Knowledge-Based Systems 245 (2022) 108588

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Backdoor attacks-resilient aggregation based on Robust Filtering of Outliers in federated learning for image classification

Nuria Rodríguez-Barroso [a,*], Eugenio Martínez-Cámara [a], M. Victoria Luzón [b], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Spain
[b] Department of Software Engineering, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Spain

ARTICLE INFO

Article history:
Received 30 July 2021
Received in revised form 3 February 2022
Accepted 11 March 2022
Available online 26 March 2022

Keywords:
Federated Learning
Backdoor attacks
Resilient aggregation
Robust filtering of outliers

ABSTRACT

Federated Learning is a distributed machine learning paradigm vulnerable to different kind of adversarial attacks, since its distributed nature and the inaccessibility of the data by the central server. In this work, we focus on model-poisoning backdoor attacks, because they are characterized by their stealth and effectiveness. We claim that the model updates of the clients of a federated learning setting follow a Gaussian distribution, and those ones with an outlier behavior in that distribution are likely to be adversarial clients. We propose a new federated aggregation operator called Robust Filtering of one-dimensional Outliers (RFOut-1d), which works as a resilient defensive mechanism to model-poisoning backdoor attacks. RFOut-1d is based on an univariate outlier detection method that filters out the model updates of the adversarial clients. The results on three federated image classification dataset show that RFOut-1d dissipates the impact of the backdoor attacks to almost nullifying them throughout all the learning rounds, as well as it keeps the performance of the federated learning model and it outperforms that state-of-the-art defenses against backdoor attacks.

1. Introduction

Federated learning (FL) is a nascent learning paradigm based on the distributed training of a learning model among a set of clients under the orchestration of a central server, while keeping the training data sequestered in those clients [1–3]. FL is vulnerable to adversarial attacks as machine learning systems are [4], but the distributed nature of FL and the inaccessibility of the data hinder the defense against those malicious attacks [5–7]. Since the capacity of the adversarial clients of misleading the behavior of the FL model, injecting a backdoor attack or breaking the data privacy, the development of robust and resilient FL aggregation operators to those adversarial clients is a real need [8].

The aim of an adversarial attack may be to poison the FL model [9], or to infer any properties of the training data as in the inference attacks [10]. Likewise, the attacks to the FL model may have an specific target [11], or they may only focus on hindering the performance of the FL model without any particular target, as the Byzantine attacks do [12,13]. The attacks to the FL model can be performed by corrupting the learning model (model-poisoning attacks) or the data (data-poisoning attacks). The latter ones pursue the perversely alteration of the data for provoking their misclassification, *e.g.* the dirty-label poisoning attack [14]. However, this kind of attack is mitigated by the distributed nature of FL and the usual reduced size of adversarial clients, since the aggregation of the local models dissipates the influence of the manipulated data points [5]. In contrast, the model-poisoning attacks may adulterate the FL model without a predefined target or by injecting a backdoor task, which tricks the model in favor of a specific target while keeping good performance on the main task [15]. Also, the backdoor task can be built upon the exploitation of data poisoning to alter the parameter updates. A broad review of each adversarial attack can be read in [16].

In this paper, we focus on the model-poisoning attacks based on data-poisoning and boosting of the model updates of the adversarial clients, specifically on the input-instance and two instances of pattern backdoor attacks, namely pattern-key backdoor attacks [17] and distributed backdoor attacks [18]. Since they are grounded in subtle alterations of the data on the clients, which are inaccessible, and the performance of the main task is not affected, they represent a high risk for FL. We claim that the model updates of the clients in a FL setting follow a Gaussian

[*] Corresponding author.
E-mail addresses: rbnuria@ugr.es (N. Rodríguez-Barroso), emcamara@decsai.ugr.es (E. Martínez-Cámara), luzon@ugr.es (M.V. Luzón), herrera@decsai.ugr.es (F. Herrera).

https://doi.org/10.1016/j.knosys.2022.108588
0950-7051/© 2022 Elsevier B.V. All rights reserved.

distribution, and those ones that have an outlier behavior in that distribution may be adversarial clients.

We propose the federated aggregation operator Robust Filtering of one-dimensional Outliers (RFOut-1d), which is able to filter out those clients whose model update represents an outlier in the Gaussian distribution of the model updates of the clients, thereby becoming a defense against model-poisoning attacks based on data-poisoning. The RFOut-1d federated aggregation operator performs the Standard Deviation Method on each dimension of the model clients updates for identifying univariate outliers [19], and it replaces them with the mean of the one-dimensional vector for dispensing with their participation in the aggregation. Since RFOut-1d filters out the adversarial clients, or outliers in our setting, the FL model converges faster and its performance is enhanced. Moreover, RFOut-1d can be combined with other FL defenses against backdoor attacks, as norm threshold of updates or weak differential privacy [17], enlarging its utility for preventing FL from backdoor attacks.

We evaluate the federated aggregation operator RFOut-1d on two settings of model-poisoning attacks, the input-instance and pattern backdoor attacks. The input-instance attack is based on modifying the label of some data points with a target label. Likewise, we define three difficulty levels of the pattern attack by modifying the pattern setting for both the pattern-key backdoor attack and the distributed backdoor attack. We conduct the evaluation on federated datasets, *i.e.* the distribution among the clients is predefined in the datasets. We compare RFOut-1d with FL aggregation operators like the classical FedAvg [20], and classical and state-of-the-art defenses against backdoor attacks in FL such as Median [21], Trimmed-mean [22], Norm Clipping and Weak Differential Privacy [17] and Robust Learning Rate [23].

The results show that RFOut-1d is the defense that highly minimizes the performance of the backdoor attacks in both attack settings. Moreover, RFOut-1d allows to reach the highest performance on the main task, and in some cases meets and even improves the performance of the FL model in a scenario without any adversarial client, which means that the defense of RFOut-1d does not hinder the learning of the FL model. Therefore, the consideration of adversarial clients as outliers on a Gaussian distribution allows (1) to minimize the influence of backdoor adversarial clients, and (2) to keep or even improve the performance of the FL model.

The rest of the work is organized as follows: Section 2 sums up the related works about adversarial attacks and defenses in FL; Section 3 presents the proposed federated aggregation operator based on the robust filtering of outliers, which works as a defense mechanism; Section 4 details the experimental set-up carried out; Section 5 analyzes the performance of the proposal and; finally, we expound the conclusions of the work in Section 6.

2. Adversarial attacks and defenses in federated learning

Machine learning is highly susceptible to adversarial attacks [24], and most of the defensive approaches are based on [25]: (1) game theory [4], (2) data sanitation [26] and (3) resilient and robust learning models, which assume that a fraction of the training data may be manipulated and consider it as outliers [27]. The first approach cannot be directly applied in FL, since the federated aggregation operator is usually agnostic in relation to the amount of adversarial client and to which one is adversarial. Likewise, the second approach is not feasible in a FL setting, since the data is inaccessible and kept in the clients. Hence, the most plausible defense approach is developing resilient and robust federated aggregation operators able to mitigate the malicious intention of the attacker. Accordingly, we introduce below a taxonomy of adversarial attacks in FL, some outstanding defenses against them and the backdoor attacks types and properties.

