

# Defending Byzantine Attacks in Ensemble Federated Learning: A Reputation-based Phishing Approach

Beibei Li, *Member, IEEE*, Peiran Wang, *Student Member, IEEE*, Qinglei Kong, *Member, IEEE*, Yuan Zhang, *Member, IEEE*, and Rongxing Lu, *Fellow, IEEE*

**Abstract**—Emerging as a promising distributed learning paradigm, federated learning (FL) has been widely adopted in many fields. Nonetheless, a big challenge for FL in real-world implementation is Byzantine attacks, where compromised clients can mislead or poison the training model by falsifying or manipulating the local model parameters. To solve this problem, in this paper, we present a reputation-based Byzantine robust-FL scheme (called FLPhish) for defending Byzantine attacks under the Ensemble Federated Learning architecture (called EFL). Specifically, we first develop a novel ensemble FL architecture, EFL, which allows FL compatible with different deep learning models in different clients. Second, we craft a phishing algorithm for the EFL architecture to identify possible Byzantine behaviors. Third, a Bayesian inference based reputation mechanism is devised to measure each client's level of confidence and to further identify Byzantine clients. Last, we strictly analyze how the FLPhish scheme defend against backdoor attacks. Extensive experiments under different settings demonstrate that the proposed FLPhish achieves great efficacy in defending Byzantine attacks in EFL. FLPhish is tested with different fractions of Byzantine clients and different degrees of distribution imbalance.

**Index Terms**—Federated learning, ensemble learning, Bayesian inference-based reputation, phishing.

## I. INTRODUCTION

MANY elements of our daily lives and society have benefited from deep learning tasks in natural language processing, computer vision, and anomaly detection. To learn complex rules, such activities necessitate a large dataset. In most cases, these huge datasets are acquired by the application developers from users, such as the shopping app users' purchase record data, patients' clinical data, and etc. Nonetheless, in recent years, there has been an explosion in social concerns about personal privacy, making it difficult to get data directly from users anymore. Under these circumstances, each individual's data is referred to as an 'Isolated

This paper is an extended version of the paper titled 'FLPhish: Reputation-Based Phishing Byzantine Defense in Ensemble Federated Learning', which was published in IEEE ISCC 2021, and awarded 'Best Paper'.

B. Li and P. Wang are with the School of Cyber Science and Engineering, Sichuan University, Chengdu, Sichuan, China 610065. Email: libeibei@scu.edu.cn; wangpeiran@stu.scu.edu.cn.

Q. Kong is with the Future Network of Intelligence Institute, The Chinese University of Hong Kong, Shenzhen, China 518172, and also with The University of Science and Technology of China, Hefei, China 230052. Email: kql8904@163.com.

Y. Zhang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China 610054. Email: zy\_joye@126.com.

R. Lu is with the Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada E3B 5A3. Email: rlu1@unb.ca.

TABLE I  
SUMMARY OF NOTATIONS

Term	Description
$s$	central server in FL
$c_i$	the $i$ th client in FL, $i = 1, 2, 3, \dots, u$
$d_i$	the local dataset preserved by the $i$ th client
$C$	the ensemble of all the clients
$u$	the number of clients
$D_t$	the unlabeled dataset chosen by $s$ in each procedure
$D$	the unlabeled dataset preserved by $s$
$n$	the number of samples in $D_t$
$B_t$	the labeled dataset ('bait') chosen by $s$ in each procedure
$B$	the labeled dataset preserved by $s$
$m$	the number of samples in $B_t$
$a_i^t$	the accuracy of predictions of $B_t$ made by $c_i$ in $t$ th procedure
$q_i$	the label of $c_i$ to judge it is a malicious client or not
$r_q$	the threshold of malicious clients
$x_l^t$	the $l$ th data point in $D_t$
$b_i$	the Byzantine attacker
$\sigma$	the 'trigger' in the backdoor attack
$\iota$	the backdoor label in the backdoor attack
$M$	global model preserved by $s$
$m_i$	local models trained by the $i$ th client
$k_i^t$	the predictions ('knowledge') made by the $i$ th client in the $t$ th procedure
$\hat{y}_l^t$	the ensembled prediction of data point $x_l^t$
$\hat{y}_l^1$	the prediction of $l$ th data point made by $i$ th client
$K_t$	the aggregated labels (predictions) of the $t$ th iteration's unlabeled dataset

Data Island'. The existence of each 'Isolated Data Island' drives the development of privacy-preserving solutions like Federated Learning (FL) [1]–[3]. Bonawitz *et al.* built the first FL system that is operated on Google's mobile phone to train a global model based on TensorFlow<sup>1</sup>. Its FL system could be operated on thousands of mobile phones. Moreover, a team from WeBank developed an FL scheme called FATE<sup>2</sup> for credit risk prediction. Furthermore, some former researchers have also applied FL to some industrial cyber-physical Systems [4]–[6].

FL is a distributed machine learning paradigm, which allows the central server in the paradigm to produce a global model

<sup>1</sup><https://federated.withgoogle.com/>

<sup>2</sup><https://github.com/FederatedAI/FATE>

without getting each individual's private data. Instead of gathering private data from each user, the central server in FL aggregates all the model gradient updates from distributed clients to its global model. In each iteration of FL, the central server sends a model to each client. Each client updates the model using its private data and sends the model gradient update back to the central server. In the central server, all the clients' updates are aggregated to a global model gradient update, and the global model gradient update is utilized to update the global model. Thus, FL not only protects each participating individual's privacy, but also leverages the capabilities of the end users' computation and storage.

Since thousands of clients from different sources may participate in the training process, security issues also exist in the distributed FL system. On one hand, former researchers have already studied the privacy problems of FL and have proposed the corresponding schemes to enhance privacy protection in FL [7]–[9]. On the other hand, FL faces threats from the poisoning attacks launched by malicious attackers among the FL clients [10], [11]. And such attacks are referred to as Byzantine attacks in wireless communication network [12]–[15]. By poisoning the clients' datasets or directly manipulating the gradient updates, the incorrect gradient updates are sent by the malicious clients to the central server, which causes the global model to learn incorrect knowledge. As a result, this process renders the central server's global model obsolete. Furthermore, Byzantine attacks can be separated into two types according to the attack consequences. In the first type of Byzantine attacks, called denial-of-service attacks (including untargeted attacks, targeted attacks, e.g), the Byzantine attackers intend to disturb the global model thus making it produce wrong predictions of the normal dataset [16]–[20]. In another type of Byzantine attack, called backdoor attacks, the disturbed global model will make wrong predictions of the data samples that have 'backdoor' in them [21]–[25].

