

# Defending against the Label-flipping Attack in Federated Learning

Najeeb Moharram Jebreel, Josep Domingo-Ferrer, David Sánchez and Alberto Blanco-Justicia

{najeeb.jebreel,josep.domingo,david.sanchez,alberto.blanco}@urv.cat

Universitat Rovira i Virgili

Catalonia, Tarragona

## Abstract

Federated learning (FL) provides autonomy and privacy by design to participating peers, who cooperatively build a machine learning (ML) model while keeping their private data in their devices. However, that same autonomy opens the door for *malicious peers* to poison the model by conducting either untargeted or targeted poisoning attacks. The *label-flipping (LF) attack* is a targeted poisoning attack where the attackers poison their training data by flipping the labels of some examples from one class (*i.e.*, the source class) to another (*i.e.*, the target class). Unfortunately, this attack is easy to perform and hard to detect and it negatively impacts on the performance of the global model. Existing defenses against LF are limited by assumptions on the distribution of the peers' data and/or do not perform well with high-dimensional models. In this paper, we deeply investigate the LF attack behavior and find that the contradicting objectives of attackers and honest peers on the source class examples are reflected in the parameter gradients corresponding to the neurons of the source and target classes in the output layer, making those gradients good discriminative features for the attack detection. Accordingly, we propose a novel defense that first dynamically extracts those gradients from the peers' local updates, and then clusters the extracted gradients, analyzes the resulting clusters and filters out potential bad updates before model aggregation. Extensive empirical analysis on three data sets shows the proposed defense's effectiveness against the LF attack regardless of the data distribution or model dimensionality. Also, the proposed defense outperforms several state-of-the-art defenses by offering lower test error, higher overall accuracy, higher source class accuracy, lower attack success rate, and higher stability of the source class accuracy.

**CCS Concepts:** • Computing methodologies → Distributed machine learning; • Security and privacy;

**Keywords:** Federated learning, Security, Poisoning attacks, Label-flipping attacks

---

Author's address: Najeeb Moharram Jebreel, Josep Domingo-Ferrer, David Sánchez and Alberto Blanco-Justicia, {najeeb.jebreel,josep.domingo,david.sanchez,alberto.blanco}@urv.cat, Universitat Rovira i Virgili, Av. Països Catalans 26, Catalonia, Tarragona, E-43007.

## 1 Introduction

Federated learning (FL) [19, 26] is an emerging machine learning (ML) paradigm that enables multiple peers to collaboratively train a shared ML model without sharing their private data with a central server. In FL, the peers train a global model received from the server on their local data, and then submit the resulting model updates to the server. The server aggregates the received updates to obtain an updated global model, which it re-distributes among the peers in the next training iteration. Therefore, FL improves privacy and scalability by keeping the peers' local data at their respective premises and by distributing the training load across the peers' devices (*e.g.*, smartphones) [6].

Despite these advantages, the distributed nature of FL opens the door for malicious peers to attack the global model [5, 17]. Since the server has no control over the peers' behavior, any of them may deviate from the prescribed training protocol to conduct either untargeted poisoning attacks [4, 41] or targeted poisoning attacks [2, 3, 11]. In the former, the attackers aim to cause model failure or non-convergence; in the latter, they aim to lead the model into misclassifying test examples with specific features into some desired labels. Whatever their nature, all these attacks result in bad updates being sent to the server.

The *label-flipping (LF) attack* [3, 11] is a type of targeted poisoning attack where the attackers poison their training data by flipping the labels of some correct examples from a source class to a target class, *e.g.*, flipping "spams" to "non-spams" or "fraudulent" activities to "non-fraudulent". Although the attack is easy for the attackers to perform, it has a significantly negative impact on the source class accuracy and, sometimes, on the overall accuracy [11, 29, 38]. Moreover, the impact of the attack increases as the ratio of attackers and their number of flipped examples increase [37, 38].

Several defenses against poisoning attacks (and LF in particular) have been proposed, which we survey in Section 4. However, they are either not practical [15, 31] or make specific assumptions about the distributions of local training data [1, 4, 9, 11, 35, 38, 42]. For example, [15] assumes the server has some data examples representing the distribution of the peers' data, which is not always a realistic assumption in FL; [4, 9, 35, 38, 42] assume the data to be independent and identically distributed (iid) among peers, which leads to poor performance when the data are non-iid [1]; [1, 11]

identify peers with a similar objective as attackers, which leads to a high rate of false positives when honest peers have similar local data [23, 32]. Moreover, some methods, such as multi-Krum (MKrum) [4] and trimmed mean (TMean) [42] assume prior knowledge of the ratio of attackers in the system, which is a strong assumption.

**Besides the assumptions on the distribution of peers’ data or their behavior, the dimensionality of the model is an essential factor that impacts on the performance of most of the above methods:** high-dimensional models are more vulnerable to poisoning attacks because an attacker can operate small but damaging changes on its local update without being detected [8]. Specifically, in the LF attack, the changes made to a bad update become less evident as the dimensionality of the update increases, because of the relatively small changes the attack causes on the whole update.

**To the best of our knowledge, there is no work that provides an effective defense against LF attacks without being limited by the data distribution and/or model dimensionality.**

**Contributions and plan.** In this paper, we present a novel defense against the LF attack that is effective regardless of the peers’ data distribution or the model dimensionality. Specifically, we make the following contributions:

- We conduct in-depth conceptual and empirical analyses of the attack behavior and we find a useful pattern that helps better discriminate between the attackers’ bad updates and the honest peers’ good updates. Specifically, we find that the contradictory objectives of attackers and honest peers on the source class’ examples are reflected in the parameters’ gradients connected to the source and target classes’ neurons in the output layer, making those gradients better discriminative features for attack detection. Moreover, we observe that those features stay robust under different data distributions and model sizes. **Also, we observe that different types of non-iid data require different strategies to defend against the LF attack.**
- We propose a novel defense that dynamically extracts the potential source and target classes’ gradients from the peers’ local updates, applies a clustering method on those gradients and analyzes the resulting clusters to filter out potential bad updates before model aggregation.
- We demonstrate the effectiveness of our defense against the LF attack through an extensive empirical analysis on three data sets with different deep learning model sizes, peers’ local data distributions and ratios of attackers up to 50%. In addition, we compare our approach with several state-of-the-art defenses and show its superiority at simultaneously delivering low test error, high overall accuracy, high source class accuracy, low attack success rate and stability of the source class accuracy.

The rest of this paper is organized as follows. Section 2 introduces preliminary notions. Section 3 formalizes the label-flipping attack and the threat model being considered. Section 4 discusses countermeasures for poisoning attacks in FL. Section 5 presents the design rationale and the methodology of the proposed defense. Section 6 details the experimental setup and reports the obtained results. Finally, conclusions and future research lines are gathered in Section 7.

## 2 Preliminaries

### 2.1 Deep neural network-based classifiers

A deep neural network (DNN) is a function  $F(x)$ , obtained by composing  $L$  functions  $f^l, l \in [1, L]$ , that maps an input  $x$  to a predicted output  $\hat{y}$ . Each  $f^l$  is a layer that is parametrized by a weight matrix  $w^l$ , a bias vector  $b^l$  and an activation function  $\sigma^l$ .  $f^l$  takes as input the output of the previous layer  $f^{l-1}$ . The output of  $f^l$  on an input  $x$  is computed as  $f^l(x) = \sigma^l(w^l \cdot x + b^l)$ . Therefore, a DNN can be formulated as

$$F(x) = \sigma^L(w^L \cdot \sigma^{L-1}(w^{L-1} \dots \sigma^1(w^1 \cdot x + b^1) \dots + b^{L-1}) + b^L).$$

DNN-based classifiers consist of a feature extraction part and a classification part [21, 27]. The classification part takes the extracted abstract features and makes the final classification decision. It usually consists of one or more fully connected layers where the output layer contains  $|C|$  neurons, with  $C$  being the set of all possible class values. The output layer’s vector  $o \in \mathbb{R}^{|C|}$  is usually fed to the softmax function that transforms it to a vector  $p$  of probabilities, which are called the confidence scores. In this paper, we use predictive DNNs as  $|C|$ -class classifiers, where the final predicted label  $\hat{y}$  is taken to be the index of the highest confidence score in  $p$ . Also, we analyze the output layer of DNNs to filter out updates resulting from the LF attack (called bad updates for short in what follows).

### 2.2 Federated learning

In federated learning (FL),  $K$  peers and an aggregator server collaboratively build a global model  $W$ . In each training iteration  $t \in [1, T]$ , the server randomly selects a subset of peers  $S$  of size  $m = C \cdot K \geq 1$  where  $C$  is the fraction of peers that are selected in  $t$ . After that, the server distributes the current global model  $W^t$  to all peers in  $S$ . Besides  $W^t$ , the server sends a set of hyper-parameters to be used to train the local models, which includes the number of local epochs  $E$ , the local batch size  $BS$  and the learning rate  $\eta$ . After receiving  $W^t$ , each peer  $k \in S$  divides her local data  $D_k$  into batches of size  $BS$  and performs  $E$  SGD training epochs on  $D_k$  to compute her update  $W_k^{t+1}$ , which she uploads to the server. Typically, the server uses the Federated Averaging (FedAvg) [26] method to aggregate the local updates and obtain the updated global model  $W^{t+1}$ . FedAvg averages the

updates proportionally to the number of training samples of each peer.

### 3 Label-flipping attack and threat model

In the label-flipping (LF) attack [3, 11, 38], the attackers poison their local training data by flipping the labels of training examples of a source class  $c_{src}$  to a target class  $c_{target} \in C$  while keeping the input data features unchanged. Each attacker poisons her local data set  $D_k$  as follows: for all examples in  $D_k$  whose class label is  $c_{src}$ , change their class label to  $c_{target}$ . After poisoning their training data, attackers train their local models using the same hyper-parameters, loss function, optimization algorithm and model architecture sent by the server. Thus, the attack only requires poisoning the training data, but the learning algorithm remains the same as for honest peers. Finally, the attackers send their bad updates to the server, so that they are aggregated with other good updates.

**Feasibility of the LF attack in FL.** Although the LF attack was introduced for centralized ML [3, 37], it is more feasible in the FL scenario because the server does not have access to the attackers' local training data. Furthermore, this attack can provoke a significant negative impact on the performance of the global model, but it cannot be easily detected because it does not influence non-targeted classes –it causes minimal changes in the poisoned model [38]. Furthermore, LF can be easily performed by non-experts and does not impose much computation overhead on attackers because it is an off-line computation that is done before training.

**Assumptions on training data distribution.** Since the local data of the peers can come from heterogeneous sources [6, 39], they may be either identically distributed (iid) or non-iid. In the iid setting, each peer holds local data representing the whole distribution, which makes the locally computed gradient an unbiased estimator of the mean of all the peers' gradients. The iid setting requires each peer to have examples of all the classes in a similar proportion as the other peers. In the non-iid setting, the distributions of the peers' local data sets can be different in terms of the classes represented in the data and/or the number of examples each peer has of each class. We assume that the distributions of the peers' training data may range from extreme non-iid to pure iid. Consequently, each peer may have local data with i) all the classes being present in a similar proportion as in the other peers' local data (iid setting), ii) some classes being present in a different proportion (mild non-iid setting), or iii) only one class (extreme non-iid setting, because the class held by a peer is likely to be different from the class held by another peer). The number of peers that have a specific class  $c$  in their training data can be denoted as  $K_c = |\{k \in K | c \in \text{Classes}(D_k)\}|$ .

**Threat model.** We consider an attacker or a coalition of  $K'_{c_{src}}$  attackers, with  $K'_{c_{src}} \leq (K_{c_{src}}/2)$ , for the iid and the mild

non-iid settings, and  $K'_{c_{src}} < K_{c_{target}}$  for the extreme non-iid setting (see Section 5.2 for a justification). The  $K'_{c_{src}}$  attackers perform the LF attack by flipping their training examples labeled  $c_{src}$  to a chosen target class  $c_{target}$  before training their local models. Furthermore, we assume the aggregator to be honest and non-compromised, and the attacker(s) to have no control over the aggregator or the honest peers. The goal of the attackers is to degrade as much as possible the performance of the global model on the source class examples at test time.