2.1. Taxonomy of adversarial attacks

According to [28], there are two types of adversarial attacks: (1) *Inference attacks* [29], which aim at inferring information from the training data; and (2) *poisoning attacks* [30], which pursue to compromise the global learning model. Concerning inference attacks, there are different types of them depending on the information being inferred. The most important ones are the property and membership inference attacks, which respectively seek to infer certain properties of the data and the membership of specific samples in the training set. Due to their nature, the defenses proposed in the literature are based on Differential Privacy [31].

Concerning model attacks, we identify two taxonomies:

1 Depending on which part of the FL schema is attacked, we differentiate between *model-poisoning* [32] and *data-poisoning attacks* [33]. In practice, both are almost equivalent, since a poisoning of the data results in a poisoned model. However, data-poisoning attacks and some of the model-poisoning attacks fail to be effective since the attack dissipates in the aggregation of many clients. For that reason, these attacks are usually combined with *model-replacement* [5] techniques, which boosts the adversarial model (or models) in order to replace the global model in the aggregation.

2 Depending on the purpose of the attack, we distinguish between *untargeted or byzantine attacks* [34], which seek to affect the model's performance, and *targeted or backdoor attacks* [5], which aim at injecting a secondary (or backdoor) task into the global model by stealth. The second ones may be more harmful, since they may be jeopardizing the integrity of the global model without been detected. Moreover, as adversarial client models optimize both the original and the adversarial task, they are also more difficult to detect in the aggregation process. Accordingly, in this paper we focus on backdoor attacks.

2.2. Defenses against adversarial attacks

The research into defenses mechanisms against adversarial attacks in FL is a booming field, and therefore many works have been published in recent years. The literature provides multiple solutions to both byzantine and backdoor attacks in classical machine learning. The vast majority of these defenses are based on data inspection methods, such as removing outliers from the training data in centralized learning [35] or, in a distributed setting, removing outliers from participant's training data or models [36,37]. In both cases, the available defenses require data inspection, which is not possible in FL. Therefore, defenses against backdoor attacks in FL must be designed ad hoc.

Regarding the state-of-the-art defenses designed to be applied in federated settings, they are based on the modification of the aggregation operator, because the attack is usually carried out by the clients. The first proposed defenses are based on a more robust aggregation of the updates such as the *Byzantine-robust aggregation rules* [38]: coordinate-wise aggregations (trimmed mean or median) [39], Krum [40] or Bulyan [41]. However, these defenses are not effective enough against backdoor attacks due to the stealthy nature of backdoor attacks [15], which stresses the need of ad hoc defenses to mitigate them.

We find some specific defenses against backdoor attacks. The most simple ones are based on the need to apply boosting, such as model-replacement, to these attacks in order to be effective. Therefore, these defenses consist of applying norm bounding of the updates (*Norm Clipping*) with the aim of weakening the effect of the most influencing clients (presumably the attacker) [17]. Moreover, these defenses can be combined with Differential Privacy [31] to get a more generalizable aggregation protection from attacks. More specific defenses are nowadays being proposed, which are based on the assumption that the attackers' updates will have different features than the rest. Some of the most influential examples are: *signSGD* [42] or Robust Learning Rate [23].

2.3. Backdoor attacks: types and properties

We subsequently introduce the backdoor attacks conducted for assessing the defensive capacity of RFOut-1d, as well as their properties. In particular, we perform three backdoor attacks based on the manipulation of the data for replacing the global model. Those attacks differ on how the data is poisoned [43], and specifically they are:

Input-instance backdoor attacks. The objective of the attack is to lead the FL model to misclassify some particular samples of the input distribution in favor of a certain target. For example, in a facial recognition system to access a room allowing access to someone (specific input) who originally did not have it.

Pattern backdoor attacks. The aim is to misclassify some modified samples according to a certain pattern in favor to a specific target. For instance, in the same facial recognition system allowing access to all people wearing purple glasses (certain pattern). The pattern can be known by all the adversarial clients or partially distributed among them, so each client fractionally knows it.

The previous backdoor attacks have a set of *hyper-parameters* or properties for configuring out their behavior. We introduce those properties that support the definition of the backdoor attack setting of the evaluation, as in [17]:

Number of backdoor tasks. In the input-instance backdoor attacks, due to the differences between clients' distributions, we consider the samples of each client as a specific backdoor task, so the number of backdoor tasks corresponds to the number of clients from which we select samples for the backdoored dataset, which we call $D_{backdoor}$. In the pattern backdoor attacks this term is not necessary, since the attack should be generalizable and it may be thus considered to be addressed by just one backdoor task (one pattern).

Number of adversarial clients. Number of clients compromised and coordinated in order to perform the backdoor tasks. The local training dataset of each adversarial client i is composed by the union of its original training dataset $D^i_{original}$ and the backdoored dataset $D_{backdoor}$. That is, $D^i_{adv} = D^i_{original} \bigcup D_{backdoor}$. In the input-instance backdoor attack, $D_{backdoor}$ will correspond to the set of samples from every backdoor task. Regarding the pattern backdoor attack, the $D_{backdoor}$ is composed of all the samples perversely altered according to a certain pattern.

Sampling of adversarial clients and frequency. The frequency of appearance of adversarial clients in the subset of clients selected for each aggregation is a key factor. In [17], the authors discuss between fixed-frequency appearance or random sampling. They conclude that the fraction of adversarial clients required for the attack to be effective is too high and unrealistic when using random sampling. Hence, we focus on the fixed-frequency attacks, in which we determine the number of adversarial clients participating in each aggregation.

3. Defense against model-poisoning backdoor attacks based on robust filtering of outliers

We consider the notations and definitions of FL as defined in [5] in order to describe the attacks discussed in this work. In particular, let G^t and L^t_i be the global model and local model of client ith at the learning round t respectively, n the total number of clients selected for each aggregation and η the server learning rate. Accordingly, the update of the global model in the learning round t is performed as follows in Eq. (1):

$$G^t = G^{t-1} + \frac{\eta}{n}\sum_{i=1}^{n}(L^t_i - G^{t-1}). \tag{1}$$

In this context, we define the backdoor attack scenario as one or several clients which are coordinated to inject a secondary or backdoor task into the global model. Typically, these attacks do not negatively affect the original task performance, which makes them harder to identify. Since the distributed character of the learning process, the high number of clients participating in each aggregation and the assumption that the proportion of adversarial clients will be significantly lower than of benign clients, the influence of the adversarial clients would be dissipated among the rest of the clients and no effective attack would take place. For that reason, we focus on model-poisoning backdoor attacks based on the model-replacement paradigm proposed in [5,15,17], which is based on boosting the influence of the adversarial attack for avoiding its dissipation among the large size of benign clients.

As we consider that only one adversarial client is selected in the learning round t, its aim is to replace the global model G^t with its backdoored model L^t_{adv}, which optimizes both original and backdoor tasks by sending to the FL server

$$\hat{L}^t_{adv} = \beta(L^t_{adv} - G^{t-1}), \tag{2}$$

where $\beta = \frac{n}{\eta}$ is the boost factor required to conduct model-replacement [5]. Then, replacing Eq. (2) in Eq. (1) we have[1]

$$G^t = G^{t-1} + \frac{\eta}{n}\frac{n}{\eta}(L^t_{adv} - G^{t-1}) + \frac{\eta}{n}\sum_{i=2}^{n}(L^t_i - G^{t-1}). \tag{3}$$

According to the definition of FL [20], eventually the FL model will converge to a solution, so we can assume that $L^t_i - G^{t-1} \approx 0$ for benign clients. Hence, we rewrite Eq. (3) as follows

$$G^t \approx G^{t-1} + \frac{\eta}{n}\frac{n}{\eta}(L^t_{adv} - G^{t-1}) = L^t_{adv}, \tag{4}$$

which replaces the global model with the *backdoored* model. If multiple adversarial clients participate in the same learning round, we assume that they can coordinate for the attack by dividing the boosting factor between the attackers. In the rest of the paper, we consider $\eta = 1$.

