.41

**Impact of the non-IID degree.** Table 2 shows the impact of the non-IID degree under LIE attack on CIFAR-10 with 30% compromised clients. We observe that as the non-IID degree  $q$  varies from 0.1 (*i.e.*, the IID case) to 0.95 (*i.e.*, the extremely non-IID case), all the schemes (including Baseline) achieve a lower and lower accuracy gradually. However, the accuracy of FPD is invariably comparable with that of Baseline (with the accuracy gap of 0.30%  $\sim$  2.08%). FLTrust and LFR perform well when  $q = 0.1$  and  $q = 0.3$ . However, when  $q \geq 0.5$ , their accuracy drops dramatically, which indicates that FLTrust and LFR do not apply to non-IID scenario. Krum, FABA, and Median cannot obtain a high-quality global model even in IID setting (*i.e.*,  $q = 0.1$ ) due to the remarkable attack effect of LIE attack.

**Ablation study on the absence of modules.** We perform an ablation study to understand the empirical effects of different modules in Table 3, where  $A, B, C, D$  indicate *reliable client selection*, *mitigating colluding attacks*, *mitigating non-colluding attacks*, and *update denoising* respectively. It can be seen that without module  $A$  the global model accuracy decreases 0.57%  $\sim$  3.37%, and without module  $D$  the global model accuracy decreases 1.39%  $\sim$  3.01%, which indicates that the two modules can slightly improve off-the-shelf defenses. Without module  $B$ , the global model accuracy under LIE attack drops 4.85%, which means that module  $B$  is effective to defend against colluding attacks. Without module  $C$ , the combination cannot achieve a desirable global model accuracy under non-colluding attacks (*i.e.*, LF and SF attacks), demonstrating the necessity of module  $C$ .

**Performance under mixed attack.** Previous experiments have demonstrated that FPD exhibits superior defense performance against individual colluding attacks or non-colluding attacks. As a result, one may naturally wonder whether FPD can withstand *mixed attacks* (MA) as well, *i.e.*, a group of attackers deploy colluding attacks while the remaining deploy non-colluding attacks. To this end, we conduct MA (half of attackers deploy LIE and the other half deploy LF) and compare it with LIE and LF, the results are shown in Tab. 4. Surprisingly, MA is not stronger than LIE, and sometimes even

**Table 4: Performance under MA on CIFAR-10 with 30% attackers**

Attack	Accuracy (%)					
	Krum	FABA	Median	FLTrust	LFR	FPD
LIE	10.00	36.50	50.47	56.20	45.10	73.43
LF	43.60	60.55	63.66	69.29	72.62	73.14
MA	48.97	65.32	64.82	62.44	59.13	73.46

weaker than LF. Specifically, our FPD performs consistently well under the three attacks with the highest accuracy, demonstrating its superiority in eliminating malicious updates. For FLTrust and LFR, MA is somewhat effective, but its impact is intermediate between that of pure LIE and LF. This is because both defenses are effective in defending against LF, but are weak in identifying LIE attackers. As for Krum, FABA, and Median, MA has the slightest effect on accuracy, we speculate that MA makes malicious updates more dispersed, thus making it easier for these similarity-based defenses to identify benign updates.

## 6 LIMITATIONS

Although our proposed FPD performs best, there are still some limitations.

**Suboptimal performance when attackers dominate.** Our defense suffers from an accuracy degrade when attackers dominate. Because the server lacks a gold standard, the server can only assume that the majority is reliable as did in existing defenses [23, 25, 29]. Though some works (*e.g.*, FLTrust) work in such an extreme case, they make a stronger assumption, *i.e.*, the server owns a clean dataset, which obviously violates the privacy requirements of FL.

**Lack of theoretical analysis.** In the literature of security studies in federated learning, it is difficult to provide a theoretical security analysis [23, 34], and our scheme is also heuristic. It is a challenging and promising topic and we leave it to our future work.

## 7 CONCLUSION

This paper proposed FPD, a four-pronged defense against Byzantine attacks. Specifically, FPD first performs reliable client selection to encourage participants to share high-quality updates. Next, a similarity-based filter is employed to prohibit the adversary from designing excessively similar malicious updates, enhancing the difficulty of launching a covert attack. Then, FPD utilizes a spectral-based outlier detector to remove the updates far from the overall distribution. Finally, an autoencoder is used to denoise the slightly noisy but harmful updates. Extensive experiments demonstrate that FPD is superior to existing defenses.

## ACKNOWLEDGMENTS

Shengshan’s work is supported in part by the National Natural Science Foundation of China (Grant No.U20A20177) and Hubei Province Key R&D Technology Special Innovation Project under Grant No.2021BAA032. Minghui’s work is supported in part by the National Natural Science Foundation of China (Grant No. 62202186) Shengshan Hu is the corresponding author.