### 4 Related work

The defenses proposed in the literature to counter poisoning attacks (and LF attacks in particular) against FL are based on one of the following principles:

- *Evaluation metrics.* Approaches under this type exclude or penalize a local update if it has a negative impact on an evaluation metric of the global model, e.g. its accuracy. Specifically, [15, 31] use a validation data set on the server to compute the loss on a designated metric caused by each local update. Then, updates that negatively impact on the metric are excluded from the global model aggregation. However, realistic validation data require server knowledge on the distribution of the peers' data, which conflicts with the FL idea whereby the server does not see the peers' data.
- *Clustering updates.* Approaches under this type cluster updates into two groups, where the smaller group is considered potentially malicious and, therefore, disregarded in the model learning process. Auror [35] and multi-Krum (MKrum) [4] assume that the peers' data are iid, which results in high false positive and false negative rates when the data are non-iid [1]. Moreover, they require previous knowledge about the characteristics of the training data distribution [35] or the number of expected attackers in the system [4].
- *Peers' behavior.* This approach assumes that malicious peers behave similarly, which means that their updates will be more similar to each other than to those of honest peers. Consequently, updates are penalized based on their similarity. For example, FoolsGold (FGold) [11] and CONTRA [1] limit the contribution of potential attackers with similar updates by reducing their learning rates or preventing them from being selected. However, they also tend to incorrectly penalize good updates that are similar, which results in substantial drops in the model performance [23, 32].
- *Update aggregation.* This approach uses robust update aggregation methods that are sensitive to outliers at the coordinate level, such as the median [42], the trimmed mean (Tmean) [42] or the repeated median (RMedian) [36]. In this way, bad updates will have little to no influence on the global model after aggregation.

Although these methods achieve good performance with updates resulting from iid data for small DL models, their performance deteriorates when updates result from non-iid data, because they discard most of the information in model aggregation. Moreover, their estimation error scales up with the size of the model in a square-root manner [8]. Furthermore, RMedian [36] involves high computational cost due to the regression process it performs, whereas Tmean [42] requires explicit knowledge about the fraction of attackers in the system.

Several works focus on analyzing specific parts of the updates to defend against poisoning attacks. [16] proposes analyzing the output layer's biases to distinguish bad updates from good ones. However, it only considers the model poisoning attacks in the iid setting. FGold [11] analyzes the output layer's weights to counter data poisoning attacks, but it has the shortcomings mentioned above. [38] uses PCA to analyze the weights associated with the possibly attacked source class and excludes potential bad updates that differ from the majority of updates in those weights. However, the method needs an explicit knowledge about the possibly attacked source class or performs a brute-force search to find it, and is only evaluated under the iid setting with simple DL models. CONTRA [1] integrates FGold [11] with a reputation-based mechanism to penalize potential bad updates and prevent peers with low reputation from being selected. However, the method is only evaluated under mild non-iid settings using different Dirichlet distributions [28]. The methods just cited share the shortcomings of (i) making assumptions on the distributions of peers' data and (ii) not providing analytical or empirical evidence of why focusing on specific parts of the updates contributes towards defending against the LF attack.

In contrast, we analytically and empirically justify why focusing on the gradients of the parameters connected to the neurons of the source and target classes in the output layer is more helpful to defend against the attack. Also, we propose a novel defense that stays robust under different data distributions and model sizes, and does not require prior knowledge about the number of attackers in the system.

## 5 Our defense against LF attacks

In this section, we first introduce the rationale of our proposal. Based on that, we present the design of an effective defense against the label-flipping attack.

### 5.1 Rationale of our defense

The effectiveness of any defense against the LF attack depends on its ability to distinguish good updates sent by honest peers from bad updates sent by attackers. In this section,

we conduct comprehensive theoretical and empirical analyses of the attack behavior to find a discriminative pattern that better differentiates good updates from bad ones.

**Theoretical analysis of the LF attack.** To understand the behavior of the LF attack from an analytical perspective, let us consider a classification task where each local model is trained with the *cross-entropy* loss over one-hot encoded labels. First, the vector  $o$  of the output layer neurons (*i.e.*, the logits) is fed into the *softmax* function to compute the vector  $p$  of probabilities as

$$p_k = \frac{e^{o_k}}{\sum_{j=1}^{|C|} e^{o_j}}, \quad k = 1, \dots, |C|.$$

Then, the loss is computed as

$$\mathcal{L}(y, p) = - \sum_{k=1}^{|C|} y_k \log(p_k),$$

一个热编码器标签的交叉熵损失

where  $y = (y_1, y_2, \dots, y_{|C|})$  is the corresponding one-hot encoded true label and  $p_k$  denotes the confidence score predicted for the  $k^{th}$  class. After that, the gradient of the loss w.r.t. the output  $o_i$  of the  $i^{th}$  neuron (*i.e.*, the  $i^{th}$  neuron error) in the output layer is computed as

$$\begin{aligned} \delta_i &= \frac{\partial \mathcal{L}(y, p)}{\partial o_i} \\ &= - \sum_{j=1}^{|C|} \frac{\partial \mathcal{L}(y, p)}{\partial p_j} \frac{\partial p_j}{\partial o_i} \\ &= - \frac{\partial \mathcal{L}(y, p)}{\partial p_i} \frac{\partial p_i}{\partial o_i} - \sum_{j \neq i} \frac{\partial \mathcal{L}(y, p)}{\partial p_j} \frac{\partial p_j}{\partial o_i} \\ &= p_i - y_i. \end{aligned}$$

未算出来。这个式子的推导。

Note that  $\delta_i$  will always be in the interval  $[0, 1]$  when  $y_i = 0$  (for the wrong class neuron), while it will always be in the interval  $[-1, 0]$  when  $y_i = 1$  (for the true class neuron).

The gradient  $\nabla b_i^L$  w.r.t. the bias  $b_i^L$  of the  $i^{th}$  neuron in the output layer can be written as

$$\nabla b_i^L = \frac{\partial \mathcal{L}(y, p)}{\partial b_i^L} = \delta_i \frac{\partial \sigma^L}{\partial (w_i^L \cdot a^{L-1} + b_i^L)}, \quad (1)$$

where  $a^{L-1}$  is the activation output of the previous layer  $L-1$ .

Likewise, the gradient  $\nabla w_i^L$  w.r.t. the weights vector  $w_i^L$  connected to the  $i^{th}$  neuron in the output layer is

$$\nabla w_i^L = \frac{\partial \mathcal{L}(y, p)}{\partial w_i^L} = \delta_i a^{L-1} \frac{\partial \sigma^L}{\partial (w_i^L \cdot a^{L-1} + b_i^L)}. \quad (2)$$

From Equations (1) and (2), we can notice that  $\delta_i$  directly and highly impacts on the gradients of its connected parameters. For example, for the ReLU activation function, which is widely used in DL models, we get

$$\nabla b_i^L = \begin{cases} \delta_i, & \text{if } (w_i^L \cdot a^{L-1} + b_i^L) > 0; \\ 0, & \text{otherwise;} \end{cases}$$

and

$$\nabla w_i^L = \begin{cases} \delta_i a^{L-1}, & \text{if } (w_i^L \cdot a^{L-1} + b_i^L) > 0; \\ 0, & \text{otherwise.} \end{cases}$$

The objective of the attackers is to minimize  $p_{c_{src}}$  and maximize  $p_{c_{target}}$  for their  $c_{src}$  examples, whereas the objective of honest peers is exactly the opposite. We notice from Expressions (5.1), (1), and (2) that these contradicting objectives will be reflected on the gradients of the parameters connected to the *relevant* source and target output neurons. For convenience, in this paper, we use the term *relevant neurons' gradients* instead of the gradients of the parameters connected to the source and target output neurons. Also, we use the term *non-relevant neurons' gradients* instead of the gradients of the parameters connected to the neurons different from source and target output neurons. As a result, as the training evolves, the magnitudes of the relevant neurons' gradients are expected to be larger than those of the non-relevant and non-contradicting neurons. Also, the angle between the relevant neurons' gradients for an honest peer and an attacker is expected to be larger than those of the non-relevant neurons' gradients. That is because the error of the non-relevant neurons will diminish as the global model training evolves, especially when it starts converging (honest and malicious participants share the same training objectives for non-targeted classes). On the other hand, the relevant neurons' errors will stay large during model training because of the contradicting objectives. Therefore, the relevant neurons' gradients are expected to carry a more valuable and discriminative pattern for an attack detection mechanism than the whole model gradients or the output layer gradients, which carry a lot of information not relevant to the attack.

**Empirical analysis of the LF attack.** To empirically validate the analytical findings discussed above and see how the model size and the data distribution impact on the detection of LF attacks, we used exploratory analysis to visualize the gradients sent by peers in five different FL scenarios under label-flipping attacks: MNIST-iid, MNIST-Mild, MNIST-Extreme, CIFAR10-iid and CIFAR10-Mild. Besides the whole updates, we visualized the output layer's gradients and the relevant neurons' gradients. The FL attacks in the MNIST benchmarks consisted of flipping class 7 to class 1, while in the CIFAR10 benchmarks they consisted of flipping *Dog* to *Cat*. For the MNIST benchmarks, we used a simple DL model which contains about 22K parameters. For the CIFAR10 benchmarks we used the ResNet18 [14] architecture, which yields large models containing about 11M parameters. The details of the experimental setup are given in Section 6.1. In order to visualize the updates, we used Principal Component Analysis (PCA) [40] on the selected gradients and we plotted the first two principal components. We next report what we observed.

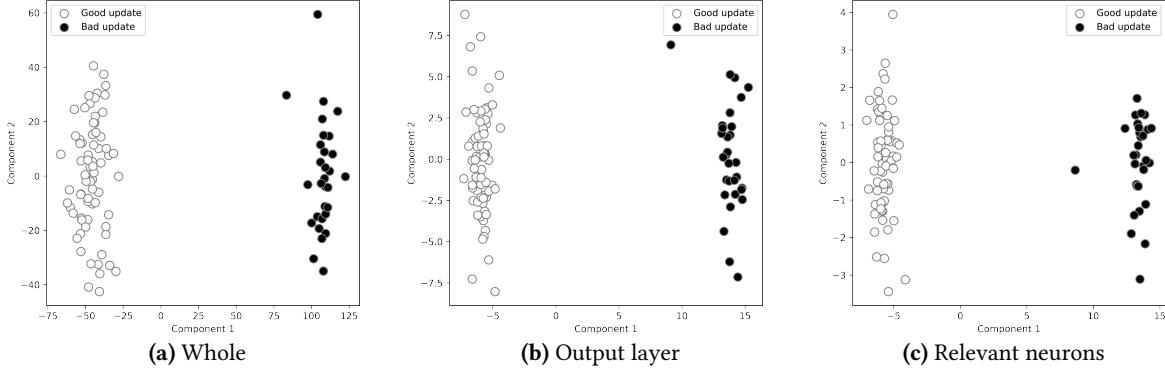
1) **Impact of model dimensionality.** Figures 1 and 2 show the gradients of whole local updates, gradients corresponding to the output layers, and relevant gradients corresponding to the source and target neurons from the MNIST-iid (30 bad updates out of 100) and the CIFAR10-iid (6 bad updates out of 20) benchmarks, respectively. In these two benchmarks, the training data were iid among peers.

The figures show that, when the model size is small (MNIST-iid), good and bad updates can be easily separated, whichever set of gradients is considered. On the other hand, when the model size is large (CIFAR10-iid), the attack's influence does not seem to be enough to distinguish good from bad updates when using whole update gradients; yet, the gradients of the output layer or those of the relevant neurons still allow for a crisp differentiation between good and bad updates.