We consider the attack scenario described below, in which the model updates of benign clients minimizes the global task loss, while the model updates of adversarial clients optimize the global and backdoor task loss. We base our proposal on the following two assumptions:

1 The model updates of the clients follow a Gaussian distribution from a certain learning round, since the global aggregated model tends to converge to a common solution. This is intuitively proven based on the Central Limit Theorem [44], which states that the sum of independent random variables closely approaches to a Gaussian distribution. Let the clients local weight's distributions be each

[1] For the sake of clarity, we assume that the adversarial client is client 1.

of the random variables, then, linear combinations of them approach closely to a Gaussian distribution. Therefore, aggregation over aggregation, the result will converge to a Gaussian distribution. In particular, the data distribution for each of the dimensions of the updates converges to an univariate Gaussian distribution.

2 Since the model update of adversarial clients has a twofold target, we assume that it represents an outlier in the distribution of client updates for a specific learning round.

Regarding the previous assumptions, we propose the **RFOut-1d** (**R**obust **F**iltering of **1-d**imensional **Out**liers) federated aggregation operator based on filtering out the outliers in the distribution of client model updates with the objective of producing a more robust aggregation in each learning round t. Since the high dimensionality of the updates (usually from neural networks), and with the aim of avoiding the loss of information by applying dimensionality reduction techniques, we perform an univariate anomaly detection for each dimension of the model updates. Therefore, for each of dimension $i \in \{1, \dots, m\}$, where m is the dimension of the vectors of the model updates, we consider the vector formed by the local model update of each client in that dimension $L_i = (L_1^t[i], L_2^t[i], \dots, L_n^t[i])$, where n is the number of clients participating in the aggregation, and we apply the Standard Deviation Method for identifying univariate outliers in Gaussian distributions. Hence, it filters out those that verify that the difference between the value and the mean is greater or equal than δ times the standard deviation. Formally, we replace by μ_i those that verify

$$abs(L_j^t[i] - \mu_i) \geq \delta\sigma_i, \tag{5}$$

where μ_i and σ_i are the mean and standard deviation of L_i, respectively, $L_j^t[i]$ is the parameter of the dimension i of the model update at the round of learning t of the client j and $\delta = 3$ according to an experimental result of [19]. We use the mean estimator since, as the model updates of the clients are subsequently aggregated, it filters out the participation of the outliers in the aggregation.

At the end, the federated aggregation RFOut-1d consists of the 1-dimensional mean of the non-filtered out parameters. Formally, the resulting aggregated model G_t in each dimension i of the parameters is

$$G_t[i] = \frac{1}{n}\sum_{i=1}^{n} \hat{L}_j^t[i] \quad \forall i \in \{1, \dots, m\}, \quad \text{where} \tag{6}$$

$$\hat{L}_j^t[i] = \begin{cases} \mu_i & \text{if } abs(L_j^t[i] - \mu_i) \geq \delta\sigma_i \\ L_j^t[i], & \text{otherwise,} \end{cases} \forall j \in \{1, \dots, n\} \tag{7}$$

where $\hat{L}_j^t[i]$ the resulting vector after applying Eq. (5) criteria to $L_j^t[i]$. Algorithm 1 depicts the proposed aggregation operator.

Note that RFOut-1d, in addition to filtering out those clients that are presumably attackers, optimizes the learning process by favoring a faster convergence towards a common solution. Moreover, it can be combined with other aggregation mechanisms proposed as defenses, such as norm threshold of updates or weak Differential Privacy [17].

4. Experimental set-up

We subsequently detail the experimental framework for assessing the RFOut-1d federated aggregation operator. We describe the datasets used in the evaluation, the configuration of the backdoor attacks and the evaluation measures. We follow the guidelines of [45] for conducting the experiments.[2]

Algorithm 1 RFOut-1d

```
Input: local updates $\{L_1^t, L_2^t, \dots, L_n^t\}$
$num\_dimensions = length(L_1^t)$
Initialize $G^t$
$\delta = 3$
for $i = 0$ to $num\_dimensions$ do
    $\hat{L}_i = (L_1^t[i], L_2^t[i], \dots, L_n^t[i])$
    $\mu_i = mean(\hat{L}_i)$
    $\sigma_i = std(\hat{L}_i)$
    for $j = 1$ to $n$ do
        if $abs(L_j[i] - \mu_i) \geq \delta\sigma_i$ then
            $L_j[i] \leftarrow \mu_i$
        end if
    end for
    $G_t[i] = mean(\hat{L}_i)$
end for
Return $G_t$
```

Table 1
Description of the FEMNIST, CelebA-S and CelebA-A datasets.

	FEMNIST	CelebA-S	CelebA-A
Clients	3579	1878	1878
k	8	30	30
Number of labels	10	2	2
Training samples	240000	56364	56364
Samples per client (mean)	67.05	30.01	30.01
Samples per client (std)	11.17	0.19	0.19
Testing samples	40000	19962	19962

4.1. Datasets

The few availability of non-simulated federated datasets is one of the difficulties for evaluating FL models. It is possible to use classical machine learning datasets and distribute them among clients according to different data distributions. However, although the non-IID character of data distribution can be simulated [8], it is quite complex to simulate the customization of data among clients, so that they represent their individual features. For that reason, we decided to use datasets that are by definition federated. We focus on the following image classification datasets included in the LEAF benchmark:

1 *Digits FEMNIST*:[3] The digits dataset of the federated version of EMNIST, where each client corresponds to an original writer.

2 *CelebA*:[4] An image classification dataset composed by famous face images with 40 binary attributes annotations per image, where we associate each famous with a client. We use it as a binary image classification dataset, selecting a specific attribute as target, in particular, *Smiling* (*CelebA-S*) and *Attractive* (*CelebA-A*).

The use of federated datasets may result in some clients with insufficient amount of data. Accordingly, we set the minimum number of samples per client k and discard the clients that do not satisfy this condition. For *CelebA* datasets, we use $k = 30$, specified as the best option in [46], and for *FEMNIST* we set $k = 8$, as it is the minimum number of samples per client. Table 1 shows the statistics per dataset.

[2] We provide the source code of RFOut-1d at this GitHubRepository.

[3] https://www.nist.gov/itl/products-and-services/emnist-dataset.

[4] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

Table 2
Definition of input-instance backdoor attacks set-up for the FEMNIST, CelebA-S and CelebA-A datasets.

	FEMNIST	CelebA-S	CelebA-A
Backdoor tasks	30	30	10
$\lvert D_{backdoor} \rvert$	213	247	228
Adversarial clients	11	20	15
Frequency of attack	1	1	1
Origin label	7	No	No
Target label	1	Yes	Yes

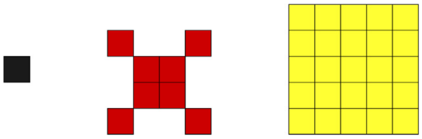


Fig. 1. Representation of the pattern-key employed. From left to right: the 1-pixel pattern, the 8-pixel pattern (a red cross of length 4) and the 25-pixel pattern (a 5 × 5 yellow square).