## REFERENCES

- [1] Sana Awan, Bo Luo, and Fengjun Li. 2021. CONTRA: Defending Against Poisoning Attacks in Federated Learning. In *Proceedings of the 26th European Symposium on Research in Computer Security (ESORICS'21)*. 455–475.
- [2] Gilad Baruch, Moran Baruch, and Yoav Goldberg. 2019. A Little Is Enough: Circumventing Defenses for Distributed Learning. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS'19)*. 8632–8642.
- [3] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS'17)*. 119–129.
- [4] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. 2021. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In *Proceedings of the 28th Annual Network and Distributed System Security Symposium (NDSS'21)*.
- [5] Xiaoyu Cao and Neil Zhenqiang Gong. 2022. MPAF: Model Poisoning Attacks to Federated Learning based on Fake Clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*. 3396–3404.
- [6] Yae Jee Cho, Samarth Gupta, Gauri Joshi, and Osman Yağan. 2020. Bandit-based Communication-Efficient Client Selection Strategies for Federated Learning. In *Proceedings of the 54th Asilomar Conference on Signals, Systems, and Computers (ACSSC'20)*. 1066–1069.
- [7] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. 2019. Sever: A Robust Meta-Algorithm for Stochastic Optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*. 1596–1606.
- [8] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2020. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security'20)*. 1605–1622.
- [9] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. 2020. The Limitations of Federated Learning in Sybil Settings. In *Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID'20)*. 301–316.
- [10] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. 2021. Learning from History for Byzantine Robust Optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, Vol. 139. 5311–5319.
- [11] Alex Krizhevsky and Geoffrey Hinton. 2009. Learning Multiple Layers of Features from Tiny Images. (2009).
- [12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [13] Minghui Li, Wei Wan, Jianrong Lu, Shengshan Hu, Junyu Shi, Leo Yu Zhang, Man Zhou, and Yifeng Zheng. 2022. Shielding Federated Learning: Mitigating Byzantine Attacks with Less Constraints. In *Proceedings of 18th International Conference on Mobility, Sensing and Networking (MSN'22)*. 178–185.
- [14] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'17)*, Vol. 54. 1273–1282.
- [15] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. 2021. Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*.
- [16] Luis Muñoz-González, Kenneth T. Co, and Emil C. Lupu. 2019. Byzantine-Robust Federated Machine Learning through Adaptive Model Averaging. *arXiv preprint arXiv:1909.05125* (2019).
- [17] Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, and Prateek Mittal. 2022. SparseFed: Mitigating Model Poisoning Attacks in Federated Learning with Sparsification. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS'22)*, Vol. 151. 7587–7624.
- [18] Saurav Prakash and Amir Salman Avestimehr. 2020. Mitigating Byzantine Attacks in Federated Learning. *CoRR abs/2010.07541* (2020).
- [19] Xinyi Shang, Yang Lu, Gang Huang, and Hanzhi Wang. 2022. Federated Learning on Heterogeneous and Long-Tailed Data via Classifier Re-Training with Federated Features. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*. 2218–2224.
- [20] Virat Shejwalkar and Amir Houmansadr. 2021. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. In *Proceedings of the 28th Annual Network and Distributed System Security Symposium (NDSS'21)*.
- [21] Junyu Shi, Wei Wan, Shengshan Hu, Jianrong Lu, and Leo Yu Zhang. 2022. Challenges and Approaches for Mitigating Byzantine Attacks in Federated Learning. In *Proceedings of International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 139–146.
- [22] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursuoy, and Ling Liu. 2020. Data Poisoning Attacks Against Federated Learning Systems. In *Proceedings of the 25th European Symposium on Research in Computer Security (ESORICS'20)*, Vol. 12308. 480–501.
- [23] Wei Wan, Shengshan Hu, Jianrong Lu, Leo Yu Zhang, Hai Jin, and Yuanyuan He. 2022. Shielding Federated Learning: Robust Aggregation with Adaptive Client Selection. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*. 753–760.
- [24] Wei Wan, Jianrong Lu, Shengshan Hu, Leo Yu Zhang, and Xiaobing Pei. 2021. Shielding Federated Learning: A New Attack Approach and Its Defense. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC'21)*. 1–7.
- [25] Qi Xia, Zeyi Tao, Ziji Hao, and Qun Li. 2019. FABA: An Algorithm for Fast Aggregation against Byzantine Attacks in Distributed Neural Networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. 4824–4830.
- [26] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747* (2017).
- [27] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. 2020. Fall of Empires: Breaking Byzantine-tolerant SGD by Inner Product Manipulation. In *Proceedings of the 36th International Conference on Uncertainty in Artificial Intelligence (UAI'20)*. 261–270.
- [28] Cong Xie, Sanmi Koyejo, and Indranil Gupta. 2019. Zeno: Distributed Stochastic Gradient Descent with Suspicion-based Fault-tolerance. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, Vol. 97. 6893–6901.
- [29] Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter L. Bartlett. 2018. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, Vol. 80. 5636–5645.
- [30] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Proceedings of the 13th European Conference on Computer Vision (ECCV'14)*. 818–833.
- [31] Hangjia Zhang, Zhijun Xie, Roozbeh Zarei, Tao Wu, and Kewei Chen. 2021. Adaptive Client Selection in Resource Constrained Federated Learning Systems: A Deep Reinforcement Learning Approach. *IEEE Access* 9 (2021), 98423–98432.
- [32] Hangtao Zhang, Zeming Yao, Leo Yu Zhang, Shengshan Hu, Chao Chen, Alan Liew, and Zhetao Li. 2023. Denial-of-Service or Fine-Grained Control: Towards Flexible Model Poisoning Attacks on Federated Learning. *arXiv preprint arXiv:2304.10783* (2023).
- [33] Wenyu Zhang, Xiumin Wang, Pan Zhou, Weiwei Wu, and Xinglin Zhang. 2021. Client Selection for Federated Learning with non-IID Data in Mobile Edge Computing. *IEEE Access* 9 (2021), 24462–24474.
- [34] Bo Zhao, Peng Sun, Tao Wang, and Keyu Jiang. 2022. FedInv: Byzantine-Robust Federated Learning by Inverting Local Model Updates. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22)*. 9171–9179.