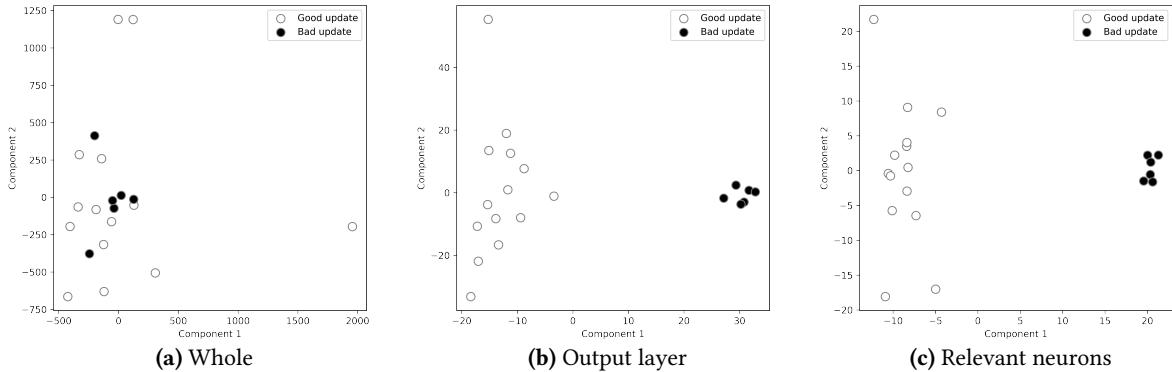
In fact, several factors make it challenging to detect LF attacks by analyzing an entire high-dimensional update. First, the computed errors for the neurons in a certain layer depend on all the errors for the neurons in the subsequent layers and their connected weights [33]. Thus, as the model size gets larger, the impact of the attack is mixed with that of the non-relevant neurons. Second, the early layers of DL models usually extract common features that are not class-specific [30]. Third, in general, most parameters in DL models are redundant [10]. **These factors cause the magnitudes of the whole gradients of good and bad updates and the angles between them to be similar, making DL models with large dimensions an ideal environment for a successful label-flipping attack.**

To confirm these observations, we performed the following experiment with the CIFAR10-iid benchmark. First, a chosen peer trained her local model on her data honestly, which yielded a good update. Then, the same peer flipped the labels of the source class *Cat* to the target class *Dog* and then trained her local model on the poisoned training data, which yielded a bad update. After that, we computed the magnitudes of and the angle between i) the whole updates, ii) the output layer gradients, iii) the relevant gradients related to  $c_{src}$  and  $c_{target}$ . Table 1 shows the obtained results, which confirm our analytical and empirical findings. It is clear that both whole gradients had approximately the same magnitude, and the angle between them was close to zero. On the other hand, the difference between the output layer gradients was large and even more significant in the case of the relevant neurons' gradients. As for non-relevant neurons, their gradients' magnitude and angle were not significantly affected because in such neurons there was no contradiction between the objectives of the good and the bad updates.

To underscore this point and see how the gradients of the non-relevant neurons vanish as the training evolves, while the gradients of the relevant neurons remain larger, we show the gradients' magnitudes during training in Figure 3. The magnitudes of those gradients for the MNIST-iid and the CIFAR10-Mild benchmarks are shown for ratios of attackers



**Figure 1.** First two PCs of the MNIST-iid benchmark gradients



**Figure 2.** First two PCs of the CIFAR10-iid benchmark gradients

**Table 1.** Comparison of the magnitudes and the angle of the gradients of a good and a bad update for the whole update, the output layer parameters, the parameters of the relevant source and target neurons, and the parameters of the non-relevant neurons.

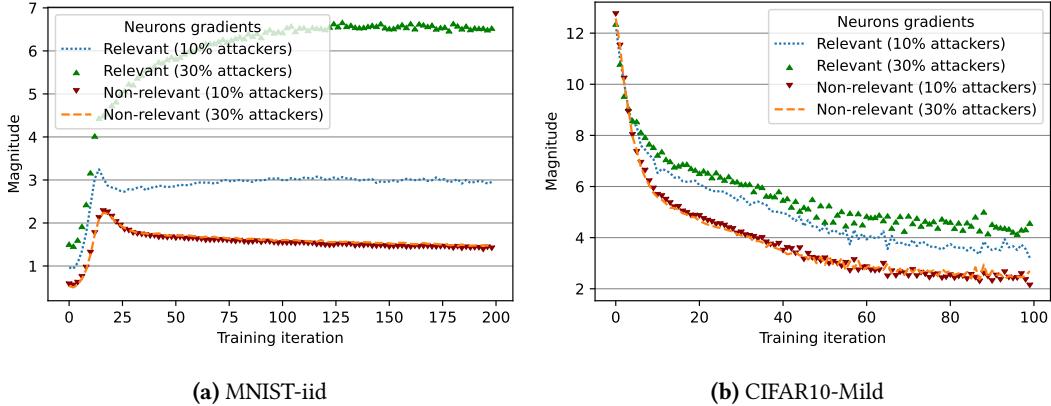
Gradients		Whole	Output layer	Relevant neurons	Non-relevant neurons
Magnitude	Good	351123	72.94	23.38	55.30
	Bad	351107	100.23	64.43	65.95
Angle		0.41	69.19	115	18

10% and 30%. We can see that, although the attackers' ratio and the data distribution had an impact on the magnitudes of those gradients, the gradients' magnitudes for the relevant source and target class neurons always remained larger.

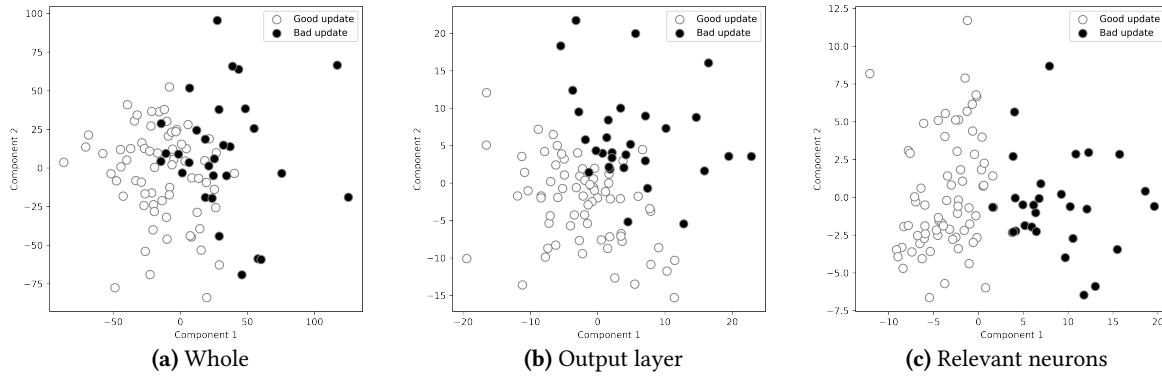
**2) Impact of data distribution.** Figure 4 shows the gradients of 100 local updates from the MNIST-Mild benchmark and their corresponding output layer and relevant neurons’ gradients, where 30 updates were bad. Figure 5 shows the same for the CIFAR10-Mild benchmark, where 6 out of 20 local updates were bad. In these two benchmarks, the training data were distributed following a mild non-iid Dirichlet distribution among peers [28] with  $\alpha = 1$ . Figure 4 shows that,

despite the model used for the MNIST-Mild benchmark being small, distinguishing between good and bad updates was harder than in the iid setting shown in Figure 1. It also shows that the use of the relevant neurons' gradients provided the best separation compared to whole update gradients or output layer gradients.

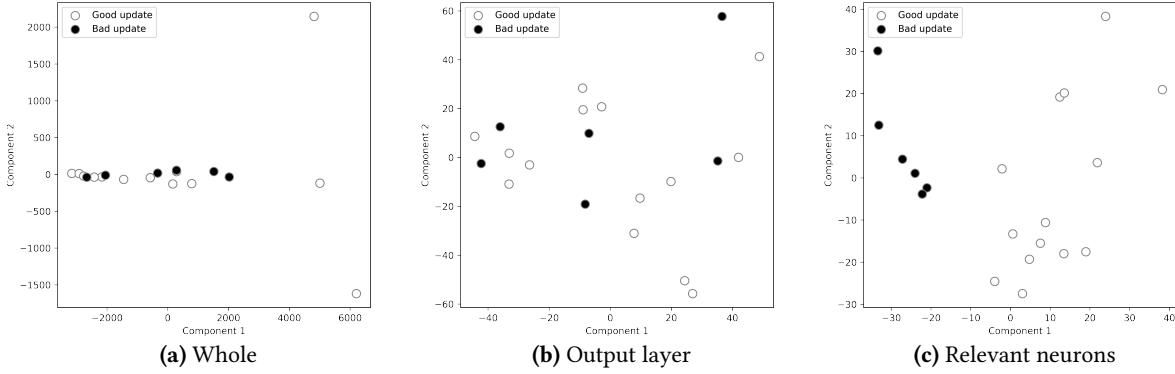
Figure 5 shows that the combined impact of model size and data distribution in the CIFAR-Mild benchmark made it very challenging to separate bad updates from good ones using whole update gradients or even using the output layer gradients. On the other hand, the relevant neurons' gradients gave a clearer separation.



**Figure 3.** Gradient magnitudes during training for relevant and non-relevant neurons



**Figure 4.** First two PCs of the MNIST-Mild benchmark gradients



**Figure 5.** First two PCs of the CIFAR10-Mild benchmark gradients

From the previous analyses, we can observe that analyzing the gradients of the parameters connected to the source and target class neurons led to better discrimination between good updates and bad ones for both the iid and the mild non-iid settings. We can also observe that, in general, those

gradients formed two clusters: one cluster for the good updates and another cluster for the bad updates. Moreover, the attackers' gradients were more similar among them and caused their clusters to be denser than the honest peers' clusters.

However, what would be the case when the data are extremely non-iid, that is, when each peer has local training data of a single class? Figure 6 shows the gradients of the relevant neurons' gradients of 100 local updates from the MNIST-Extreme benchmark, where each peer provided examples of a single class. In this experiment, 4 attackers out of the 10 peers who had examples of the class 7, flipped the labels of their training examples from the source class 7 to the target class 1. The figure shows that the gradients of the updates of each class form an individual cluster, and the 4 bad updates form a cluster that is very close to the cluster of the target class 1 updates. The explanation is that, in the extreme non-iid setting, most peers have classes different from  $c_{src}$  in their data, and hence, the honest peers have less influence on  $\delta_{c_{src}}$  than in the iid or the mild non-iid settings. Therefore, the alteration of  $\delta_{c_{src}}$  via decrease of  $p_{c_{src}}$  is less detectable than the alteration of  $\delta_{c_{target}}$  via increase of  $p_{c_{target}}$ , because averaging local updates should decrease both  $p_{c_{src}}$  and  $p_{c_{target}}$  (in this extreme non-iid setting, each class is absent from the local data of most peers). In this benchmark, bad updates got close to the target class cluster because both the attackers and class 1 honest peers shared a common objective, which is maximizing  $p_{c_1}$ . On the other hand, what made them different is the attackers' aim to minimize  $p_{c_7}$  to mitigate the remaining impact of the honest peers in  $K_{c_7}$  after the global model aggregation.