4.2. Backdoor attacks set-up

According to the definition of backdoor attacks, the design of such attacks has a wide range of options as the backdoor task depends on the aim of the injected task. We define an input-instance and the two pattern backdoor attacks settings to assess RFOut-1d in each dataset.

4.2.1. Input-instance backdoor attacks set-up

We set a target label and a set of samples ($D_{backdoor}$) from clients which belong to another class (original label). The attack consists in classifying the highest amount of these samples with the target label without modifying any sample. Due to the particularity of each client, we set that the number of backdoor tasks corresponds to the number of clients from whom samples have been taken for the backdoored dataset $D_{backdoor}$. We set the number of adversarial clients as the number of clients who have the backdoored dataset among their data and the frequency of attack. Based on these parameters, we define these attacks in Table 2.

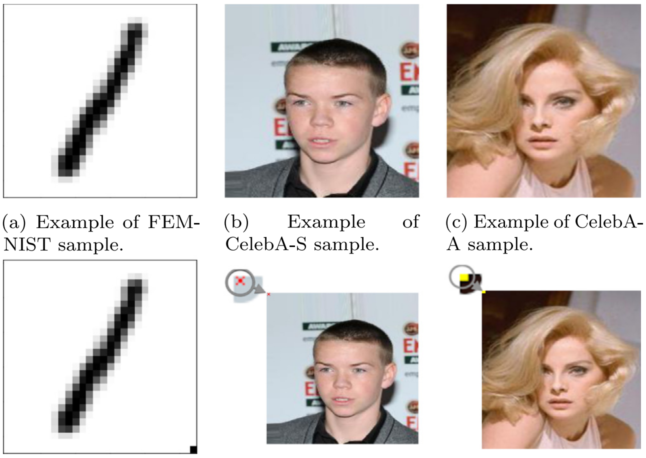
4.2.2. Pattern backdoor attacks set-up

We evaluate RFOut-1d in two types of pattern backdoor attack: (1) Pattern-key backdoor attack, in which all the clients know the complete pattern and use it in their training process and (2) Distributed backdoor attack [18], in which each client knows the pattern partially and the aim is to coordinate to inject the complete pattern.

Pattern-key backdoor attacks. We set a target label and a pattern-key. Thus, the attack consists in classifying any sample poisoned with the pattern-key as the target label. In this case, the number of backdoor tasks corresponds to the number of adversarial clients, because only the adversarial clients poison some of their samples with the pattern-key. In order to show that the behavior of RFOut-1d is agnostic of the pattern-key, we use three patterns of different levels of difficulty expressed in numbers of pixels (see Fig. 1): (1) one single black pixel, (2) a red cross of length 4 and (3) a yellow square of side 5 × 5. Analogously, we define these attacks in Table 3, and we show the patterns used for poisoning implemented in Fig. 2. When the pattern-key is small in comparison with the original image we add a zoom of the pattern in the corner.

Table 3
Definition of pattern-key backdoor attacks set-up for the FEMNIST, CelebA-S and CelebA-A datasets.

	FEMNIST	CelebA-S	CelebA-A
Adversarial clients	30	15	15
Frequency of attack	1	1	1
Target label	0	Yes	Yes
Pixels of the pattern	1	8	25



(a) Example of FEMNIST sample.

(b) Example of CelebA-S sample.

(c) Example of CelebA-A sample.

(d) Backdoored FEMNIST sample (1-pixel pattern).

(e) Backdoored CelebA-S sample (8-pixel pattern).

(f) Backdoored CelebA-A sample (25-pixel pattern).

Fig. 2. Examples of original (a, b and c) and backdoored samples (d, e and f) of each dataset.

Table 4
Definition of distributed backdoor attacks set-up for the FEMNIST, CelebA-S and CelebA-A datasets.

	FEMNIST	CelebA-S	CelebA-A
Adversarial clients	4	2	2
Frequency of attack	4	2	2
Target label	0	Yes	Yes
Pixels of the complete pattern	4	10	10
Pixels of each partial pattern	1	5	5

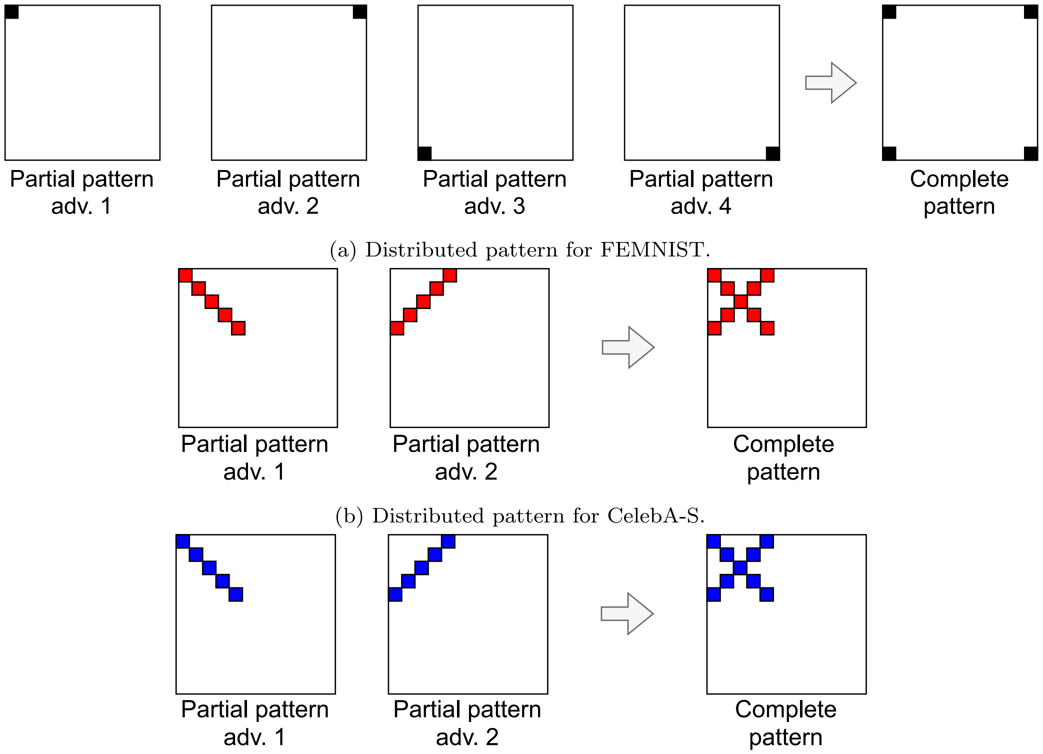
Distributed backdoor attack. We set the target label, the complete pattern and the partial pattern of each adversarial client. Clearly, the attack consists in classifying each sample poisoned with the complete pattern as the target label, not the partial ones. For that reason, in each aggregation participates one adversarial client from each partial pattern, thus involving multiple adversarial clients in each learning round. In order to show that the behavior of RFOut-1d is agnostic of the pattern, we use different patterns for each database (see Fig. 3 and Table 4):

1 Black corners. Four single black pixels distributed among the four corners of the image for FEMNIST. We distribute the pattern by setting 4 adversarial clients and assigning each corner to one of them.

2 Monocolor cross. A cross of length 5 in the upper left corner red for CelebA-S and blue for CelebA-A. We distribute the pattern by setting 2 adversarial clients and assigning each diagonal of the cross to one of them.