Former researchers have offered certain Byzantine-robust techniques to deal with malicious Byzantine clients under the FL application settings [26]–[35]. Byzantine-robust techniques try to construct a global model with high accuracy in the presence of a finite number of malicious clients. According to their different mechanisms, we divide Byzantine-robust approaches into two major types. The first (named Byzantine-Detection) is based on the development of a Byzantine-robust aggregation algorithm that distinguishes suspected clients from benign clients. The suspected clients' gradient updates are subsequently removed from the aggregation process by the server. For instance, in the DRACO scheme proposed by Chen *et al.*, each node analyzes duplicate gradients that the parameter server uses to mitigate the effects of adversarial updates [27]. Another Byzantine-robust technique (named Byzantine-Tolerance) seeks to ensure that the aggregation process is tolerant to poisoned updates from Byzantine clients without excluding Byzantine clients like Median [29]. In Median, the FL server sorts the values of each parameter and picks the median value of each parameter as the value to be utilized in global model updates. In this study, we provide a unique reputation-based phishing method (named FLPhish) to protect against Byzantine attacks in EFL. Our contributions are three-

fold:

- We design a new FL architecture, Ensemble Federated Learning (called EFL), which utilizes an unlabeled dataset to replace the gradient updates in typical FL. This architecture is flexible because it is compatible with different deep learning models on different clients.
- We craft a 'phishing' method based on EFL to detect Byzantine attacks. The 'phishing' method employs the labeled dataset to detect the potential Byzantine clients in the EFL architecture, which preserves the security of the EFL.
- We present a Bayesian inference-based reputation mechanism to promote FLPhish's aggregation. The reputation mechanism gives each client a reputation to measure its confidence value and identifies the clients with low reputation values as Byzantine clients, which helps FLPhish identify the Byzantine clients with higher accuracy.

## II. RELATED WORK

In this section, we discuss about the related research work about the proposed Byzantine defense methods in FL and the proposed reputation mechanism in cybersecurity research.

### A. Byzantine Defense Methods in Federated Learning

Byzantine-robust schemes are very important for FL to enhance its security as Byzantine attacks can cause great damages to the FL system. Recent years have witnessed the increasing interest in the research of Byzantine-robust schemes in the context of FL. Most of the current Byzantine-robust FL methods tend to make a more robust aggregation rule which aims to tolerate the presence of Byzantine clients. For example, in 2017, Chen *et al.* developed an approach called Krum, which selects one client's update as a global model based on a square-distance score in each iteration [26]. In the same year, Blanchard *et al.* proposed two Byzantine-tolerant FL aggregation rules called Trimmed mean and Median [29]. Trimmed Mean considers each parameter of the model update individually. Trimmed Mean sorts the parameter of the model updates collected, and cuts off the largest ones and the smallest ones. Median sorts the values of each parameter of all local model updates as well. Besides it considers the median value of each parameter as the value of the parameter in the global model update. In 2018, Chen *et al.* designed an approach called DRACO to evaluate redundant gradients that are used by the parameter server to eliminate the effects of adversarial updates. In 2019, Xie *et al.* proposed Zeno, which uses a ranking-based preference mechanism [28]. The server computes a score for each client by using the stochastic zero-order oracle. Then Zeno presents a ranking list of clients based on the estimated descent of the loss function and the magnitudes. At last, Zeno computes the global model update by aggregating the clients with the highest scores. In 2020, SLSGD developed by Xie *et al.* also uses trimmed mean as the robust aggregation rules for Byzantine-robust FL [36]. In the same year, Cao *et al.* proposed a Byzantine-tolerant scheme: FLTrust to introduce the use of trust [30]. In each iteration, the server calculates a trust score for each client at first and

lowers the trust score if the client's local model update's direction deviates more from the direction of the global model update. The client with a trust score lower than the threshold is considered a malicious client. In 2021, a privacy-enhanced FL (PEFL) framework is presented by Liu *et al.* [37]. PEFL takes advantage of homomorphic encryption to protect the privacy of the clients. Furthermore, a channel using the effective gradient data extraction is provided for the server to punish poisoners.

### B. Reputation Mechanism in Cybersecurity

The reputation mechanism is valued as a way to measure an entity's performance in a long term, such as in an online social network [38], and in a smart grid system, [39], [40]. In 2012, Das *et al.* first presented a dynamic trust computation model called SecuredTrust. This framework is used to distribute the workload and deal with the altering behavior of malicious clients [41]. To calculate and manage trust and reputation of CSP and SNP services, Zhu *et al.* proposed an authenticated trust and reputation calculation and management system in wireless sensor networks and cloud computing in 2015 [42]. Lei *et al.* presented a reputation-based Byzantine Fault Tolerance rule in 2018, which uses a reputation model to assess the performance of each node in the blockchain system [43]. If the system detects any malicious behavior, the nodes' discourse rights and reputation in the voting process are reduced. They also provided a primary change method based on reputation. The node with a higher reputation would have more chances to generate fresh valid blocks, lowering the system's security risk. In 2020, Chouikhi *et al.* developed a reputation computing and credibility model to improve network efficiency [44]. They measured a vehicle's behavior toward other vehicles and network services using its reputation score or worth. And a vehicle's credibility is utilized to determine the correctness of a reputation score it offers. In the same year, Wen *et al.* designed a Dirichlet reputation-based approach, and used the reputation score to choose a trustworthy Helper as a friendly jammer in a wireless cooperative system (WCS) [45]. Furthermore, they devised a multi-threshold fake noise detection approach. They gave ratings on a scale of one to ten. The graded ratings were directly represented and reflected in the generated reputation scores in the Dirichlet reputation-based method. In 2021, Liang *et al.* introduced an intrusion detection system with a Markov-based reputation algorithm [46]. The RS-HgMTD model of the Hidden Generalized Mixture Transition Distribution (HgMTD) is designed to help each vehicle in the VANET assess the creditworthiness of its neighbors.

## III. MODELS AND DESIGN GOALS

In this section, we discuss the system model, show the threat model and identify our design goals.

### A. System Model

We first show the system architecture of a typical FL with two entities, FL server, and a group of FL clients.

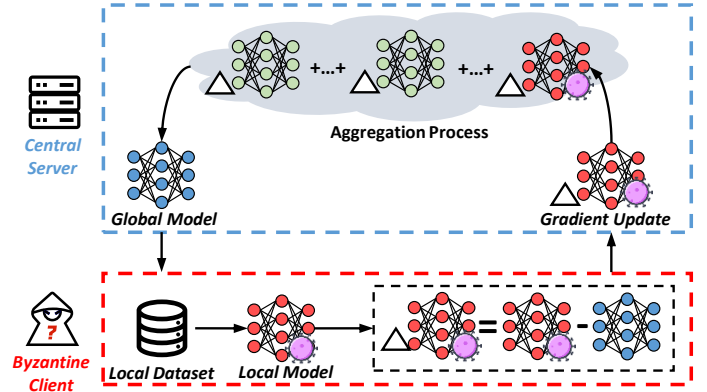


Fig. 1. System Model & Threat Model.

1) *FL Server*: In each iteration, the FL server  $s$  provides a global model to each client. The FL server aggregates all of the gradient updates to a global update based on FedAvg after receiving them from all of the clients. The FL server updates the global model after the aggregation process.

2) *FL Client*: The local dataset  $d_i$  gathered by each FL client  $c_i$  ( $c_i$  denotes the  $i$ th client in FL) is preserved by each FL client  $c_i$ . To update the model obtained from the FL server, the FL client  $c_i$  uses its local dataset  $d_i$ . The model gradient updates are then sent back to the FL server. Meanwhile, it repeats the preceding steps throughout the FL process until the FL server  $s$  stops transmitting new models.