Based on the analyses and observations presented so far, we concluded that *an effective defense against the label-flipping attack needs to consider the following aspects:*

- Only the gradients of the parameters connected to the source and target class neurons in the output layer must be extracted and analyzed.
- If the data are iid or mild non-iid, the extracted gradients need to be separated into two clusters that are compared to identify which of them contains the bad updates.
- If the data are extremely non-iid, the extracted gradients need to be dynamically clustered so that the gradients of the peers that have data belonging to the same class fall in the same individual cluster. **Then, the bad updates cluster must be compared with the target class's cluster.**

## 5.2 Design of our defense

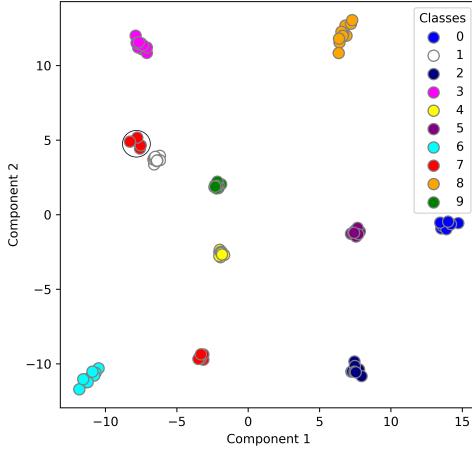
Considering the observations and conclusions discussed in the previous section, we present our proposed defense against the label-flipping attack in federated learning systems.

**Unlike other defenses, our proposal does not require a prior assumption on the peers' data distribution, is not affected by model dimensionality, and does not require prior knowledge about the proportion of attackers.** In each training iteration, our defense first separates the gradients of

the output layer parameters from the local updates. Then it dynamically identifies the two neurons with the highest gradient magnitudes as the potential source and target class neurons, and extracts the gradients of the parameters connected to them. Next, it applies a proper clustering method on those extracted gradients based on the peers' data distribution. Unlike existing approaches, we do not use a fixed strategy to address all types of local data distributions. Instead, we cluster the extracted gradients into two clusters using k-means [13] for the iid and mild non-iid settings, while we cluster them into multiple clusters using HDBSCAN [7] for the extreme non-iid setting. Thereafter, we further analyze the resulting clusters to identify the bad cluster. In the iid and the mild non-iid settings, we consider the size and the density of clusters. The smaller and/or the denser cluster is identified as a potentially bad cluster. In the extreme non-iid setting, we compare the two clusters with the same highest neuron gradients' magnitudes. The smaller cluster is identified as a potentially bad cluster. Finally, we exclude the updates corresponding to the potentially bad cluster from the aggregation phase. Note that discovering whether the data of peers are iid, mild non-iid, or extreme non-iid can be achieved by either i) projecting the extracted gradients into two dimensions and seeing the shape of the formed clusters, ii) asking each peer what classes she holds, or iii) using the sign of the bias gradient of each class output neuron (as mentioned in the previous section, the error of a peer's output neuron lies within  $[-1, 0]$  for the classes she holds, while it lies within  $[0, 1]$  for the classes she does not hold). In any case, assuming knowledge on the peers' class distribution is a much weaker requirement than assuming the peers' data follow certain distributions (*i.e.* many related works directly assume the iid setting).

We formalize our method in Algorithm 1. The aggregator server  $A$  starts a federated learning task by selecting a random set  $S$  of  $m$  peers, initializes the global model  $W^0$  and sends it to the  $m$  selected peers. Then, each peer  $k \in S$  locally trains  $W^t$  on her data  $D_k$  and sends her local update  $W_k^{t+1}$  back to  $A$ . Once  $A$  receives the  $m$  local updates, it computes their corresponding gradients as  $\{\nabla W_k^t = (W^t - W_k^{t+1})/\eta | k \in S\}$ . After that,  $A$  separates the gradients connected to the output layer neurons to obtain the set  $\{\nabla_k^{L,t} | k \in S\}$ .

**Identifying potential source and target classes.** After separating the gradients of the output layer, we need to identify the potential source and target classes, which is key to our defense. As we have shown in the previous section, the magnitudes of the gradients connected to the source and target class neurons for the attackers and honest peers are expected to be larger than the magnitudes of the other non-relevant classes. Thus, we can dynamically identify the potential source and target class neurons by analyzing the magnitudes of the gradients connected to the output layer neurons. To do so, for each peer  $k \in S$ , we



**Figure 6.** Extreme non-iid setting. First two PCs of the MNIST-Extreme source and target classes’ neurons gradients. Circled updates are bad.

compute the neuron-wise magnitude of each output layer neuron’s gradients  $\|\nabla_{i,k}^{L,t}\|$  and identify the two neurons with the highest two magnitudes  $imax_{1,k}$  and  $imax_{2,k}$  as potential source and target class for that peer under the extreme non-iid setting. For the iid or mild non-iid settings, after computing the output layer neuron magnitudes for all peers in  $S$ , we aggregate their neuron wise gradient magnitudes into the vector  $(\|\nabla_{1,S}^{L,t}\|, \dots, \|\nabla_{i,S}^{L,t}\|, \dots, \|\nabla_{|C|,S}^{L,t}\|)$ . We then identify the potential source and target class neurons  $imax_{1,S}$  and  $imax_{2,S}$  as the two neurons with the highest two magnitudes in the aggregated vector.

**Filtering in case of iid and mild non-iid updates.** For the local updates resulting from the iid and mild non-iid settings, we filter out bad updates by using the FILTER\_MILD procedure detailed in Procedure 1. First, we start by extracting the gradients connected to the identified potential source and target classes  $imax_{1,S}$  and  $imax_{2,S}$  from the output layer gradients of each peer. Then, we use the k-means [13] method with two clusters to group the extracted gradients into two clusters  $cl_1$  and  $cl_2$ . Once the two clusters are formed, we need to decide which of them contains the potential bad updates. To make this critical decision, we consider two factors: the size and the density of clusters. We mark the smaller and/or denser cluster as potentially bad. This makes sense: when the two clusters have similar densities, the smaller is probably the bad one, but if the two clusters are close in size, the denser and more homogeneous cluster is probably the bad one. The higher similarity between the attackers makes their cluster usually denser, as shown in the previous section. To compute the density of a cluster, we compute the pairwise angle  $\theta_{ij}$  between each pair of gradient vectors  $i$  and  $j$  in the cluster. Then, for each gradient vector  $i$  in the cluster, we

find  $\theta_{max,i}$  as the maximum pairwise angle for that vector. That is because no matter how far apart two attackers’ gradients are, they will be closer to each other due to the larger similarity of their directions compared to that of two honest peers. After that, we compute the average of the maximum pairwise angles for the cluster to obtain the inverse density value  $dns$  of the cluster. This way, the denser the cluster, the lower  $dns$  will be. After computing  $dns_1$  and  $dns_2$  for  $cl_1$  and  $cl_2$ , we compute  $score_1$  and  $score_2$  by re-weighting the computed clusters’ inverse densities proportionally to their sizes. If both clusters have similar inverse densities, the smaller cluster will probably have the lower score, or if they have similar sizes, the denser cluster will probably have the lower score. Finally, we use  $score_1$  and  $score_2$  to decide which cluster contains the potential bad updates. We compute the set  $bad\_peers$  as the peers in the cluster with the minimum score.

**Filtering in case of extreme non-iid updates.** For the local updates resulting from extreme non-iid data, we filter out potential bad updates by using the FILTER\_EXTREME procedure described in Procedure 2. First, from each peer’s output layer gradients, we extract the gradients connected to the potential source and target class neurons of that peer ( $imax_{1,k}$  and  $imax_{2,k}$ ). After that, we use HDBSCAN [7], that clusters its inputs based on their density and dynamically determines the required number of clusters. This method fits our need: we do not know how many classes there are, but we need to separate the gradients resulting from each class training data into an individual cluster. Another interesting feature of HDBSCAN is that it requires only one main parameter to build clusters: the minimum cluster size, which must be greater than or equal to 2. We use a minimum

---

**Algorithm 1:** Defending against the label-flipping attack

---

**Input:**  $K, C, BS, E, \eta, T$   
**Output:**  $W^T$ , the global model after  $T$  training rounds

- 1 A initializes  $W^0$
- 2 **for** each round  $t \in [0, T - 1]$  **do**
- 3      $m \leftarrow \max(C \cdot K, 1)$
- 4      $S \leftarrow$  random set of  $m$  peers
- 5     A sends  $W^t$  to all peers in  $S$
- 6     **for** each peer  $k \in S$  **in parallel do**
- 7          $W_k^{t+1} \leftarrow \text{PEER\_UPDATE}(k, W^t)$  // A sends  $W^t$  to each peer  $k$  who trains  $W^t$  using her data  $D_k$  locally, and sends her local update  $W_k^{t+1}$  back to the aggregator
- 8     **end**
- 9     Let  $\{\nabla_k^{L,t} | k \in S\}$  be the peers' output layer gradients at iteration  $t$
- 10    **for** each peer  $k \in S$  **do**
- 11      **for** each neuron  $i \in [1, |C|]$  **do**
- 12         Let  $\|\nabla_{i,k}^{L,t}\|$  be the magnitude of the gradients connected to the output layer neuron  $i$  of the peer  $k$  at iteration  $t$
- 13      **end**
- 14      Let  $imax_{1,k}, imax_{2,k}$  be the neurons with the highest two magnitudes in peer  $k$ 's output layer
- 15    **end**
- 16   **if**  $data$  are iid or mild non-iid **then**
- 17      Let  $\|\nabla_{i,S}^{L,t}\| = \sum_{k \in S} \|\nabla_{i,k}^{L,t}\|$  // Neuron-wise magnitude aggregation
- 18      Let  $imax_{1,S}, imax_{2,S}$  be the neurons with the highest two magnitudes in  $(\|\nabla_{1,S}^{L,t}\|, \dots, \|\nabla_{i,S}^{L,t}\|, \dots, \|\nabla_{|C|,S}^{L,t}\|)$  // Identifying potential source and target classes
- 19       $bad\_peers \leftarrow \text{FILTER\_MILD}(\{\nabla_k^{L,t} | k \in S\}, imax_{1,S}, imax_{2,S})$
- 20   **else**
- 21       $bad\_peers \leftarrow \text{FILTER\_EXTREME}(\{\nabla_{imax_{1,k}}^{L,t}, \nabla_{imax_{2,k}}^{L,t} | k \in S\})$
- 22   **end**
- 23   A aggregates  $W^{t+1} \leftarrow \text{FedAvg}(\{W_k^{t+1} | k \notin bad\_peers\})$ .
- 24 **end**

---

cluster size of 2 because, in this way, a cluster can be formed if there are at least two gradient vectors with similar input features. Otherwise, a gradient vector will be marked as an outlier and assigned the label  $-1$  if it does not belong to any formed cluster. After clustering the extracted gradients, we

---

**Procedure 1:** Filtering iid and mild non-iid updates

---

- 1 **FILTER\_MILD**( $\{\nabla_k^{L,t} | k \in S\}, imax_{1,S}, imax_{2,S}$ ):
- 2      $data \leftarrow \{\nabla_{i,k}^{L,t} | (k \in S, i \in \{imax_{1,S}, imax_{2,S}\})\}$
- 3      $cl_1, cl_2 \leftarrow \text{kmeans}(data, num\_clusters = 2)$
- 4     // Computing cluster inverse densities
- 5      $dns_1 \leftarrow \text{CLUSTER\_INVERSE\_DENSITY}(cl_1)$
- 6      $dns_2 \leftarrow \text{CLUSTER\_INVERSE\_DENSITY}(cl_2)$
- 7     // Re-weighting clusters inverse densities
- 8      $score_1 = |cl_1| / (|cl_1| + |cl_2|) * dns_1$
- 9      $score_2 = |cl_2| / (|cl_1| + |cl_2|) * dns_2$
- 10    **if**  $score_1 < score_2$  **then**
- 11       $bad\_peers \leftarrow \{k | k \in cl_1\}$
- 12    **else**
- 13       $bad\_peers \leftarrow \{k | k \in cl_2\}$
- 14   **return**  $bad\_peers$
- 15 **CLUSTER\_INVERSE\_DENSITY**( $\{\nabla_i\}_{i=1}^n$ ):
- 16   **for** each  $\nabla_i$  **do**
- 17      **for** each  $\nabla_j$  **do**
- 18         Let  $\theta_{ij}$  be the angle between  $\nabla_i$  and  $\nabla_j$
- 19         Let  $\theta_{max,i} = \max_j(\theta_{ij})$
- 20          $dns = \frac{1}{n} \sum_i \theta_{max,i}$
- 21   **return**  $dns$

---

compute the neuron-wise mean of the output layer gradients for the peers in each cluster. Then, for each mean, we compute the magnitudes of the gradient vectors corresponding to the parameters of the mean's output neurons. That is, for the mean  $\mu_j$  corresponding to the  $j$ -th cluster, we compute  $\|\nabla_{1,j}\|, \dots, \|\nabla_{i,j}\|, \dots, \|\nabla_{|C|,j}\|$  where  $\|\nabla_{i,j}\|$  is the magnitude of the gradient vector of the  $i$ -th neuron in  $\mu_j$ . After that, for each mean  $\mu_j$ , we identify the index of the neuron that has the maximum gradient vector magnitude as  $imax_{1,j}$ . In the extreme non-iid setting, each cluster corresponds to a specific class and its maximum magnitude corresponds to the neuron of that specific class, as we have discussed in the previous section. As a result, when the means of two clusters have the same  $imax_1$ , one of them could be a potential attackers' cluster. We assume that the attackers' cluster must have a smaller size than the target class cluster, and therefore we identify the smaller cluster as a potential bad cluster. Note that even if honest peers who hold *source* class examples are a minority compared to the attackers, our defense will preserve the contributions of that minority provided that the number of attackers is less than the number of peers of the *target* class. Also, we identify gradient vectors that do not belong to any cluster (labeled  $-1$  by HDBSCAN) as potential bad gradients. This ensures that even if there is only one attacker in the system, we can also detect him/her. Finally, we compute the set  $bad\_peers$  as the peers with gradients in the cluster identified as bad or labeled  $-1$ .