4.3. Evaluations metrics and baselines

The task of defending against backdoor attacks is a twofold task, and its evaluation thus requires of measuring the prevention



(a) Distributed pattern for FEMNIST.

(b) Distributed pattern for CelebA-S.

(c) Distributed pattern for CelebA-A.

Fig. 3. Distributed pattern for CelebA-A.

against the attack and the performance of the resulting model in the original task. The aim of the defense mechanism is to reduce the effects of the attack as much as possible without compromising the performance of the model in the original task. We consider two test datasets:

- *Original task test*. The original test of the dataset used for measuring the performance in terms of accuracy in the original task.
- *Backdoor task test*. Dataset which represents the attack in order to measure the performance in terms of accuracy in the backdoor task. Regarding the input-instance backdoor attacks, we consider the backdoored dataset $D_{backdoor}$ as in [17]. Concerning the pattern backdoor attacks, we consider two test datasets [43]: (1) *Backdoor task test* as in the input-instance backdoor attack situation to measure the effectiveness of the attack; and (2) *Global backdoor task test*, consisting of the test instances not originally belonging to the target class, but poisoned using the pattern in order to measure the capability of generalization of the attack.

Since the results can be highly heterogeneous in each of the learning rounds depending on the defense mechanism and in order to show robust results, we use the average of each of these measures throughout the last ten learning rounds.

We compare RFOut-1d with the following federated aggregation operators and backdoor defense mechanisms, which represent the classical baselines and the state-of-the-art in defenses against backdoor attacks:

1 *Federated Averaging (FedAvg)* [20]. It is based on the (weighted) averaging of the local models. We use this aggregation operator as the simplest baseline due to it represents the *no-defense* situation.

2 *Median* [21]. It is one of the Byzantine-robust aggregation rules which is based on replacing the mean with the median in the aggregation method. We use it as a baseline, due to the higher robustness of the median with respect to the mean in the presence of extreme values.

3 *Trimmed-mean* [22]. It represents another Byzantine-robust aggregation rule. It relies on using a more robust version of the mean that consists in eliminating a fixed percentage of extreme values both below and above the data distribution.

4 *Norm Clipping of updates* [17]. Since model-poisoning backdoor attacks produce updates with large norms because of the boosting factor, norm clipping of updates is widely used as a simple defense mechanism. It consists in clipping the update by dividing it with the appropriate scalar if it exceeds a fixed threshold M, as in Eq. (8), where $\Delta L_i^t = L_i^{t+1} - G^t$.

$$G^{t+1} = G^t + \frac{\eta}{n}\sum_{i=1}^{n}\frac{\Delta L_i^t}{\max(1, \|\Delta L_i^t\|_2/\mathbf{M})} \tag{8}$$

5 *Weak Differential Privacy (WDP)* [17]. This defense is based on Differential Privacy [31], which is commonly used to defend against backdoor attacks [47]. This mechanism consists of applying norm techniques combined with a little amount of Gaussian noise as a function of σ according to Eq. (9).

$$G^{t+1} = G^t + \frac{\eta}{n}\sum_{i=1}^{n}\frac{\Delta L_i^t}{\max(1, \|\Delta L_i^t\|_2/\mathbf{M})} + \mathcal{N}(0, \frac{\sigma\mathbf{M}}{n}) \tag{9}$$

6 *Robust Learning Rate (RLR)* [23]. They determine the direction of the update, for each dimension, in form of signs

Table 5
Parameters used in our experiments according to the parameters recommended by the authors.

	M	σ	θ
Norm Clipping	3/1	0	–
WDP	3/1	0.0025	–
RLR	0.5/1	0.0001	7
RFOut-1d	0.5/1	0.0001	–

of the updates using a threshold parameter θ. Hence, if the sum of the signs of the updates is less than θ, they change the direction of the update by multiplying by -1. They assert that this defense can be combined with the two previous ones, by means applying the norm clipping and noise addition specified in Eq. (9) to the modified models' updates, producing a better performance.

We did not compare RFOut-1d with the defenses Krum [40] and Bulyan [41], because they are design against Byzantine attacks, while RFOut-1d works against backdoor attacks.

We use the configuration values specified in [17,23] in our experiments. In addition, we evaluate the use of *norm clipping* and *noise addition* in RFOut-1d following Eq. (9) with the same parameter values as RLR. Table 5 shows these parameters, where M is the threshold for the updates norm, σ the Gaussian noise parameter and θ the threshold for RLR.

Since the main aim of this work is to propose a robust federated aggregation operator to defense against backdoor attacks, we use an standard CNN-based image classification model composed of two CNN layers followed by its corresponding max-pooling layers, a dense layer and the output layer with a softmax activation function. In particular, we use the models and the hyperparameters included in the LEAF[5] benchmark for each dataset. All details concerning hyperparameters, number of epochs or batch size can be found in the GitHub repository.

5. Analysis of the results

We evaluate RFOut-1d on the datasets described in Section 4.1, and in the two backdoor attack settings described in Section 4.2 during 100 rounds of learning. Subsequently, we expose the assessment in each backdoor attack, and we analyze the capacity of RFOut-1d of enhancing the FL model convergence.

5.1. Analysis of the performance against input-instance backdoor attacks

Table 6 shows that RFOut-1d outperforms all the baselines in the twofold goal of minimizing the backdoor task performance and maximizing the performance of the original task (image classification), which means that filtering out the parameters that represent outliers in the distribution of updates mitigates these attacks.

Generally, as we use a more complex defense, the results obtained improve notably in favor of the defense. In particular, RLR is the most powerful baseline (especially the norm clipping and noise version), namely as far as the accuracy of the original task is concerned.

The highest result in all test sets is always achieved by RFOut-1d. On the one hand, the ability to mitigate the attack is shown, achieving a null effect of the attack (0.0 of backdoor accuracy) in two of the three datasets. On the other hand, we show that it does not compromise the performance in the original task, even improving the result of the task without attack in the case of *FEMNIST* and *CelebA-A*, which proves that it also filters out low-value information. This suggest that it may not be only filtering out adversarial clients, but those clients who have such poor training that they confuse the model rather than contributing to its convergence towards a global solution.

Regarding to the combination of the defenses with *norm clipping* and *noise*, both RLR and RFOut-1d can be combined. However, for RLR it seems to be a necessity as the results improve markedly while RFOut-1d obtains strong results on its own, which confirms the robustness of our proposal.

Therefore, the results show that RFOut-1d is a robust federated aggregation operator against input-instance backdoor attack, and it does not need any additional operation to preserve the FL model from this kind of adversarial attack.

5.2. Analysis of the performance against pattern backdoor attacks

We analyze the behavior of RFOut-1d in two different pattern backdoor attacks: (1) the analysis of the performance of the pattern-key backdoor attacks, in which only one adversarial client participates in each aggregation process; and (2) the analysis of the distributed backdoor attacks, in which participate as many clients as different partial patterns defined in each aggregation process.

Pattern-key backdoor attacks. Analogously, the results in Tables 7, 8, 9 show the higher performance of RFOut-1d compared to the baselines in FEMNIST, CelebA-A and CelebA-S respectively, which proves that our claim is also confirmed for pattern-key backdoor attacks and, moreover, for patterns of different level of difficulty.

If we compare the effectiveness of these pattern-key attacks without any defense (*FedAvg*) with the same condition as in the input-instance attacks, we find that the first ones are, generally, more effective. This is due to the alteration of images with a pattern is a more sophisticated attack, and it allows to reach its aims with a higher success than the input-instance backdoor attack.