### B. Threat Model

Byzantine attacks are a problem in the current system. We separate Byzantine attacks into two types according to the attack consequences:

1) *Denial-of-Service Byzantine Attack*: In denial-of-service attacks, the Byzantine attackers intend to disturb the global model, thus making it produce wrong predictions of the normal dataset [16]–[20]. The label flipping attack in the current system model can be used by a malicious Byzantine client  $b_i$  to launch Byzantine attacks against the global model. Label flipping attacks require  $b_i$  to change the labels of training data while ensuring that the data's features remain unchanged [47]. The local model of the Byzantine client,  $b_i$  is trained with incorrect labels, resulting in a 'poisoned' model with low accuracy. Then Byzantine client  $b_i$  dispatches the false model gradient updates to the central server. Therefore, the false model gradient updates cause the central server to learn the falsely distilled knowledge from clients. The server  $s$ 's aggregation process is performed on FedAvg, which takes each client  $c_i$ 's dataset  $d_i$ 's size as the aggregation weight for  $c_i$ . This means that a client  $c_i$  with a larger size of  $d_i$  gets a larger aggregation weight. Meanwhile, FedAvg takes the size of  $d_i$  declared by  $c_i$  as  $d_i$ 's real size, which means  $b_i$  can declare a fake size value larger than  $d_i$ 's real size value to increase the impact of the attack. If the weight of the malicious clients reaches a threshold, the central server is misguided to produce false predictions.

In section V, we first divided denial-of-service attacks into 2 types of attacks: The first one is the untargeted attack,

which makes the Byzantine attackers change the labels of data samples from one type into another type (for instance, all the samples with label ‘5’ are changed to label ‘0’). Another one is the random attack which makes the Byzantine attackers randomly change the label of data samples. Then we evaluated our framework against the 2 types of denial-of-service attacks.

2) *Backdoor Byzantine Attack*: In backdoor attacks, called backdoor attacks, the disturbed global model will make wrong predictions of the data samples that have ‘backdoor’ in them [21]–[25]. Backdoor attackers in FL need to embed some ‘triggers’ (noted as  $\sigma$ ) in their local dataset. Then they relabel the data samples with  $\sigma$  as the target label  $\iota$ . Each backdoor attacker adopts the preprocessed local dataset with  $\sigma$  to update the global model it received from the FL server. Then it transfers the poisoned model updates, which contain the information that the data sample with  $\sigma$  is predicted as the target label  $\iota$  to the FL server. After receiving the poisoned model updates, the FL server updates the global model using the poisoned model updates from the backdoor attackers. After the update, the FL server’s model misclassifies the data with the  $\sigma$  to the target label  $\iota$  of backdoor attackers. Take a backdoor attack process towards the construction of an FL on CIFAR-10 as an example. The backdoor attacker adds a grey square as  $\sigma$  in each data sample. Each data sample with a  $\sigma$  is labeled as ‘cat’. Then the backdoor attacker uses these data samples to update the global model from the central server and transfers the model gradient updates containing the  $\sigma$  information to the central server. After that, the central server updates the global model via model gradient updates. Thus, the global model learns the  $\sigma$  information from the backdoor attacker. It misclassifies the data sample with the  $\sigma$  as ‘cat’ as well.

In section IV.D, we did a theoretical analysis of how our FLPhish scheme can defend against backdoor Byzantine attacks.

### C. Design Goals

The proposed FLPhish scheme’s main goal is to provide a reliable method for accurately resisting opportunistic Byzantine attacks in our EFL architecture. The following are our design goals:

1) The typical FL design has several flaws, including incompatibility with various deep learning models in different clients and significant communication costs. As a result, we created EFL, a novel FL architecture inspired by the idea of ensemble learning. It lowers the cost of network transfers and expands our ability to defend against Byzantine attacks in FL.

2) The proposed EFL architecture currently needs a robust Byzantine attack protection mechanism. We urgently seek an efficient solution to deal with malicious Byzantine clients in FL because they cannot be trusted. In our proposed EFL architecture, we describe a phishing-based approach to guard against Byzantine attacks.

3) As the performance of each client remains not stable in each iteration, it is important for our scheme to accurately measure each client’s level of confidence in the long term. Therefore, we further propose an effective Bayesian inference-based reputation scheme based on our phishing-based model to spot Byzantine attacks compromised by malicious users.

---

### Algorithm 1 EFL

---

**Input:** each FL client  $c_i$ ,  $i = 1, 2, 3, \dots, u$  with its local dataset  $d_i$ ,  $i = 1, 2, 3, \dots, u$ ; an FL central server  $s$  with the unlabeled dataset  $D$  preserved by itself; number of the FL training iterations  $T$ ; the size of unlabeled batch  $n$  in each iteration;

**Output:**

```

1:  $\mathbf{m}_i \leftarrow$  each FL client  $c_i$  trains a local model using its
   local dataset  $d_i$  collected by itself;
2: for  $t=1, 2, 3, \dots, T$  do
3:    $s$  selects  $D_t$  (containing  $n$  data samples) from  $D$ ;
4:   for  $i=1, 2, 3, \dots, u$  do
5:      $s \xrightarrow{D_t} c_i$ ;
6:      $c_i$  makes predictions  $\mathbf{k}_i^t$  of the  $D_t$ ;
7:      $c_i \xrightarrow{\mathbf{k}_i^t} s$ ;
8:   end for
9:    $Y_t = \text{KnowledgeEnsemble}(\mathbf{k}_1^t, \mathbf{k}_2^t, \mathbf{k}_3^t, \dots, \mathbf{k}_u^t)$ ;
10:   $\mathbf{M} = \text{ModelUpdate}(Y_t, D_t, \mathbf{M})$ ;
11: end for
12: return  $\mathbf{M}$ .
```

---



---

### Algorithm 2 KnowledgeEnsemble

---

**Input:** the ensemble of each client’s prediction (‘distilled knowledge’)  $\mathbf{k}_i^t$   $\{i=1, 2, 3, \dots, u\}$ ; the size of each client’s local dataset  $e_i$   $\{i=1, 2, 3, \dots, u\}$ ; the unlabeled dataset  $D_t$  used in  $t$ th iteration;  $\hat{\mathbf{y}}^t$  is the ensembled prediction of dataset  $D_t$ ;  $\hat{\mathbf{y}}_i^t$  denotes the prediction of the dataset  $D_t$  made by  $i$ th client;

**Output:**

```

1: for  $l = 1, 2, 3, \dots, n$  ( $l = 1, 2, 3, \dots, n$ , denotes the data
   point in the unlabeled dataset) do
2:    $\hat{\mathbf{k}}^t \leftarrow \sum_{i=1}^u \frac{e_i}{\sum_{i=1}^u e_i} \hat{\mathbf{k}}_i^t$ ;
3:    $\hat{\mathbf{y}}^t \leftarrow \text{argmax}(\hat{\mathbf{k}}^t)$ ;
4: end for
5: return  $\hat{\mathbf{y}}^t$ .
```

---

## IV. PROPOSED FLPHISH SCHEME

In this section, we show the proposed FLPhish scheme including the EFL architecture, the phishing mechanism, the reputation mechanism, and the security analysis of backdoor attacks.