---

**Procedure 2:** Filtering extreme non-iid updates

---

```

1 FILTER_EXTREME( $\{\nabla_{imax_{1,k}}^{L,t}, \nabla_{imax_{2,k}}^{L,t} | k \in S\}$ ):
2    $data \leftarrow \{\nabla_{imax_{1,k}}^{L,t}, \nabla_{imax_{2,k}}^{L,t} | k \in S\}$ 
3    $\{cl_j\}_{j=1}^Z \leftarrow \text{HDBSCAN}(data, min\_cluster\_size = 2)$ 
   // Z is the number of clusters formed
4    $\{\mu_j\}_{j=1}^Z \leftarrow (\text{MEAN}(cl_1), \dots, \text{MEAN}(cl_j), \dots, \text{MEAN}(cl_Z))$ 
   //  $\mu_j$  is the neuron-wise mean of the output
      layer magnitudes of the  $j^{th}$  cluster
5   Let  $imax_{1,1}, \dots, imax_{1,j}, imax_{1,Z}$  be the indices of
      the neurons with the highest magnitude for the
      computed means //  $imax_{1,j}$  is the index of
      the neuron with the highest magnitude in  $\mu_j$ 
6   for  $i \in [1, Z]$  do
7     for  $j \in [i, Z]$  do
8       if ( $imax_{1,i} = imax_{1,j}$ ) and ( $|cl_i| > |cl_j|$ )
9         then
10        | bad_cluster  $\leftarrow cl_j$ 
11
12   bad_peers  $\leftarrow \{k | k \in S, k \in \{bad\_cluster, -1\}\}$ 
13   return bad_peers.

```

---

**Aggregating potential good updates.** After identifying the potentially bad peers, the server  $A$  computes  $\text{FedAvg}(\{W_k^{t+1} | k \notin bad\_peers\})$  to obtain the updated global model  $W^{t+1}$ .

## 6 Empirical analysis

In this section we compare the performance of our method with that of several state-of-the-art countermeasures against poisoning attacks. Our code and data are available for reproducibility purposes at <https://github.com/anonymized1/LF-Fighter>.

### 6.1 Experimental setup

We used the PyTorch framework to implement the experiments on an AMD Ryzen 5 3600 6-core CPU with 32 GB RAM, an NVIDIA GTX 1660 GPU, and Windows 10 OS.

**Data sets and models.** We tested the proposed method on three data sets (see Table 2):

- MNIST. It contains 70K handwritten digit images from 0 to 9 [22]. The images are divided into a training set (60K examples) and a testing set (10K examples). We used a two-layer convolutional neural network (CNN) with two fully connected layers on this data set.
- CIFAR10. It consists of 60K colored images of 10 different classes [20]. The data set is divided into 50K training examples and 10K testing examples. We used the ResNet18 CNN model with one fully connected layer [14] on this data set.
- IMDB. Specifically, we used the IMDB Large Movie Review data set [25] for binary sentiment classification. The data set is a collection of 50K movie reviews and

their corresponding sentiment binary labels (either positive or negative). We divided the data set into 40K training examples and 10K testing examples. We used a Bidirectional Long/Short-Term Memory (BiLSTM) model with an embedding layer that maps each word to a 100-dimensional vector. The model ends with a fully connected layer followed by a sigmoid function to produce the final predicted sentiment for an input review.

**Table 2.** Data sets and models used in the experiments

Task	Data set	# Examples	Model	# Parameters
Image classification	MNIST	70K	CNN	~22K
	CIFAR10	60K	ResNet18	~11M
Sentiment analysis	IMDB	50K	BiLSTM	~12M

**Data distribution and training.** We defined the following benchmarks by distributing the data from the data sets above among the participating peers in the following way:

- MNIST-iid. We randomly and uniformly divided the MNIST training data among 100 peers. The CNN model was trained for 200 iterations. In each iteration, the FL server asked the peers to train their models for 3 local epochs and a local batch size of 64. The participants used the cross-entropy loss function and the stochastic gradient descent (SGD) optimizer with learning rate = 0.001 and momentum = 0.9 to train their models.
- MNIST-Mild. We adopted a Dirichlet distribution [28] with a hyperparameter  $\alpha = 1$  to generate *mild non-iid* data for 100 participating peers. The training settings were the same as for MNIST-iid.
- MNIST-Extreme. We simulated an *extreme non-iid* setting with 100 peers where each peer was randomly assigned examples of a single class out of the MNIST data set. Out of the 100 peers, only 10 had examples of the source class 7 and 10 had examples of the target class 1. The training settings were the same as for MNIST-iid.
- CIFAR10-iid. We randomly and uniformly divided the CIFAR10 training data among 20 peers. The ResNet18 model was trained during 100 iterations. In each iteration, the FL server asked the 20 peers to train the model for 3 local epochs and a local batch size 32. The peers used the cross-entropy loss function and the SGD optimizer with learning rate = 0.01 and momentum = 0.9.
- CIFAR10-Mild. We adopted a Dirichlet distribution [28] with a hyperparameter  $\alpha = 1$  to generate *mild non-iid* data for 20 participating peers. The training settings were the same as for CIFAR10-iid.
- IMDB. We randomly and uniformly split the 40K training examples among 20 peers to simulate an *iid* setting. The BiLSTM was trained during 50 iterations. In each

iteration, the FL server asked the 20 peers to train the model for 1 local epoch and a local batch size of 32. The peers used the binary cross-entropy with logit loss function and the *Adam* optimizer with learning rate = 0.001.

**Attack scenarios.** In all the experiments with MNIST, the attackers flipped the examples with the source class 7 to the target class 1. In the CIFAR10 experiments, the attackers flipped the examples with the label *Dog* to *Cat* before training their local models, whereas for IMDB, the attackers flipped the examples with the label *positive* to *negative*.

In all benchmarks, the ratio of attackers ranged in {0%, 10%, 20%, 30%, 40%, 50%}. Note that in all the benchmarks, the 40% ratio of attackers corresponds to  $m' = (m/2) - 2$ , which is the theoretical upper bound of the number of attackers MKrum [4] can defend against.

**Evaluation metrics.** We used the following evaluation metrics on the test set examples for each benchmark to assess the impact of the LF attack on the learned model and the performance of the proposed method w.r.t. the state-of-the-art methods:

- *Test error (TE)*. Error resulting from the loss function used in training. The lower TE, the better.
- *Overall accuracy (All-Acc)*. Number of correct predictions divided by the total number of predictions for all the examples. The greater All-Acc, the better.
- *Source class accuracy (Src-Acc)*. Number of the source class examples correctly predicted divided by the total number of the source class examples. The greater Src-Acc, the better.
- *Attack success rate (ASR)*. Proportion of the source class examples incorrectly classified as the target class. The lower ASR, the better.
- *Coefficient of variation (CV)*. Ratio of the standard deviation  $\sigma$  to the mean  $\mu$ , that is,  $CV = \frac{\sigma}{\mu}$ . The lower CV, the better.

While TE, All-Acc, Src-Acc and ASR are used in previous works to evaluate robustness against poisoning attacks [4, 11, 38], we also use the CV metric to assess the stability of Src-Acc during training. We justify our choice of this metric in Section 6.2. An effective defense needs to simultaneously perform well in terms of TE, All-Acc, Src-Acc, ASR and CV.

## 6.2 Results

First, we report the robustness against the attack in terms of TE, All-Acc, Src-Acc and ASR for different ratios of attackers. Then, we report the stability of the source class accuracy under the LF attack. Finally, we report the runtime of our defense. In all the experiments, along with the results of our method, we also give the results of several countermeasures discussed in Section 4, including the standard FedAvg [26] aggregation method (not meant to counter poisoning attacks), the median [42], the repeated median (RMedian) [36], the

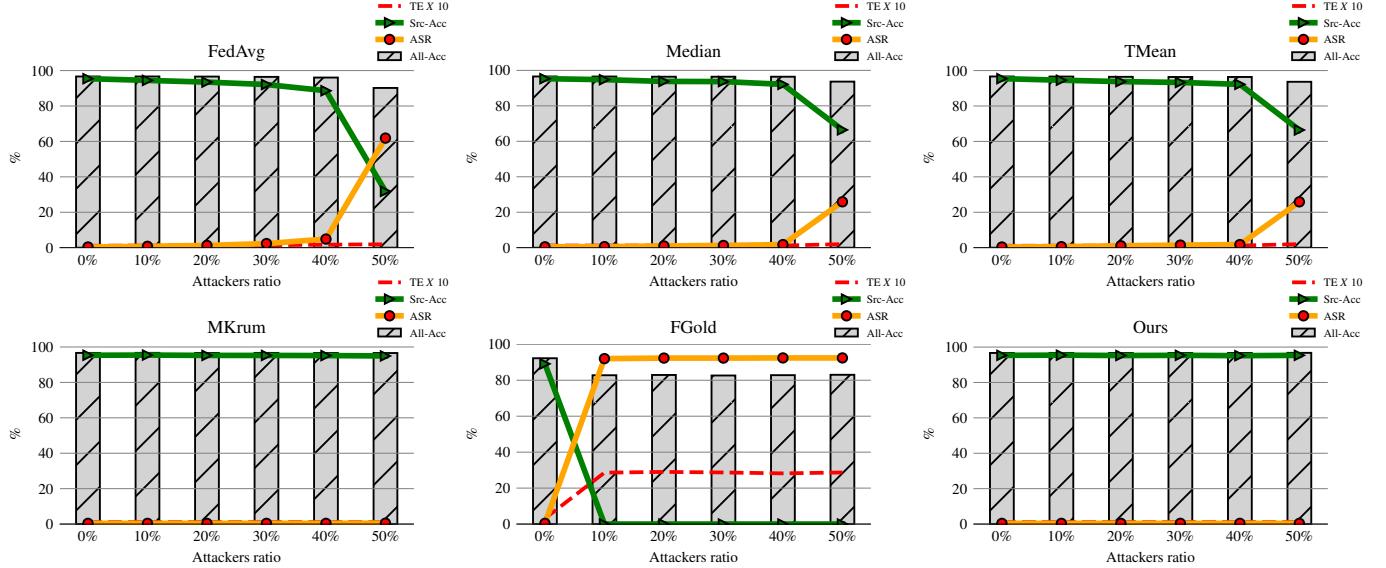
trimmed mean (TMean) [42], multi-Krum (MKrum) [4], and FoolsGold (FGold) [11].