Despite being more powerful attacks, the defenses, the baselines and RFOut-1d, show similar behavior, improving as we use a more complex defense. In particular, the defense that outperforms in both tasks of maximizing the performance of the global task and minimizing the performance of the backdoor task is, again, RFOut-1d. Therefore, RFOut-1d outperforms all the baselines in the target of defending the FL model against the pattern-key backdoor task, which means that our claim holds in this kind of backdoor attack. In this case, it also outperforms the results without any attack, which confirms its proper performance as a federated aggregation operator even without the presence of adversarial clients.

Distributed backdoor attacks. The results of Tables 10–12 show the outperforming of RFOut-1d compared with the baselines in FEMNIST, CelebA-S and CelebA-A respectively. It is worth mentioning that the backdoor set is less significant in this case as it represents the effectiveness of the partial patterns, while we are interested in the effectiveness of the complete pattern.

Regarding the effectiveness of the attack, we find that distributed backdoor attack has achieved a lower performance on the backdoor task in FEMNIST than the pattern-key backdoor attack. However, the behavior in both partitions of CelebA is comparable. We attribute this phenomenon to the fact that the distributed backdoor attack is more complicated to be successful, being too challenging to carry it out in a multi-class problem as FEMNIST. However, even in this case the presence of the defenses is notable, significantly diminishing the effectiveness of the backdoor attacks in test.

[5] https://github.com/TalwalkarLab/leaf.

Table 6
Mean results for the input-instance backdoor attack in terms of accuracy. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	*σ*	FEMNIST		CelebA-S		CelebA-A	
			Original	Backdoor	Original	Backdoor	Original	Backdoor
No attack	0	0	0.9657	–	**0.7900**	–	0.7973	–
FedAvg	0	0	0.8661	0.8230	0.3630	0.9738	0.5140	0.5194
Median	0	0	0.9448	0.0306	0.7881	0.0457	0.7961	0.0152
Trimmed-mean	0	0	0.9526	0.0256	0.7852	0.0423	0.7961	0.0221
NormClip	3	0	0.9606	0.6373	0.6852	0.1431	0.6078	0.2558
WDP	3	0.0025	0.9374	0.1578	0.7204	0.1195	0.6119	0.2399
RLR	0	0	0.8404	0.0288	0.6539	0.0457	0.7877	0.0451
RLR†	0.5/0.5/1	0.0001	0.9546	0.0128	0.7852	0.0388	0.7934	0.0043
RFOut-1d	0	0	0.9629	**0.0048**	0.7883	0.0046	0.7973	**0.0**
RFOut-1d†	0.5/0.5/1	0.0001	**0.9670**	0.0054	0.7892	**0.0**	**0.7975**	**0.0**

Table 7
Mean results for the pattern-key backdoor attack in terms of accuracy in FEMNIST. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	*σ*	FEMNIST		
			Original	Backdoor	Test
No attack	0	0	0.9657	–	–
FedAvg	0	0	0.9741	1.0	1.0
Median	0	0	0.9540	0.0091	0.0154
Trimmed-mean	0	0	0.9664	0.0114	0.0148
NormClip	1	0	0.9687	0.0553	0.0538
WDP	1	0.0025	0.9357	0.0938	0.0175
RLR	0	0	0.9039	0.0407	0.0575
RLR†	0.5/1	0.0001	0.9265	0.0089	0.0085
RFOut-1d	0	0	0.9741	**0.0043**	0.0072
RFOut-1d†	0.5/1	0.0001	**0.9753**	0.0059	**0.0051**

Table 8
Mean results for the pattern-key backdoor attack in terms of accuracy in CelebA-S. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	*σ*	CelebA-S		
			Original	Backdoor	Test
No attack	0	0	0.7900	–	–
FedAvg	0	0	0.6858	1.0	0.9999
Median	0	0	0.6978	0.0678	0.0532
Trimmed-mean	0	0	0.7013	0.0521	0.0654
NormClip	1	0	0.6798	0.1433	0.1647
WDP	1	0.0025	0.7413	0.0538	0.0743
RLR	0	0	0.7132	0.0574	0.0469
RLR†	0.5/1	0.0001	0.7714	0.0205	0.0316
RFOut-1d	0	0	**0.7900**	**0.0**	**0.0**
RFOut-1d†	0.5/1	0.0001	0.7896	**0.0**	0.0010

Table 9
Mean results for the pattern-key backdoor attack in terms of accuracy in CelebA-A. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	*σ*	CelebA-A		
			Original	Backdoor	Test
No attack	0	0	0.7973	–	–
FedAvg	0	0	0.7375	1.0	0.99
Median	0	0	0.7452	0.0163	0.0189
Trimmed-mean	0	0	0.7498	0.0092	0.0101
NormClip	1	0	0.7126	0.1433	0.1316
WDP	1	0.0025	0.6609	0.1440	0.1707
RLR	0	0	0.6657	0.0280	0.0286
RLR†	0.5/1	0.0001	0.7923	0.0031	0.0016
RFOut-1d	0	0	**0.7967**	**0.0023**	**0.0015**
RFOut-1d†	0.5/1	0.0001	0.7874	0.0054	0.0124

Table 10
Mean results for the distributed backdoor attack in terms of accuracy in FEMNIST. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	*σ*	FEMNIST		
			Original	Backdoor	Test
No attack	0	0	0.9657	–	–
FedAvg	0	0	0.9678	0.8556	0.1649
Median	0	0	0.9437	0.0114	0.0053
Trimmed-mean	0	0	0.9649	0.0102	0.0046
NormClip	1	0	0.9731	0.3289	0.0526
WDP	1	0.0025	0.9729	0.3342	0.0211
RLR	0	0	0.9518	0.7821	0.0263
RLR†	0.5/1	0.0001	0.9614	0.0107	0.0062
RFOut-1d	0	0	0.9721	0.2130	0.0089
RFOut-1d†	0.5/1	0.0001	**0.9737**	**0.0000**	**0.0032**

Concerning the evaluation of the different defenses, both the proposal and baselines, the results further confirm the satisfactory performance of RFOut-1d in backdoor attacks. To conclude, it is worthy noting that RLR, which in the evaluation of the input-instance and pattern-key backdoor attacks had achieved quite successful results, is outperformed by the other simpler baselines in both partitions of CelebA. It shows that it may not be useful for this type of distributed backdoor attacks, or at least with the parameters used in the experimentation.