### A. Designed Ensemble Federated Learning

Inspired by ensemble learning, we propose a new FL architecture, called EFL.

Typical FL architecture collects FL clients’ model gradient updates, and aggregates them for the global model updates. While unlike them, our proposed FLPhish scheme applies an unlabeled dataset (the central server preserves this dataset, while does not have its label or lacks enough resources to get its label) preserved by the central server and all the clients’ predictions of this dataset for global model updates.

1) *Client*: Each FL client  $c_i$  ( $i$  denotes the client’s serial number) collects and labels its local data from various data sources, resulting in its local dataset. To train its local model,  $c_i$  uses the local dataset. When  $c_i$  receives a public dataset

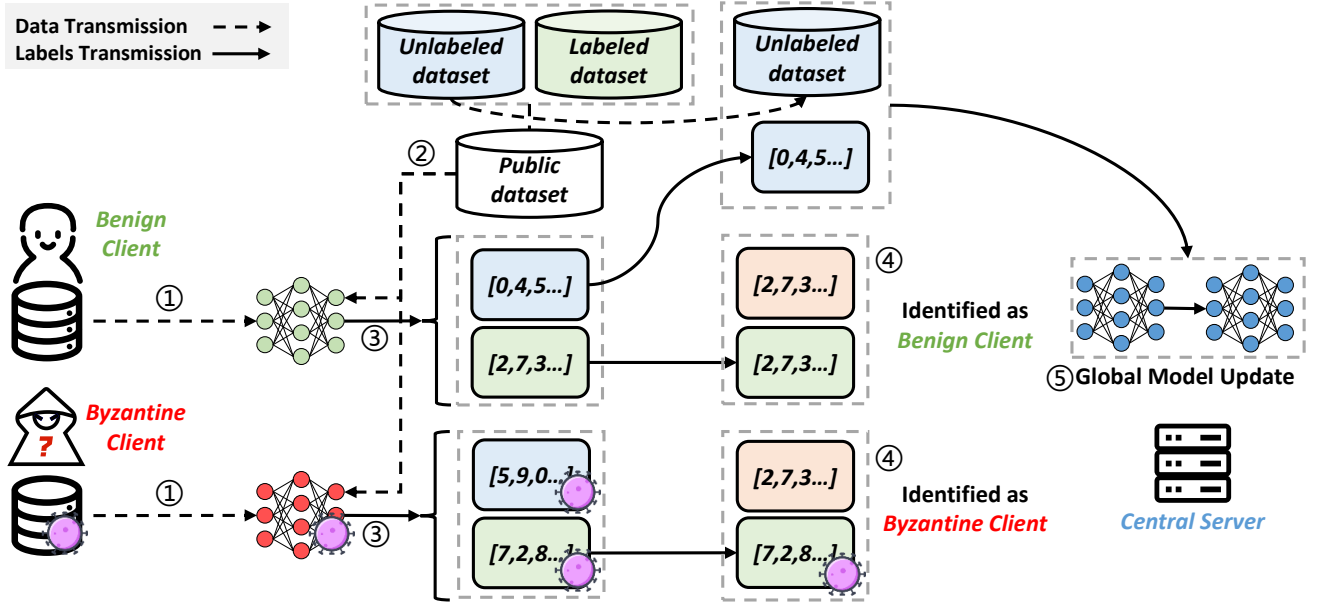


Fig. 2. Our Proposed FLPhish Scheme.

from the central server in each iteration, it uses its local model to produce predictions for the dataset and sends the predictions back to the central server  $s$ .

2) *Central Server*: The public dataset and the global model are built and preserved by the central server  $s$ . The public dataset is made up of unlabeled data that the central server does not have labels for it.  $s$  either collects data from other data sources or generates data by itself to build the public dataset. In each iteration, the public dataset is transmitted to each client by the central server  $s$  after its construction. Each client  $c_i$  uses its local model preserved by itself to make predictions about the public dataset and sends the prediction results back to  $s$ . After getting all of the predictions from all of the customers,  $s$  aggregates them into a single set of predictions

$$\hat{\mathbf{k}}^t = \sum_{i=1}^u \frac{e_i}{\sum_{i=1}^u e_i} \hat{\mathbf{k}}_i^t. \quad (1)$$

Then the central server  $s$  utilizes the aggregated predictions  $\hat{\mathbf{k}}^t$  to compute the aggregated labels

$$\hat{\mathbf{y}}^t \leftarrow \text{argmax}(\hat{\mathbf{k}}^t). \quad (2)$$

Then the central server  $s$  employs the public dataset and the aggregated labels as the dataset's labels to update the global model.

Compared with typical FL architectures, this proposed architecture has a variety of advantages:

- Because model gradient updates are used in typical FL, the global model and each local model preserved by each client must be the same sort of model. Otherwise, the FL aggregation process for the various shapes of model parameters will not be possible. However, in our approach, alternative types of neural architectures can be deployed by using distilled knowledge (client-generated predictions of unlabeled data) instead of model gradient

updates as aggregation ingredients. Various clients' local datasets with different selected features are also allowed.

- In comparison to the typical FL architecture, the overhead and latency of the communication process in our developed EFL are greatly decreased. Due to the reduced size of data and labels, transferring data and labels is much faster than transferring models and model gradient updates.
- The design of EFL can prevent the backdoor attack in FL. The detail of theoretical proof is in section IV.D.

### B. Phishing Mechanism-based Detection

The proposed EFL is still vulnerable to byzantine attacks. Malicious clients can modify their local models through label flipping attacks. In order to build a 'poisoned' local model, they assign inaccurate labels to the local dataset. When hostile clients receive unlabeled data from the central server, they make inaccurate predictions (called 'poisoned knowledge') and send them to the central server. The inaccurate predictions are subsequently aggregated as labels for the unlabeled dataset by the central server. The global model is subsequently trained using these unlabeled datasets and the erroneous aggregation predictions by the central server. As a result, an incorrect global model emerges. To deal with Byzantine attacks in EFL, we propose using labeled data in the design of EFL. We called labeled data 'bait'.

**Step-1: Local Model Training** The ensemble of clients  $C = \{c_1, c_2, \dots, c_{n-1}, c_n\}$ . A local dataset  $d_i$  is preserved by each client  $c_i$ . At the beginning of the EFL process, each local model  $\mathbf{m}_i$  is trained by each client  $c_i$  via its local dataset  $d_i$  as

$$\mathbf{m}_i = \text{Train}(d_i). \quad (3)$$

**Step-2: Dataset Transferring** From the unlabeled dataset  $D$  and the labeled dataset  $B$ , the central server randomly selects

**Algorithm 3** Phishing Mechanism

**Input:** each FL client  $c_i$ ,  $i = 1, 2, 3, \dots, u$  with its local dataset  $d_i$ ,  $i = 1, 2, 3, \dots, u$ ; an FL central server  $s$  with the unlabeled dataset  $D$  and labeled dataset  $B$  preserved by itself; the number of FL training iterations  $T$ ; the size of the unlabeled batch  $n$  used in each training iteration; the size of the labeled batch  $m$  used in each training iteration.