Note that for a 50% ratio of attackers, we employ the median instead of the trimmed mean (both are equivalent in this case). Also, due to space restrictions and because the results of the repeated median were almost identical to those of the median, we do not include the former results in this section. We refer the reader to the paper repository on GitHub<sup>1</sup> for more detailed numerical results of all the benchmarks and defenses.

**Robustness against the label-flipping attack.** We evaluated the robustness of our method against the LF attack on the used benchmarks using the attack scenarios described in Section 6.1. We report the average results of the last 10 training rounds to ensure a fair comparison among methods. Note that we scaled the TE by 10 to make its results fit in the figures. Figure 7 shows the results obtained with the MNIST-iid benchmark. We can see all the defenses, except FoolsGold, perfectly defended against the attack with ratios of attackers up to 40%. However, when the attackers’ ratio was 50%, most failed to counter the attack, except MKrum and our defense, which stayed robust against the attack with all ratios. On the other hand, FoolsGold achieved the worst performance in the presence of attackers in all the metrics. Once the attackers appeared in the system, the accuracy of the source class plummeted, while the attackers had the highest success rate (close to 100%) compared to the other defenses. That happened because of the high similarity between the honest peers’ output layer gradients which FoolsGold takes into account. That led to wrongly penalizing honest peers’ updates and wrongly including some of the attackers’ bad updates in the global model. Note that, due to the small variability of the MNIST data set, each local update of an honest peer was an unbiased estimator of the mean of all the good local updates. Therefore, the coordinate-wise aggregation methods, like the median and the trimmed mean, or the update-wise aggregation methods, like MKrum, achieved such good performance in this benchmark. Another interesting note is that, although FedAvg is not meant to mitigate poisoning attacks, it achieved a good performance in this benchmark. That was also observed in [34], where the authors argued that, in some cases, FedAvg is more robust against poisoning attacks than many of the state-of-the-art countermeasures.

Figure 8 shows the results obtained with the MNIST-Mild benchmark. Although the data were non-iid, the performance was close to that in MNIST-iid. The reasons are the simplicity of the MNIST dataset and the small size of the model. It is also worth noting that FoolsGold performed better in this benchmark than in MNIST-iid. The honest peers’ gradients were more diverse in this benchmark due to their

<sup>1</sup><https://github.com/anonymized1/LF-Fighter>



**Figure 7.** Robustness against the label-flipping attack with the MNIST-iid benchmark

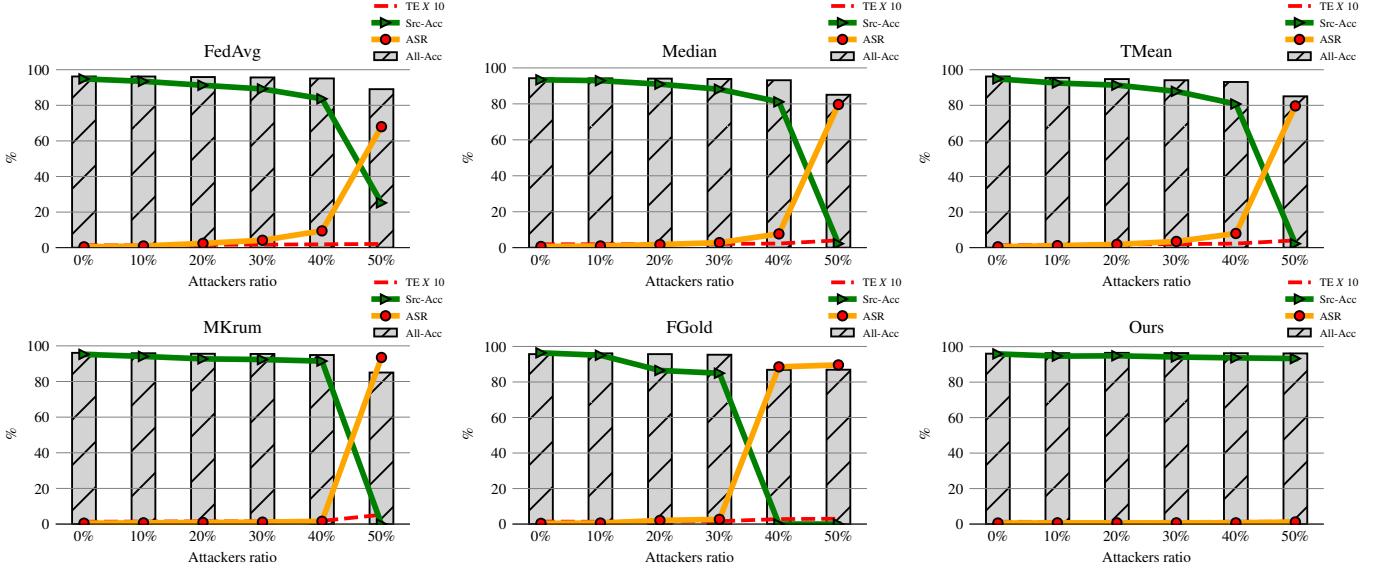
data distribution. On the other hand, our defense outperformed all the defenses and stayed robust even when the attackers' ratio reached 50%.

Figure 9 shows the results obtained on MNIST-Extreme. Our defense was effective for all considered attacker ratios. In comparison to other methods, our defense achieved simultaneous perfect performance in terms of the test error, overall accuracy, the source class (class 7) accuracy and the attack success rate. Thanks to comparing the attackers with the target class cluster, we achieved 0 false positives, and we preserved the contributions of all the other good clusters, including the source class cluster. It is also worth noting that the trimmed mean and MKrum were less affected by the attack than the other methods. This is because they considered a larger number of peer updates. On the other hand, the median and the repeated median had poor performance because the data were extremely non-iid. Thus, these methods discarded a lot of information in the global model aggregation. Also, FoolsGold performed poorly because the clusters of good updates were penalized due to their high similarity. Note that some defenses may sometimes perform well on a subset of metrics but perform poorly on the rest. For example, the median achieved an ASR of 0% in most cases, but did poorly regarding TE, All-Acc or Src-Acc. Also, FoolsGold did well for Src-Acc and ASR in some cases, but failed for TE and All-Acc. As we have mentioned, it is essential to provide good performance for all metrics, which our defense did.

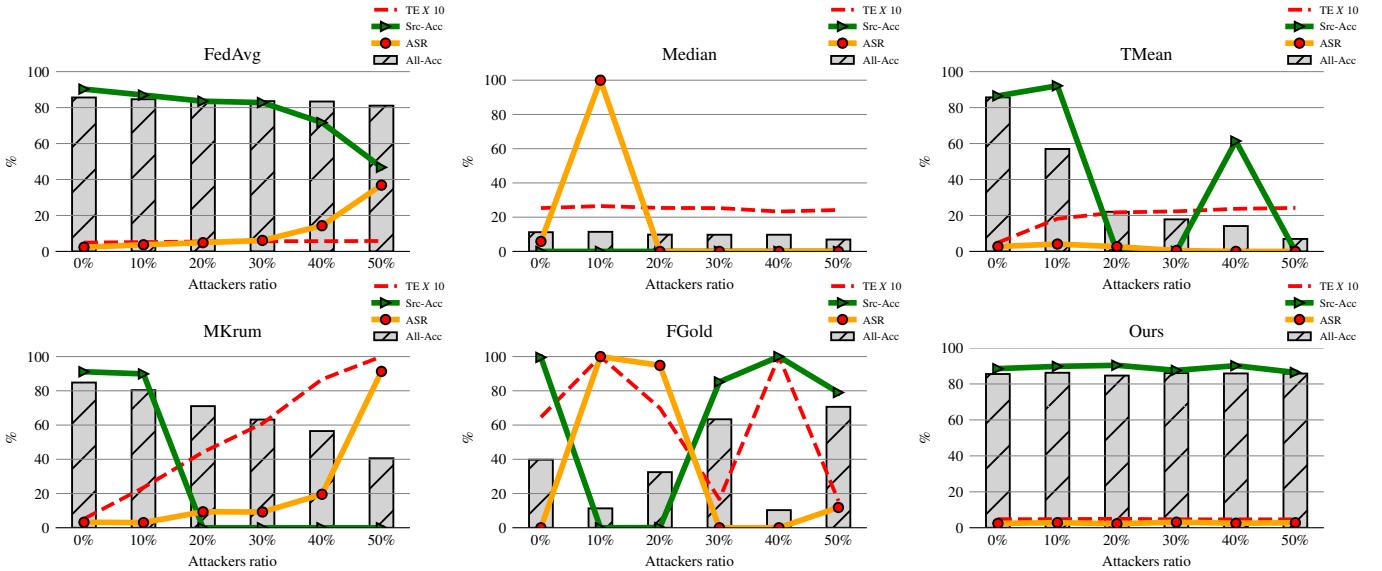
Figure 10 shows the results on the CIFAR10-iid benchmark. We can see that the performance of all methods, except FoolsGold and ours, was highly affected and degraded as the ratio of the attackers increased. This is because these methods considered the whole local updates, and the size of the used model was large (about 11M parameters). In this vast amount

of information they could not properly distinguish the good updates from the bad ones. FoolsGold effectively defended against the attack in general because it only analyzed the output layer gradients. Since the data were iid, and the CIFAR10 data set is more varied than MNIST, the attackers' output layer gradients were more similar than the honest peers' ones, and therefore, FoolsGold penalized the attackers and kept the honest peers contributions. Our defense stayed robust against the attack and achieved the best simultaneous performance for all the metrics. This is because it was able to perfectly maintain the honest peers' contributions and exclude all the attackers even when the attackers' ratio was 50%.

Figure 11 shows the results on the CIFAR10-Mild benchmark. We can see that, in this benchmark, the performance of all the methods except ours was worse due to the combined impact of the data distribution and the model dimensionality on the differentiation between the good updates and the bad ones. Although FoolsGold performed well in CIFAR10-iid, its performance substantially degraded in this benchmark. The reason for this is the combined impact on the output layer gradients of the CIFAR10 data set variability and the non-iid distribution of the data among peers, as shown in Figure 5b. This made all the output layer gradients highly diverse, and thus the gradients of the source and the target classes did not make a big difference in the distribution of the output layer gradients. On the other hand, thanks to the robust discriminative pattern we used to distinguish between updates, our defense stayed robust against the attack for all attacker ratios. In fact, it offered best simultaneous performance on all the metrics. Since our method considered only the source and target class neuron gradients (the gradients relevant to the attack) and excluded the non-relevant gradients, it was