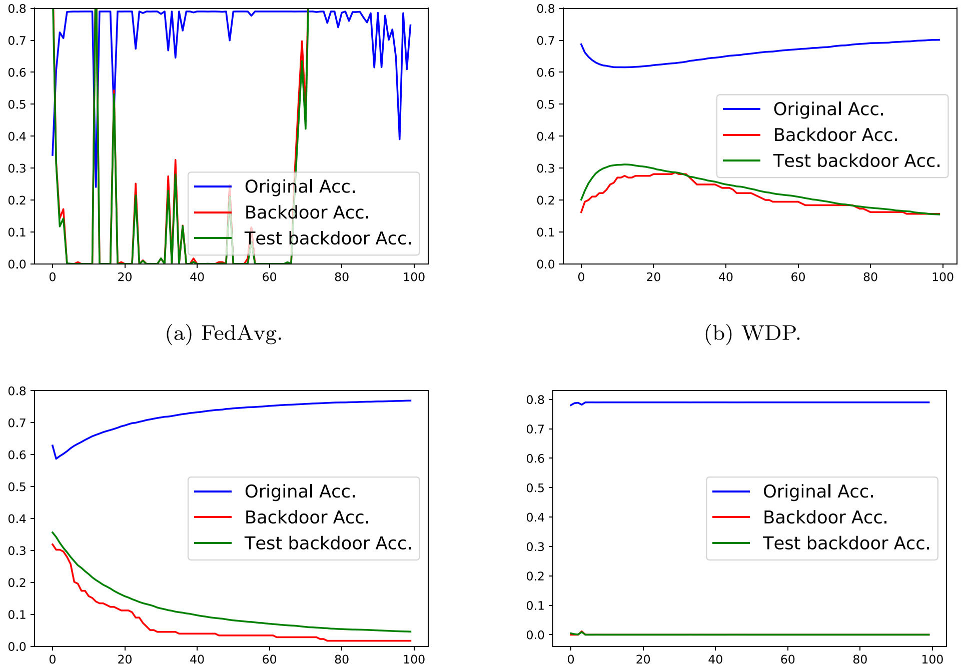
5.3. Analysis of the convergence with RFOut-1d

We claim that RFOut-1d, in addition to being an effective defense in FL, allows the global model to converge to a common solution in less rounds of learning, by means filtering out those parameters that deviate from the solution set by majority. We show it by analyzing the convergence of the models in both the original and the backdoor task throughout the learning rounds.

We choose the pattern-key backdoor attack on CelebA-S and show only two classical aggregation operators (*FedAvg* and *WDP*), *RLR* in its best version including *norm clipping* and *noise* and RFOut-1d in order to reduce the number of figures.

The convergence of the chosen models is presented in Fig. 4. Clearly, FedAvg shows the worst performance while RFOut-1d outperforms all baselines in two ways:

1 Regarding the accuracy of the original task, RFOut-1d ensures that it is not compromised in any of the attack



(a) FedAvg.

(b) WDP.

(c) RLR.

(d) RFOut-1d.

Fig. 4. Convergence plots in the CelebA-S pattern-key backdoor attack experiment. We show both the convergence of the original task (Original task accuracy, in blue) and the backdoor task (Backdoor accuracy and Test backdoor accuracy, in red and green respectively).

Table 11
Mean results for the distributed backdoor attack in terms of accuracy in CelebA-S. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	σ	CelebA-S		
			Original	Backdoor	Test
No attack	0	0	0.7900	–	–
FedAvg	0	0	0.6793	0.9772	0.9944
Median	0	0	0.3701	0.8636	0.8178
Trimmed-mean	0	0	0.7831	**0.0000**	0.0014
NormClip	1	0	0.7604	**0.0000**	0.0499
WDP	1	0.0025	0.7896	**0.0000**	0.0031
RLR	0	0	0.2276	0.9454	0.9704
RLR†	0.5/1	0.0001	0.2686	0.8878	0.9351
RFOut-1d	0	0	0.7602	0.0021	0.0076
RFOut-1d†	0.5/1	0.0001	**0.7897**	**0.0000**	**0.0000**

Table 12
Mean results for the distributed backdoor attack in terms of accuracy in CelebA-A. The symbol † denotes the combination of a defense with norm clipping and noise addition. We also show, in the first row, the expected accuracy with *FedAvg* but without any attack. The best result for each of the test sets is highlighted in bold.

	M	σ	CelebA-A		
			Original	Backdoor	Test
No attack	0	0	0.7973	–	–
FedAvg	0	0	0.5796	0.9643	0.9871
Median	0	0	0.7759	0.0363	0.0525
Trimmed-mean	0	0	0.7888	0.0714	0.0166
NormClip	1	0	0,7954	0.0666	0.0079
WDP	1	0.0025	0.7087	0.1764	0.1602
RLR	0	0	0.2113	0.9765	0.9523
RLR†	0.5/1	0.0001	0.4127	0.6154	0.6433
RFOut-1d	0	0	0.6223	0.0284	0.0367
RFOut-1d†	0.5/1	0.0001	**0.7997**	**0.0000**	**0.0013**

attempts, while in the rest of the baselines the performance is more unstable, becoming the global model's compromised in several rounds of learning.

2 Regarding the backdoor tasks, RFOut-1d demonstrates an outstanding performance and shows a clear improvement over the rest of the baselines. In fact, the attack is not successful in any learning round.

We stress out the relevance of this fact because, despite the acceptable results achieved by the rest of the defenses, the attack is relatively successful at certain learning rounds, which also compromise the integrity of the model. This fact, combined with the fast convergence provided by RFOut-1d, further highlights the success of this approach as an aggregation operator as well as a defense in FL.

6. Conclusions

We addressed the defense against model-poisoning backdoor attacks, which is a real challenge of FL. Based on the claim that the updates from adversarial clients would represent outliers in the Gaussian distribution of clients' updates, we propose RFOut-1d, a defense mechanism based on a robust filtering of one-dimensional outliers in the federated aggregation operator. After evaluating RFOut-1d in a variety of settings under different backdoor attacks, and comparing it with the state of the art defenses, the results shows that our claim holds. Therefore, we state that:

- RFOut-1d is a highly effective defense that dissipates the impact of the backdoor attacks to the point of (almost) nullifying them throughout all the learning rounds.

- In some scenarios, RFOut-1d outperforms the results achieved without any attack, which shows its capacity to filter out clients who are hindering the training process.
- In contrast to other defenses, it does not hinder the FL process by keeping (or even improving) the performance of the model in the original task.
- The convergence of the model to the common solution is accelerated and optimized by filtering out clients that diverge from this solution.

To conclude, we have shown that RFOut-1d is a high quality defense as well as a proper federated aggregation operator by effectively stopping the effect of attacks while favoring the learning of the global model.

CRediT authorship contribution statement

Nuria Rodríguez-Barroso: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Eugenio Martínez-Cámara:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **M. Victoria Luzón:** Conceptualization, Methodology, Writing – review & editing. **Francisco Herrera:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research work is supported by the R&D&I, Spain grants PID2020-119478GB-I00, PID2020-116118GA-I00 and EQC2018-005084-P funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”, Spain. Nuria Rodríguez Barroso were supported by the grant FPU18/04475 funded by MCIN/AEI/10.13039/501100011033 and by “ESF Investing in your future”, Spain. Eugenio Martínez Cámara, Spain were supported by the grant IJC2018-036092-I funded by MCIN/AEI/10.13039/501100011033.

References

[1] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, H. Yu, Federated learning, Synth. Lect. Artif. Intell. Mach. Learn. 13 (3) (2019) 1–207.

[2] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, Knowl.-Based Syst. 216 (2021) 106775.

[3] J. Pang, Y. Huang, Z. Xie, J. Li, Z. Cai, Collaborative city digital twin for the COVID-19 pandemic: A federated learning solution, Tsinghua Sci. Technol. 26 (5) (2021) 759–771.

[4] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, Adversarial classification, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 99–108.

[5] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Vol. 108, 2020, pp. 2938–2948.

[6] Z. Xiong, Z. Cai, D. Takabi, W. Li, Privacy threat and defense for federated learning with Non-i.i.d. Data in AIoT, IEEE Trans. Ind. Inf. (Early Access) (2021).