**Output:** output result

```

1:  $\mathbf{m}_i \leftarrow$  each client  $c_i$  train a local model using its local
   dataset  $d_i$ .
2: for  $t=1, 2, 3, \dots, T$  do
3:    $s$  selects  $D_t$  (containing  $n$  samples) from  $D$  and  $B_t$ 
   (containing  $m$  samples) from  $B$ .
4:   for  $i=1, 2, 3, \dots, u$  do
5:      $s$  sends  $D_t$  and  $B_t$  to  $c_i$ .
6:      $c_i$  makes predictions  $\mathbf{k}_i^t$  of the  $D_t$  and  $B_t$ .
7:      $c_i$  sends  $\mathbf{k}_i^t$  to  $s$ .
8:      $s$  calculates the accuracy  $a_i^t$  of the predictions of
        $B_t$  made by  $c_i$  in  $t$ th procedure.
9:     if  $a_i^t > r_q$  then
10:       $q_i = 1$ .
11:      for  $j=i+1, i+2, \dots, u-1$  do
12:         $(d, c, \mathbf{k}^t, q, a^t)_j \leftarrow (d, c, \mathbf{k}^t, q, a^t)_{j-1}$ .
13:      end for
14:    end if
15:  end for
16:   $\mathbf{K}_t = \text{KnowledgeEnsemble}(\mathbf{k}_1^t, \mathbf{k}_2^t, \mathbf{k}_3^t, \dots, \mathbf{k}_u^t)$ .
17:   $\mathbf{M} = \text{ModelUpdate}(\mathbf{K}_t, D_t, \mathbf{M})$ .
18: end for
19: return  $\mathbf{M}$ 

```

some data samples to construct the unlabeled dataset  $D_t$  and the labeled dataset  $B_t$  to construct the  $t$ th FL iteration's public dataset.  $D_t$  has  $n$  data samples, while  $B_t$  has  $m$  data samples. Then  $D_t$  and  $B_t$  are sent to each client  $c_i$  by  $s$  as

$$c_i \xleftarrow{D_t, B_t} s. \quad (4)$$

**Step-3: Label Predicting** The predictions of  $D_t$  and  $B_t$  are made by each client  $c_i$  as

$$\mathbf{k}_i^t = \text{Predict}(D_t, B_t, \mathbf{m}_i) \quad (5)$$

( $D_t$  and  $B_t$  are mixed together,  $c_i$  can not distinguish between them) via its local model and the predictions  $\mathbf{k}_i^t$ :

$$\mathbf{k}_i^t = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,g-1} & p_{1,g} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,g-1} & p_{n,g} \end{bmatrix} \quad (6)$$

is sent back to the central server.

Unlike benign clients, the malicious clients return the incorrect prediction to the central server  $s$  to disturb the aggregation process.

**Step-4: Byzantine Identifying** After accepting the predictions  $\mathbf{k}_i^t$  from each client  $c_i$ ,  $s$  needs to identify the Byzantine attackers among all the clients. The predictions of the 'bait' (the labeled dataset  $B_t$ ) are extracted apart from  $D_t$  from  $\mathbf{k}_i^t$ .

Next,  $s$  calculates the accuracy of each client's predictions of  $B_t$  using true label of the  $B_t$ :

$$a_i^t = \text{AccuracyCal}(\mathbf{k}_i^t, B_t). \quad (7)$$

Then the clients with a low accuracy value over predicting  $B_t$  are identified as malicious Byzantine clients.

**Step-5: Global Model Updating** After identifying the malicious clients within all clients,  $s$  aggregates the knowledge  $\hat{\mathbf{k}}^t$  from all clients as

$$\hat{\mathbf{k}}^t = \sum_{i=1}^u \frac{e_i}{\sum_{i=1}^u e_i} \mathbf{k}_i^t. \quad (8)$$

Then the aggregated knowledge  $\hat{\mathbf{k}}^t$  is used to derive the labels by the central server  $s$

$$\hat{\mathbf{y}}^t = \text{argmax}(\hat{\mathbf{k}}^t). \quad (9)$$

**C. Bayesian Inference-based Reputation Mechanism**

The server  $s$  keeps track of the reputation of all the clients  $C$  in the model in a reputation list. Let  $X_i$  be the  $c_i$  client's reputation, which represents  $s$ 's opinion of how probable client  $c_i$  is to be a Byzantine client. The accuracy  $a_i^t$  of client  $c_i$  from the first iteration to the  $t$ th iteration is used to compute  $X_i$ . When client  $c_i$  sends a new update to server  $s$ ,  $s$  uses the accuracy  $a_i^t$  to update the value of  $X_i$ .

At first, the reputation is neutral. With a probability of 50%, server  $s$  considers each client  $c_i$  to be a benign client. The reputation of the server  $s$  is updated when a new update  $\mathbf{k}_i^t$  arrives at the central server. When the client  $c_i$ 's reputation falls below the threshold  $r$ ,  $s$  considers the client  $c_i$  to be a Byzantine client and discards the client's update. When the  $X_i$  surpasses the threshold  $r$ , the client  $c_i$  updates are reviewed in the aggregation.

We employ Bayesian inference to construct our reputation mechanism. Each data prediction made by client  $c_i$  faces two situations: wrong predictions or correct predictions. Thus, the use of binomial parameter distributions becomes a natural choice for our reputation mechanism. Let  $Y_i$  be the event that the number of wrong predictions and correct predictions made by client  $c_i$  is  $\alpha_i$  and  $\beta_i$ . Given  $X_i = \gamma_i$ , the conditional probability is

$$\Pr(Y_i | X_i = \gamma_i) = \binom{\alpha_i + \beta_i}{\alpha_i} \gamma_i^{\alpha_i} (1 - \gamma_i)^{\beta_i}. \quad (10)$$

where  $\alpha_i$  and  $\beta_i$  is the available evidence for the estimation of the  $X$ , and  $\gamma_i$  is unknown. Eq. 10 indicates the likelihood function for  $X$ . According to Bayesian inference we compute the posterior probability as

$$\Pr(X_i = \gamma_i | Y_i) = \frac{\Pr(Y_i | X_i = \gamma_i) \Pr(X_i = \gamma_i)}{\int_0^1 \Pr(Y_i | X_i = x) \Pr(X_i = x) dx}. \quad (11)$$

Given the posterior probability function, the final value for the expectation value of reputation  $X$  is computed as

$$E(X) = \int_0^1 \frac{\Pr(Y_i | X_i = \gamma_i) \Pr(X_i = \gamma_i)}{\int_0^1 \Pr(Y_i | X_i = x) \Pr(X_i = x) dx} \gamma_i d\gamma_i. \quad (12)$$

Furthermore, we decide to use the binomial parameter beta distribution to describe the distribution of  $X$ . Assume  $X$  is a



random variable of the beta distribution with the parameters  $(\alpha, \beta)$ , the density function is

$$f(x, \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}. \quad (13)$$

, where  $B(\alpha, \beta)$  function is the beta function. Therefore we compute the expectation of  $X$  as

$$E(X) = \frac{\alpha}{\alpha + \beta}. \quad (14)$$

Thus, the central server  $s$  computes the reputation  $X_i$  of each client  $c_i$  as  $E(x_i) = \frac{\alpha_i}{\alpha_i + \beta_i}$ . We set the reputation's initial value to 50% by setting the initial value of  $\alpha$  and  $\beta$  to 1. This means that the chances for each client  $c_i$  of being a benign or malicious Byzantine client are the same. When a new iteration of EFL is completed, new evidence for the computation of the reputation is presented in the form of  $\alpha_i'$  and  $\beta_i'$ . The reputation is then computed by

$$E(x_i)' = \frac{(\alpha_i + \alpha_i')}{(\alpha_i + \alpha_i') + (\beta_i + \beta_i')}. \quad (15)$$

In our conference version, we designed an aggregation rule which utilizes a threshold to identify the Byzantine clients. While in this work, we propose a new Byzantine-tolerant aggregation rule, FLPhish-weight.