**Figure 8.** Robustness against the label-flipping attack with the MNIST-Mild benchmark

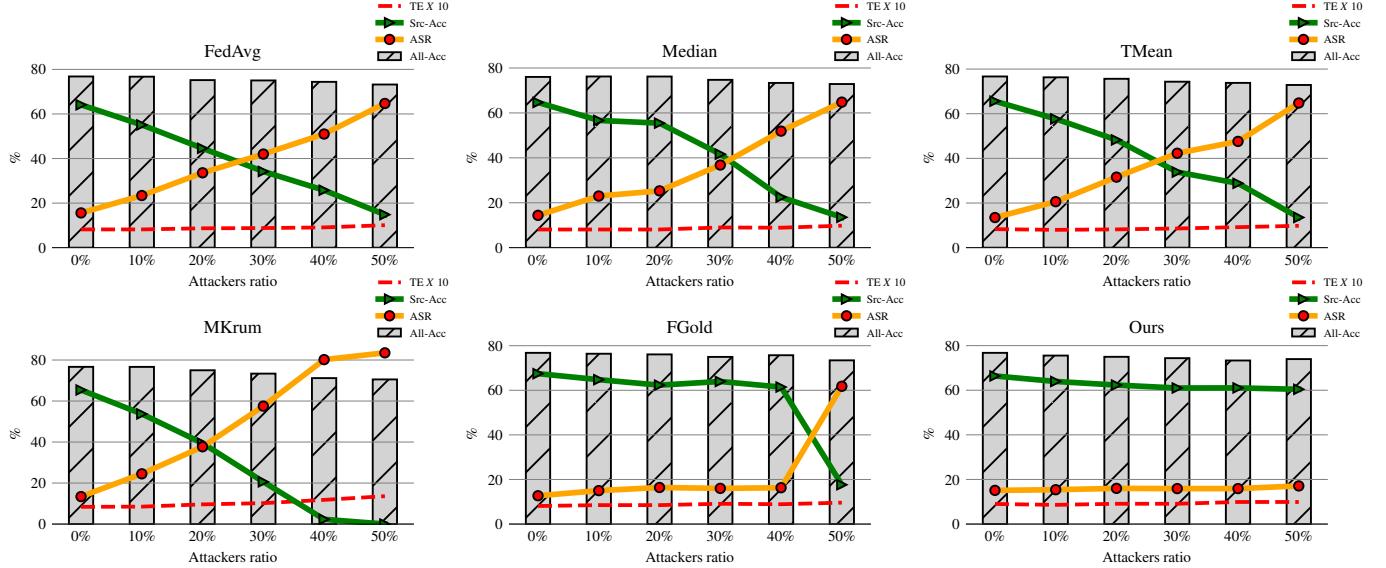


**Figure 9.** Robustness against the label-flipping attack with the MNIST-Extreme benchmark

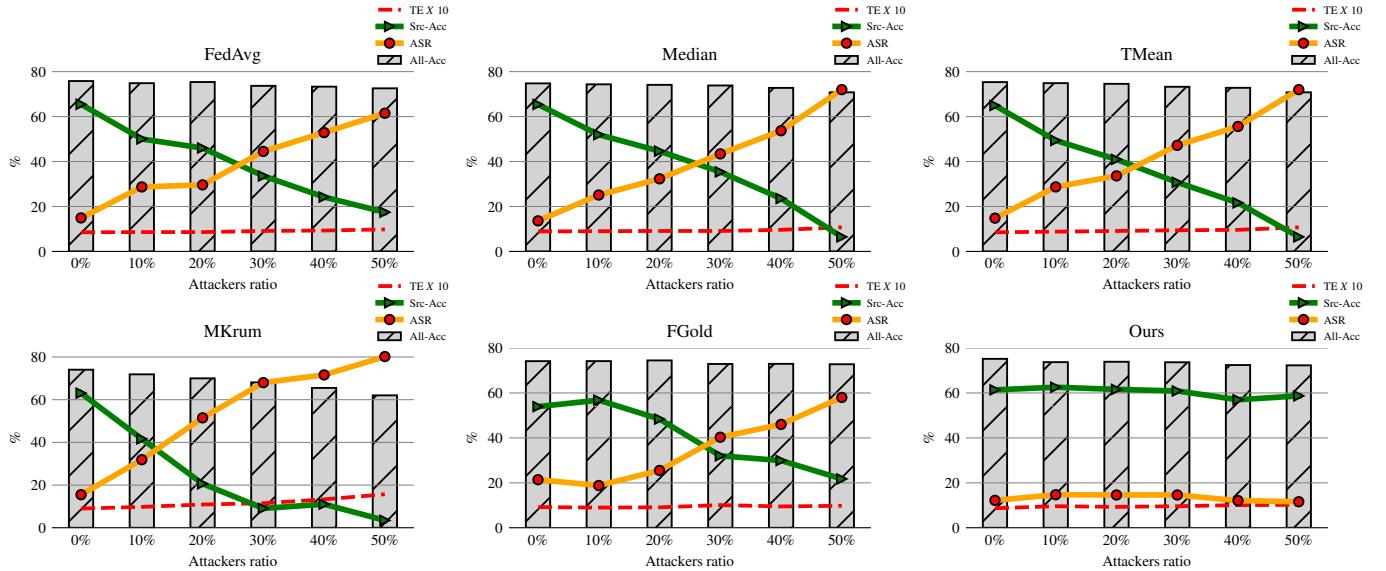
able to adequately differentiate between the good updates and the bad ones.

Figure 12 shows the results on the IMDB benchmark. Our defense and FoolsGold had almost the same performance and outperformed the other methods for all the metrics. FoolsGold performed well in this benchmark because it is its ideal setting: updates for honest peers were somewhat different due to the different reviews they gave, while updates for attackers became very close to each other because they shared the same objective. Another reason was that the number of classes in the output layer was only two. Hence, all the parameters' gradients in the output layer were relevant to the attack.

**Accuracy stability.** The stability of the global model convergence (and its accuracy in particular) during training is a problem in FL, especially when training data are non-iid [18, 24]. Furthermore, with an LF attack targeting a particular source class, the evolution of the accuracy of the source class becomes more unstable. Since an updated global model may be used after some intermediate training rounds, as in [12], this may entail degradation of the accuracy of the source class at inference time. Keeping the accuracy stable is needed to prevent such consequences. For this reason, we decided to use the CV metric to measure the stability of the source class accuracy. Table 3 shows the CV of the accuracy of the source class in the used benchmarks for the



**Figure 10.** Robustness against the label-flipping attack with the CIFAR10-iid benchmark



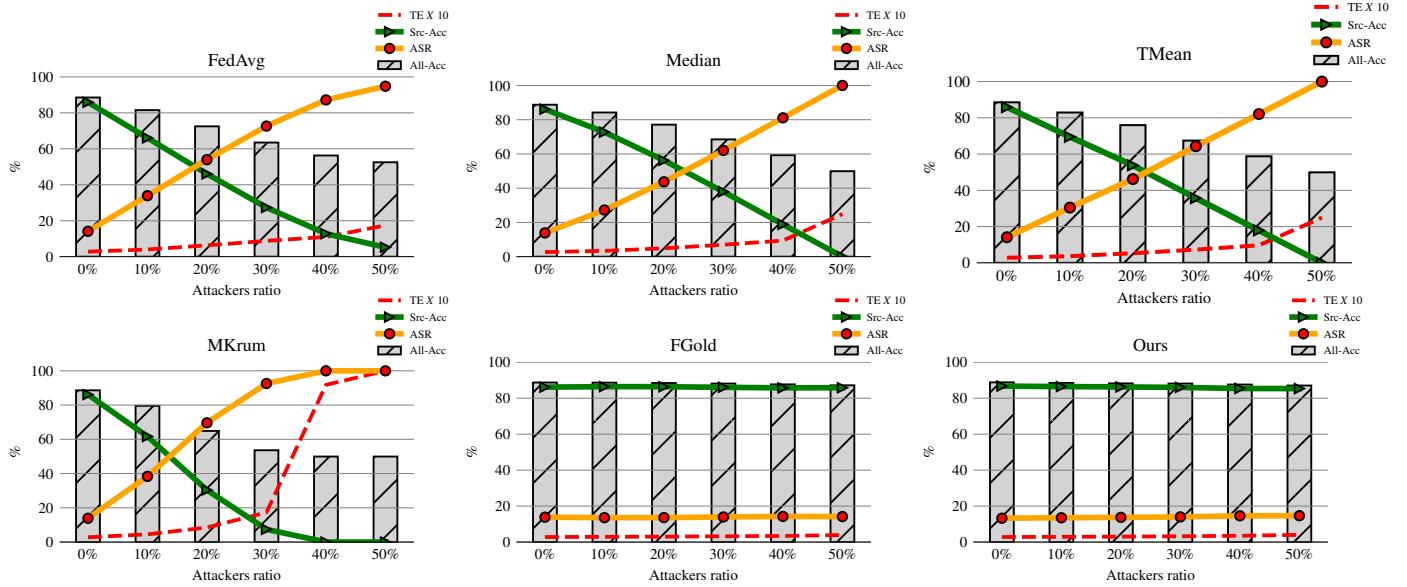
**Figure 11.** Robustness against the label-flipping attack with the CIFAR10-Mild benchmark

different defense mechanisms. We can see that our proposal outperformed the other methods in most cases, and achieved a stability very close to that of FedAvg when the attackers' ratio was 0% (*i.e.*, absence of attack). This is thanks to the perfect protection our method provided for the source class. NaN values in the table resulted from zero values of the source class accuracy in all the training rounds for those cases.

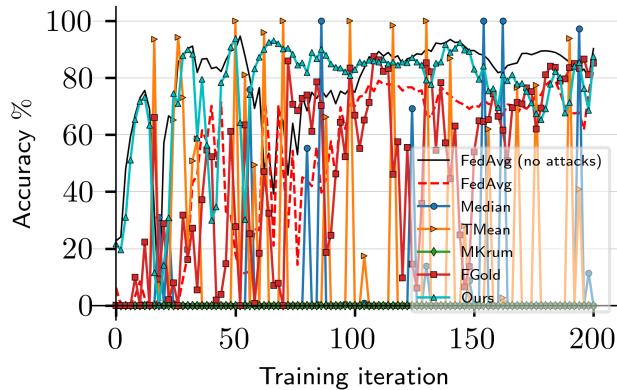
To provide a clearer picture of the effectiveness of our defense, Figure 13 shows the evolution of the accuracy of the source class as training progresses when the attacker ratio was 30% in the MNIST-Extreme, CIFAR10-iid and CIFAR10-Mild benchmarks. It is clear from the figure that the accuracy

achieved by our defense was the most similar to the accuracy of the FedAvg when no attacks were performed.

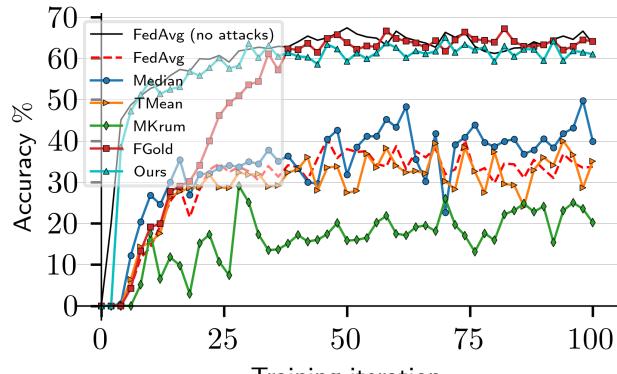
**Runtime overhead.** Finally, we measured the CPU runtime of our method and compared it with the runtime of the other methods. Figure 14 shows the total runtime in seconds (log scale) of each method during the whole training iterations. The results show that FoolsGold had the smallest runtime in all cases, excluding FedAvg, which just averages updates. The repeated median had the highest runtime due to the regression calculations it does to estimate the median points. On the other hand, the runtime of our method was similar to that of the median and the trimmed mean when the model size was small. For the large models used in the



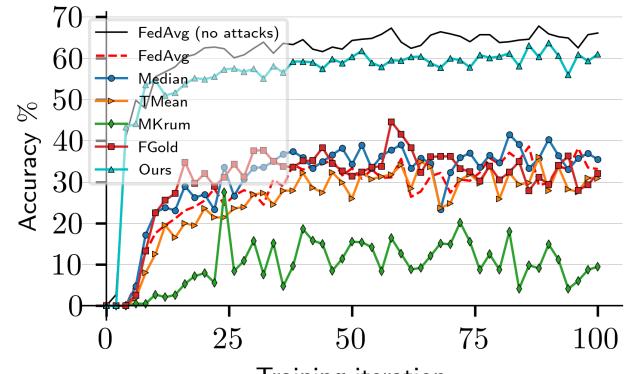
**Figure 12.** Robustness against the label-flipping attack with the IMDB benchmark



**(a)** MNIST-Extreme



**(b)** CIFAR10-iid



**(c)** CIFAR10-Mild

**Figure 13.** Evolution of the source class accuracy with 30% attackers ratio.

**Table 3.** Coefficient of variation (CV) of the source class accuracy during training for the considered benchmarks with different attacker ratios. The best figure in each column is shown in boldface.