[7] Y. Song, T. Liu, T. Wei, X. Wang, Z. Tao, M. Chen, FDA3: Federated defense against adversarial attacks for cloud-based iIoT applications, IEEE Trans. Ind. Inf. (Early Access) (2020).

[8] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210.

[9] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, D. Papailiopoulos, Attack of the tails: Yes, you really can backdoor federated learning, in: Advances in Neural Information Processing Systems, Vol. 33, 2020.

[10] M. Nasr, R. Shokri, A. Houmansadr, 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, 2019, pp. 739–753.

[11] F. Suya, S. Mahloujifar, D. Evans, Y. Tian, Model-targeted poisoning attacks: Provable convergence and certified bounds, 2020, CoRR arXiv:2006.16469.

[12] L. Li, W. Xu, T. Chen, G.B. Giannakis, Q. Ling, RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. No. 01, 2019, pp. 1544–1551.

[13] J. So, B. Güler, A.S. Avestimehr, Byzantine-resilient secure federated learning, IEEE J. Sel. Areas Commun. Early access (2020).

[14] C. Fung, C.J. Yoon, I. Beschastnikh, The limitations of federated learning in sybil settings, in: 23rd International Symposium on Research in Attacks, Intrusions and Defenses, 2020, pp. 301–316.

[15] A.N. Bhagoji, S. Chakraborty, P. Mittal, S. Calo, Analyzing federated learning through an adversarial lens, in: Proceedings of the 36th International Conference on Machine Learning, Vol. 97, 2019, pp. 634–643.

[16] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, P.S. Yu, Privacy and robustness in federated learning: Attacks and defenses, 2020, CoRR arXiv:2012.06337.

[17] Z. Sun, P. Kairouz, A.T. Suresh, H.B. McMahan, Can you really backdoor federated learning? 2019, CoRR arXiv:1911.07963.

[18] C. Xie, K. Huang, P.-Y. Chen, B. Li, DBA: Distributed backdoor attacks against federated learning, in: International Conference on Learning Representations, 2020.

[19] I.F. Ilyas, X. Chu, Data Cleaning, Association for Computing Machinery, 2019.

[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Vol. 54, 2017, pp. 1273–1282.

[21] Y. Chen, L. Su, J. Xu, Distributed statistical machine learning in adversarial settings: Byzantine gradient descent, 2017, CoRR arXiv:1705.05491.

[22] D. Yin, Y. Chen, K. Ramchandran, P.L. Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, 2018, CoRR arXiv:1803.01498.

[23] M.S. Ozdayi, M. Kantarcioglu, Y.R. Gel, Defending against backdoors in federated learning with robust learning rate, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 10, 2021, pp. 9268–9276.

[24] P. Laskov, R. Lippmann, Machine learning in adversarial environments, Mach. Learn. 81 (2010) 115–119.

[25] L. Huang, A.D. Joseph, B. Nelson, B.I. Rubinstein, J.D. Tygar, Adversarial machine learning, in: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, in: AISec '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 43–58.

[26] B. Nelson, M. Barreno, F. Jack Chi, A.D. Joseph, B.I.P. Rubinstein, U. Saini, C. Sutton, J.D. Tygar, K. Xia, Misleading learners: Co-opting your spam filter, in: Machine Learning in Cyber Trust: Security, Privacy, and Reliability, Springer US, Boston, MA, 2009, pp. 17–51.

[27] C. Croux, P. Filzmoser, M. Oliveira, Algorithms for projection–Pursuit robust principal component analysis, Chemometr. Intell. Lab. Syst. (ISSN: 0169-7439) 87 (2) (2007) 218–225.

[28] L. Lyu, H. Yu, X. Ma, L. Sun, J. Zhao, Q. Yang, P.S. Yu, Privacy and robustness in federated learning: Attacks and defenses, 2020, CoRR arXiv:2012.06337.

[29] M. Nasr, R. Shokri, A. Houmansadr, 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, 2019.

[30] D. Cao, S. Chang, Z. Lin, G. Liu, D. Sun, Understanding distributed poisoning attack in federated learning, in: 2019 IEEE 25th International Conference on Parallel and Distributed Systems, ICPADS, 2019, pp. 233–239.

[31] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends® Theor. Comput. Sci. 9 (3–4) (2014) 211–407.

[32] X. Zhou, M. Xu, Y. Wu, N. Zheng, Deep model poisoning attack on federated learning, Future Internet (ISSN: 1999-5903) 13 (3) (2021).

[33] G. Sun, Y. Cong, J. Dong, Q. Wang, J. Liu, Data poisoning attacks on federated machine learning, 2020, CoRR arXiv:2004.10020.

[34] L. Lamport, R. Shostak, M. Pease, The Byzantine generals problem, in: Concurrency: The Works of Leslie Lamport, Association for Computing Machinery, New York, NY, USA, 2019, pp. 203–226.

[35] J. Steinhardt, P.W. Koh, P. Liang, Certified defenses for data poisoning attacks, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 3520–3532.

[36] M. Shayan, C. Fung, C.J.M. Yoon, I. Beschastnikh, Biscotti: A ledger for private and secure peer-to-peer machine learning, 2018, CoRR arXiv:1811.09904.

[37] S. Shen, S. Tople, P. Saxena, Auror: defending against poisoning attacks in collaborative deep learning systems, in: Proceedings of the 32nd Annual Conference on Computer Security Applications, 2016, pp. 508–519.

[38] X. Cao, M. Fang, J. Liu, N.Z. Gong, FLTrust: Byzantine-robust federated learning via trust bootstrapping, in: ISOC Network and Distributed System Security Symposium, 2021.

[39] D. Yin, Y. Chen, R. Kannan, P. Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, in: Proceedings of the 35th International Conference on Machine Learning, Vol. 80, 2018, pp. 5650–5659.

[40] P. Blanchard, E.M. El Mhamdi, R. Guerraoui, J. Stainer, Machine learning with adversaries: Byzantine tolerant gradient descent, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 119–129.

[41] E.M. El Mhamdi, R. Guerraoui, S. Rouault, The hidden vulnerability of distributed learning in byzantium, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 3521–3530.

[42] J. Bernstein, J. Zhao, K. Azizzadenesheli, A. Anandkumar, SignSGD with majority vote is communication efficient and fault tolerant, in: International Conference on Learning Representations, 2019.

[43] X. Chen, C. Liu, B. Li, K. Lu, D. Song, Targeted backdoor attacks on deep learning systems using data poisoning, 2017, CoRR arXiv:1712.05526.

[44] B.P. Roe, Central limit theorem, in: The Concise Encyclopedia of Statistics, Springer New York, New York, NY, 2008, pp. 66–68.

[45] N. Rodríguez-Barroso, G. Stipcich, D. Jiménez-López, J.A. Ruiz-Millán, E. Martínez-Cámara, G. González-Seco, M.V. Luzón, M.A. Veganzones, F. Herrera, Federated learning and differential privacy: Software tools analysis, the sherpa.ai FL framework and methodological guidelines for preserving data privacy, Inf. Fusion 64 (2020) 270–292.

[46] S. Caldas, P. Wu, T. Li, J. Konecný, H.B. McMahan, V. Smith, A. Talwalkar, LEAF: a benchmark for federated settings, 2018, CoRR arXiv:1812.01097.

[47] Y. Ma, X. Zhu, J. Hsu, Data poisoning against differentially-private learners: Attacks and defenses, in: International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 4732–4738.