**FLPhish-weight** Unlike the aggregation rule designed in our conference version, FLPhish-weight does not discard the potential Byzantine clients' updates. On the contrary, it enables the Byzantine clients to participate the aggregation using its reputation value as the aggregation weight. Due to our reputation mechanism, the Byzantine clients are offered a low reputation, so it has a lower influence on the aggregation process. Give a reputation list  $R = [x_1, x_1, x_1, \dots, x_{n-1}, x_n]$ , the prediction results is

$$\mathbf{k}_i^t = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,g-1} & p_{1,g} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,g-1} & p_{n,g} \end{bmatrix}, \quad (16)$$

Meanwhile, FLPhish-weight computes the aggregated knowledge, which is

$$\hat{\mathbf{k}}^t = \sum_{i=1}^u \frac{e_i}{\sum_{i=1}^u e_i} \hat{\mathbf{k}}_i^t \times x_i. \quad (17)$$

Then the server  $s$  uses the aggregated knowledge  $\hat{\mathbf{k}}^t$  to get the labels

$$\hat{\mathbf{y}}^t \leftarrow \text{argmax}(\hat{\mathbf{k}}^t). \quad (18)$$

In a word, FLPhish-weight does not identify the Byzantine clients but treats the clients with a low reputation as 'bad' clients and gives them lower weights.

#### D. Security Analysis of Backdoor Byzantine Attacks

In the former discussion, we have designed the FLPhish scheme which can defend against denial-of-service attacks. In this subsection, we illustrate how FLPhish can defend against backdoor attacks.

Backdoor attackers in FL need to embed some 'triggers' (noted as  $\sigma$ ) in their local dataset. Then they relabel the data

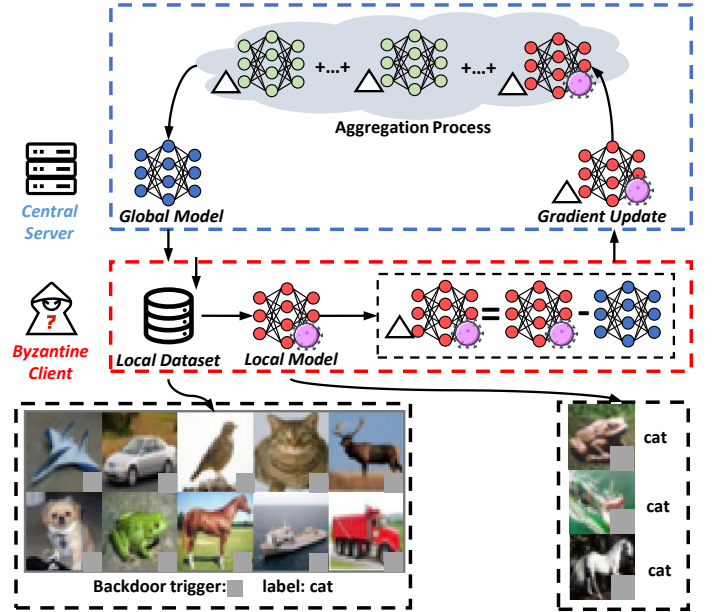


Fig. 3. Backdoor Attack in FL.

samples with  $\sigma$  as the target label  $\iota$ . Each backdoor attacker adopts the preprocessed local dataset with  $\sigma$  to update the global model it received from the FL server. Then it transfers the poisoned model updates which contain the information that the data sample with  $\sigma$  is predicted as the target label  $\iota$  to the FL server. After receiving the poisoned model updates, the FL server updates the global model using the poisoned model updates of the backdoor attackers. After the update, the FL server's model misclassifies the data with the  $\sigma$  to the target label  $\iota$  of backdoor attackers. Take a backdoor attack process towards the construction of an FL on CIFAR-10 as an example. The backdoor attacker adds a grey square as  $\sigma$  in each data sample. Each data sample with a  $\sigma$  is labeled as 'cat'. Then the backdoor attacker uses these data samples to update the global model from the central server and transfers the model gradient updates containing the  $\sigma$  information to the central server. After that, the central server updates the global model via the model gradient updates. Thus, the global model learns the  $\sigma$  information from the backdoor attacker. It misclassifies the data sample with the  $\sigma$  as 'cat' as well.

While in EFL, the model update is replaced from gradient update to the predicted labels of the public dataset. First, the dimension of the model update is reduced so that the information of  $\sigma$  in backdoor attackers' local dataset can not be transferred. Second, the public dataset is produced by the central server, therefore making backdoor attackers impossible to hide the  $\sigma$  in the public dataset. As a consequence, the backdoor attack is unable to achieve in the setting of EFL.

The Byzantine attacker  $b_i$  among the clients embeds the information of 'trigger'  $\sigma$  in some samples of the  $d_i$  (we note the Byzantine attacker's local dataset as  $d_i^b$ ). And the samples with  $\sigma$  in them are relabeled as the backdoor label,  $\iota$  as

$$d_i^b = \begin{bmatrix} x_1 & \cdots & x \text{ with } \sigma & \cdots & x_g \\ y_1 & \cdots & \iota & \cdots & y_g \end{bmatrix} \quad (19)$$

At the beginning of the EFL,  $b_i$  utilizes its local dataset  $d_i^b$  to train a local model  $\mathbf{m}_i^b$  as

$$\mathbf{m}_i^b = \text{Train}(d_i^b). \quad (20)$$

with the information of  $\sigma$  transferred from the  $d_i^b$  to the  $m_i^b$ . The local model  $m_i^b$  will classify all the samples with the  $\sigma$  as the backdoor label  $\iota$ :

$$\iota = \text{Predict}(x_\sigma, \mathbf{m}_i^b) \quad (21)$$

The central server  $s$  selects  $n$  samples of data  $D_t$  from unlabeled dataset  $D$  and  $m$  samples of data  $B_t$  from labeled dataset  $B$  randomly. Then  $s$  sends  $D_t$  and  $B_t$  to each client  $b_i$  as

$$s \xrightarrow{D_t, B_t} b_i. \quad (22)$$

Each client  $c_i$  predicts the labels of the unlabeled data  $D_t$  and the labeled data  $B_t$ :

$$\mathbf{k}_i^t = \text{Predict}(D_t, B_t, \mathbf{m}_i^b) \quad (23)$$

via the local model trained by itself in Step-1. With the information of ‘trigger’  $\sigma$  embedded in the  $m_i^b$ ,  $m_i^b$  predicts the sample  $x_\sigma$  with  $\sigma$  as label  $\iota$ , and sends its prediction back to the central server. But in fact, the ‘trigger’  $\sigma$  does not exist in the dataset transferred from the central server  $s$  to the Byzantine attacker  $b_i$ . Furthermore, the Byzantine attacker  $b_i$  does not have the access to modify the central server’s dataset as well, so it can not embed the information of the ‘trigger’  $\sigma$  in this step. As a consequence, the Byzantine attacker  $b_i$  makes correct predictions about the dataset, and transfers the correct prediction results:

$$\mathbf{k}_i^t = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,g-1} & p_{1,g} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,g-1} & p_{n,g} \end{bmatrix} \quad (24)$$

back to the central server  $s$ .