Attackers ratio/ Method	FedAvg	Median	RMedian	TMean	MKrum	FGold	Ours	FedAvg	Median	RMedian	TMean	MKrum	FGold	Ours
<b>MNIST-iid</b>														
0%	<b>0.11</b>	<b>0.11</b>	<b>0.11</b>	<b>0.11</b>	<b>0.11</b>	0.31	<b>0.11</b>	<b>0.08</b>	0.12	0.12	<b>0.08</b>	0.09	0.09	0.10
10%	0.14	0.12	0.12	0.12	<b>0.11</b>	7.31	<b>0.11</b>	0.12	0.20	0.20	0.16	0.15	<b>0.10</b>	0.15
20%	0.19	0.14	0.14	0.14	<b>0.11</b>	6.46	<b>0.11</b>	0.19	0.28	0.28	0.25	0.24	<b>0.17</b>	<b>0.17</b>
30%	0.24	0.16	0.16	0.17	<b>0.11</b>	6.89	<b>0.11</b>	0.24	0.33	0.33	0.32	0.36	<b>0.19</b>	<b>0.19</b>
40%	0.31	0.20	0.20	0.20	<b>0.11</b>	6.39	<b>0.11</b>	0.27	0.43	0.43	0.43	0.44	1.35	<b>0.24</b>
50%	0.42	0.56	0.56	0.56	0.29	6.34	<b>0.21</b>	0.36	0.63	0.63	0.63	3.14	2.75	<b>0.27</b>
<b>MNIST-Extreme</b>														
0%	<b>0.27</b>	2.50	2.49	<b>0.27</b>	0.32	0.38	0.29	0.10	0.10	0.10	0.10	0.10	<b>0.09</b>	
10%	<b>0.33</b>	2.70	3.14	1.15	0.41	NaN	0.34	0.15	0.11	0.11	0.12	0.16	<b>0.09</b>	<b>0.09</b>
20%	0.44	2.90	2.96	1.86	14.14	1.65	<b>0.29</b>	0.22	0.15	0.15	0.16	0.28	<b>0.10</b>	<b>0.10</b>
30%	0.53	3.14	2.82	2.28	NaN	0.58	<b>0.26</b>	0.29	0.17	0.17	0.18	0.43	<b>0.10</b>	<b>0.10</b>
40%	0.77	3.07	2.88	3.43	NaN	0.43	<b>0.31</b>	0.35	0.27	0.27	0.27	NaN	<b>0.09</b>	<b>0.09</b>
50%	1.25	3.43	3.23	3.43	NaN	2.74	<b>0.37</b>	0.42	0.84	0.75	0.84	NaN	<b>0.10</b>	<b>0.10</b>
<b>CIFAR10-iid</b>														
0%	0.13	<b>0.12</b>	<b>0.12</b>	0.13	<b>0.12</b>	0.13	0.13	0.14	0.14	0.14	0.14	0.14	<b>0.13</b>	0.14
10%	0.16	0.16	0.16	0.16	0.17	0.16	<b>0.14</b>	0.19	0.18	0.19	0.17	0.20	0.17	<b>0.14</b>
20%	0.19	0.21	0.20	0.18	0.23	0.20	<b>0.13</b>	0.21	0.21	0.20	0.21	0.26	0.19	<b>0.14</b>
30%	0.23	0.23	0.24	0.23	0.32	0.27	<b>0.14</b>	0.23	0.22	0.23	0.25	0.43	0.22	<b>0.15</b>
40%	0.26	0.31	0.26	0.29	1.25	0.40	<b>0.14</b>	0.28	0.28	0.27	0.28	0.45	0.26	<b>0.15</b>
50%	0.36	0.45	0.45	0.45	6.33	0.35	<b>0.24</b>	0.38	0.42	0.43	0.42	0.54	0.40	<b>0.17</b>
<b>CIFAR10-Mild</b>														

CIFAR10 and the IMDB benchmarks, our method had the second smallest runtime after FoolsGold. In fact, the runtime incurred by our method can be viewed as very tolerable, given its effectiveness at countering the LF attack.

## 7 Conclusions and future work

In this paper, we have conducted comprehensive analyses of the label-flipping attack behavior. We have observed that the contradictory objectives of attackers and honest peers turn the parameter gradients connected to the source and target class neurons into robust discriminative features to detect the attack. **Besides, we have observed that settings with different local data distributions require different strategies to defend against the attack. Accordingly, we have presented a novel defense that uses those gradients as input features to a suitable clustering method to detect attackers.** The empirical results we report show that our defense is very effective and performs very well simultaneously regarding test error, overall accuracy, source class accuracy, and attack success rate. In fact, our defense significantly improves on the state of the art.

As future work, we plan to test and expand our method to detect other targeted attacks such as backdoor attacks.

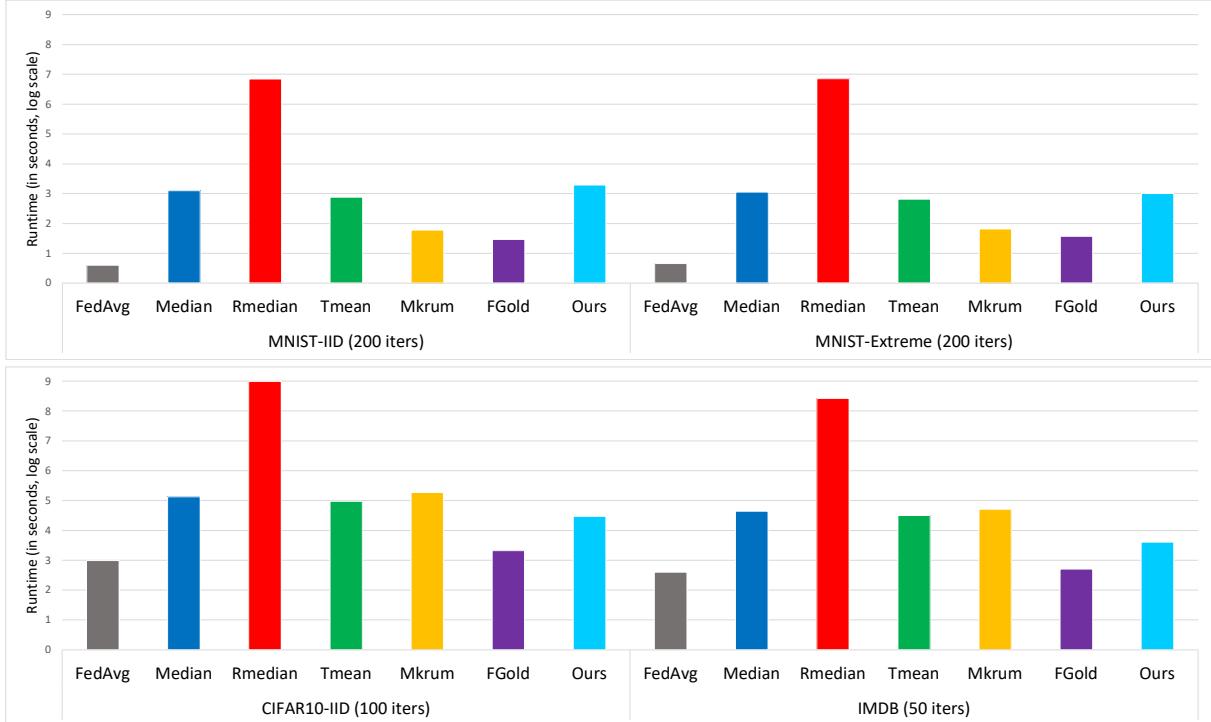
## Acknowledgments

This research was funded by the European Commission (projects H2020-871042 “SoBigData++” and H2020-101006879 “MobiDataLab”), the Government of Catalonia (ICREA Acadèmia Prizes to J.Domingo-Ferrer and D. Sánchez, and FI grant to N. Jebreel), and MCIN/AEI /10.13039/501100011033 /FEDER, UE under project PID2021-123637NB-I00 “CURLING”. The

authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and are not necessarily shared by UNESCO.

## References

- [1] Sana Awan, Bo Luo, and Fengjun Li. 2021. CONTRA: Defending against Poisoning Attacks in Federated Learning. In *In European Symposium on Research in Computer Security (ESORICS)*. Springer, 455–475.
- [2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2938–2948.
- [3] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389* (2012).
- [4] 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 International Conference on Neural Information Processing Systems*. 118–128.
- [5] Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martínez, David Sánchez, Adrian Flanagan, and Kuan Eeik Tan. 2021. Achieving security and privacy in federated learning systems: Survey, research challenges and future directions. *Engineering Applications of Artificial Intelligence* 106 (2021), 104468.
- [6] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingberman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H Brendan McMahan, et al. 2019. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046* (2019).
- [7] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 160–172.
- [8] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. 2019. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279* (2019).



**Figure 14.** Runtime overhead.

- [9] Yudong Chen, Lili Su, and Jiaming Xu. 2017. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1, 2 (2017), 1–25.
- [10] Mishra Denil, Babak Shakibi, Laurent Dinh, Marc’Aurelio Ranzato, and Nando De Freitas. 2013. Predicting parameters in deep learning. *arXiv preprint arXiv:1306.0543* (2013).
- [11] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. 2020. The Limitations of Federated Learning in Sybil Settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*. 301–316.
- [12] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Fran oise Beaufays, Sean Augenstein, Hubert Eichner, Ch o e Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).
- [13] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [15] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 19–35.
- [16] Najeeb Jebreel, Alberto Blanco-Justicia, David S nchez, and Josep Domingo-Ferrer. 2020. Efficient Detection of Byzantine Attacks in Federated Learning Using Last Layer Biases. In *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 154–165.
- [17] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aur lien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.
- [18] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*. PMLR, 5132–5143.
- [19] Jakub Kone n y, Brendan McMahan, and Daniel Ramage. 2015. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575* (2015).
- [20] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
- [22] Yann LeCun, Patrick Haffner, L on Bottou, and Yoshua Bengio. 1999. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*. Springer, 319–345.
- [23] Shenghui Li, Edith Ngai, Fanghua Ye, and Thimo Voigt. 2021. Auto-weighted Robust Federated Learning with Corrupted Data Sources. *arXiv preprint arXiv:2101.05880* (2021).
- [24] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127* (2018).
- [25] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.
- [26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera-Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*. PMLR, 1273–1282.
- [27] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep Learning-based Text Classification: A Comprehensive Review. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–40.
- [28] Thomas Minka. 2000. Estimating a Dirichlet distribution.

- [29] 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).
- [30] Milad Nasri, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 739–753.
- [31] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D Joseph, Benjamin IP Rubinstein, Udam Saini, Charles Sutton, J Doug Tygar, and Kai Xia. 2008. Exploiting machine learning to subvert your spam filter. *LEET* 8 (2008), 1–9.
- [32] Thien Duc Nguyen, Phillip Rieger, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhosseini, Ahmad-Reza Sadeghi, Thomas Schneider, et al. 2021. FLGUARD: Secure and Private Federated Learning. *arXiv preprint arXiv:2101.02281* (2021).
- [33] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
- [34] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramadge. 2021. Back to the drawing board: A critical evaluation of poisoning attacks on federated learning. *arXiv preprint arXiv:2108.10241* (2021).
- [35] Shiqi Shen, Shruti Tople, and Prateek Saxena. 2016. Auror: Defending against poisoning attacks in collaborative deep learning systems. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*. 508–519.
- [36] Andrew F Siegel. 1982. Robust regression using repeated medians. *Biometrika* 69, 1 (1982), 242–244.
- [37] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. 2017. Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3520–3532.
- [38] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security*. Springer, 480–501.
- [39] Xiaofei Wang, Yiwen Han, Chenyang Wang, Qiyang Zhao, Xu Chen, and Min Chen. 2019. In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Network* 33, 5 (2019), 156–165.
- [40] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1–3 (1987), 37–52.
- [41] Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B Giannakis. 2020. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *IEEE Transactions on Signal Processing* 68 (2020), 4583–4596.
- [42] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*. PMLR, 5650–5659.