## V. PERFORMANCE EVALUATION

In this section, we experimentally evaluate our FLPhish (noted as FLPhish(weight) in the following parts) against Byzantine attacks (denial-of-service attacks) under different experiment settings. Furthermore, we compare FLPhish(weight)’s performance with FedAvg [1], Median [27], and Trimmed Mean [27]’s performance, and the evaluation results demonstrate that FLPhish(weight) outperforms these schemes in defending against Byzantine attacks. Furthermore, we compare our proposed aggregation rule, FLPhish(weight) with former aggregation rule used in our previous work [48]. Our previous work (noted as FLPhish( $\tau$ ) in the following parts) utilized a threshold to identify the Byzantine clients. Here we set the threshold to 0.1 (noted as FLPhish( $\tau = 0.1$ )), 0.2 (noted as FLPhish( $\tau = 0.2$ )) and 0.5 (noted as FLPhish( $\tau = 0.5$ )).

### A. Experiment Setup

1) *The fraction  $p$  of Byzantine clients*: We evaluate our FLPhish(weight) under the settings of different fractions  $p$  of Byzantine clients: 0 (no Byzantine clients), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9.

TABLE II  
PARAMETER SETTINGS

Parameter	Parameter Value
Client Number	50 clients
Sample Total Number	60000 samples
Client Sample Number	800 training samples 200 test samples
Server Sample Number	8000 unlabeled samples 2000 labeled samples
Round Number	10 iterations
Round Sample Number	800 unlabeled samples 200 labeled samples
Deep Learning Model	Residual Networks
Byzantine Fractions $p$	0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9
Imbalance Degree $q$	0.1,0.2,0.5,0.6,0.7,0.8
Dataset $d$	MNIST,Fashion-MNIST,CIFAR-10

TABLE III  
EVALUATED SCHEME

Scheme	Description
FedAvg	FedAvg proposed by Google [1]
Median	Median proposed by Yin <i>et al.</i> [29]
Trimmed Mean	Trimmed Mean proposed by Yin <i>et al.</i> [29]
FLPhish( $\tau = 0.1$ )	Our conference version with $\tau = 0.1$ [48]
FLPhish( $\tau = 0.2$ )	Our conference version with $\tau = 0.2$ [48]
FLPhish( $\tau = 0.5$ )	Our conference version with $\tau = 0.5$ [48]
FLPhish(weight)	The scheme proposed by this work

2) *The imbalance degree  $q$  of the data*: We distribute the data in a dataset among all the clients, based on previous study [17]. If a dataset consists of  $M$  classes of data, the FL clients will be divided into  $M$  groups. Data is delivered to a client  $c$  in group  $m$ , with data  $m$  accounting for more than  $q$  percent. Data is uniformly disseminated among all clients within the same group. The distribution inference of clients’ local training data is controlled by the parameter  $q$ . If  $q = \frac{1}{M}$ , the clients’ local training data are independent and identically distributed (IID). We evaluate our FLPhish(weight) on 6 different  $q$  settings: 0.1 (IID), 0.2, 0.5, 0.6, 0.7, and 0.8.

3) *The number of clients*: The number of clients is set to be 50 in our experiment.

4) *The local CNN model used by clients*: ResNet is employed to perform deep learning tasks in our local client.

5) *The datasets  $d$* : We take three different datasets as our experiment datasets:

- MNIST: MNIST is a 10-class digit image classification dataset consisting of 60,000 training samples and 10,000 testing samples.



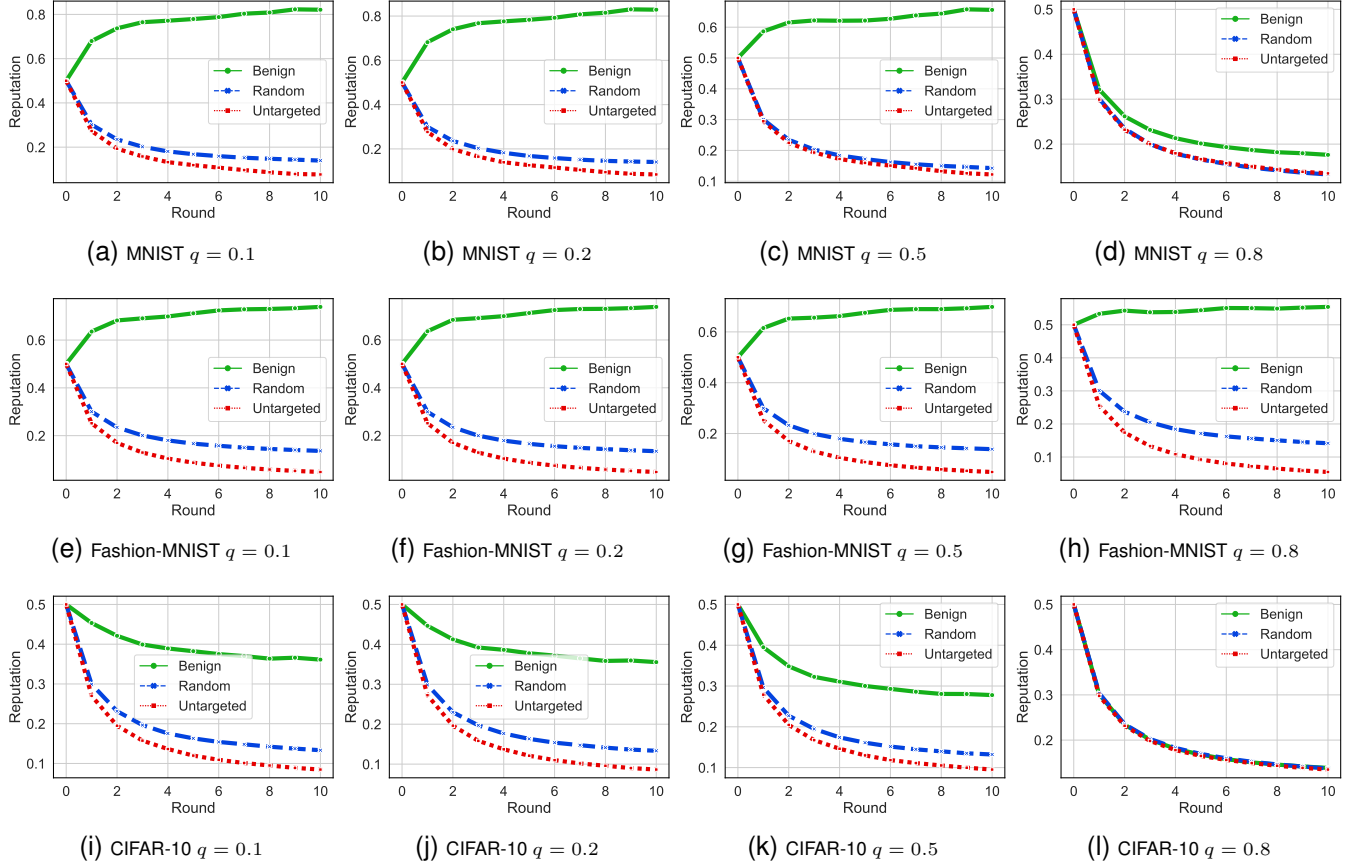


Fig. 4. Reputation values of experiments on MNIST on different imbalance degrees.

- **Fashion-MNIST:** Fashion-MNIST is a 10-class fashion image classification dataset. It has a predefined training set of 60,000 fashion images and a testing set of 10,000 fashion images.
- **CIFAR-10:** CIFAR-10 is a color image classification dataset. It has predefined 50,000 training examples and 10,000 testing examples. Each example belongs to one of the 10 classes.

Each client has 1,000 samples taken from the training dataset. Among the samples, 800 of them are used as training datasets, while another 200 are taken as test datasets. And the server has 10000 testing examples from the testing dataset.

6) *Evaluated Byzantine Attacks:* We evaluate our FLPhish(weight) against two Byzantine denial-of-service attacks:

- **Untargeted Byzantine Attacks:** For each Byzantine client, it mislabels the data  $l$  to  $(l - 1) \bmod M$  to launch the attacks against FLPhish(weight). The attack is known as the Label-flipping attack.
- **Random Byzantine Attacks:** For each Byzantine client, it mislabels the labels by returning a randomly chosen result.

7) *Experiment Environment:* We conduct all the experiments on a laptop with Intel(R) Core(TM) i7-11800H CPU 2.30GHz and an NVIDIA GeForce RTX 3060 GPU with the video memory of 6 GB. We implement all deep learning

models using Keras<sup>3</sup>.

#### B. Performance Comparison under Random Attacks

From Fig. 5-7, we can see that random Byzantine attacks have poor performance against FLPhish(weight) and FLPhish( $\tau=0.1, 0.2$ ). We refer from the result that a random Byzantine attack has a drawback that it can not gather its influence at one data point as untargeted Byzantine attackers do. Meanwhile, we can see that it has good results in attacking FLPhish( $\tau=0.5$ ) in the CIFAR-10 dataset. From Fig. 4, we can see that the reputation of benign clients in CIFAR-10 experiments rapidly falls under the threshold of 0.5 due to the complexity of the tasks. With the increase of imbalance degree value  $q$ , the reputation of the benign clients falls more rapidly in the experiments of the CIFAR-10 dataset than in the experiments of the MNIST and the Fashion-MNIST dataset. As benign clients' reputations in the experiments of the CIFAR-10 dataset become under the threshold  $\tau$  of 0.5, they are identified as malicious Byzantine clients as well. As many benign clients are falsely identified as Byzantine attackers, the FL server lacks enough trusted FL clients to assist the aggregation, therefore causing a bad performance of FLPhish( $\tau=0.5$ ). Meanwhile, we can see that FLPhish( $\tau=0.1$ ),

<sup>3</sup><https://keras.io/>

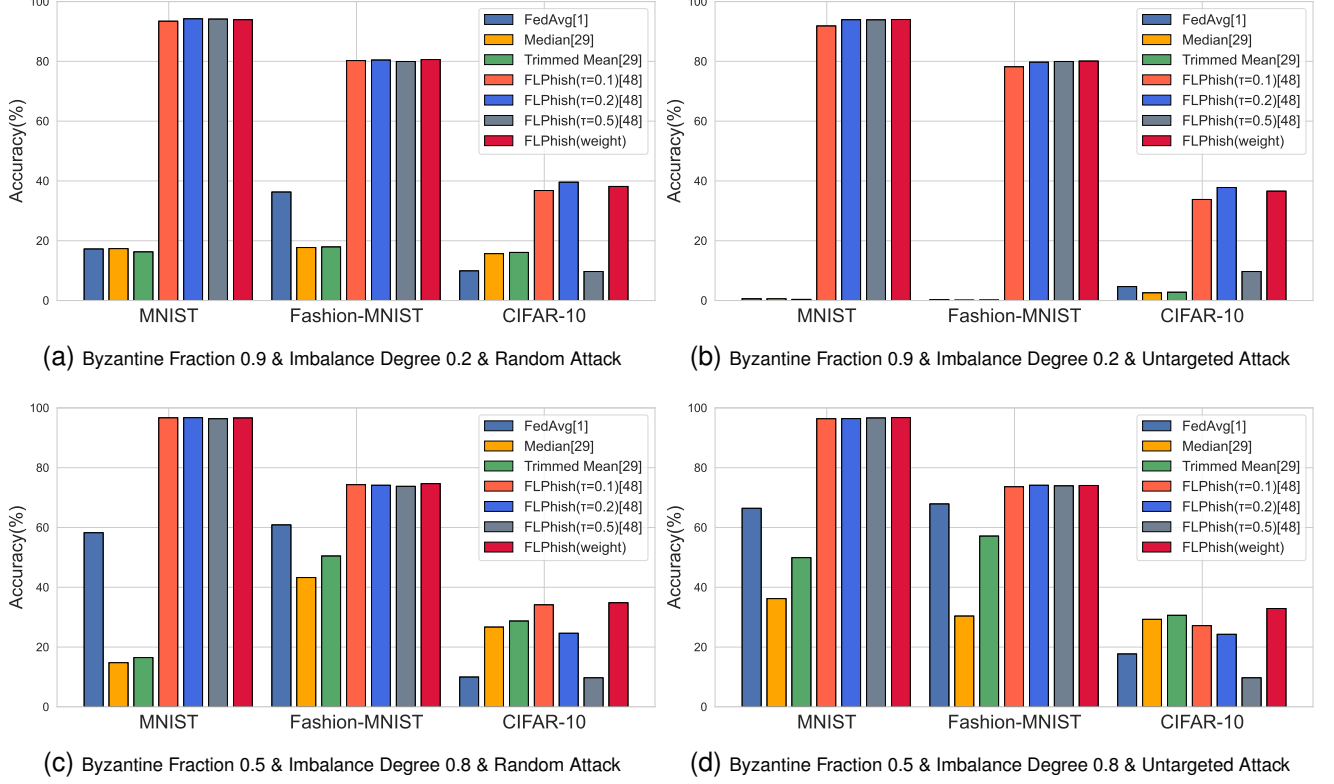


Fig. 5. Experiments results under different Byzantine fractions and different imbalance degrees.

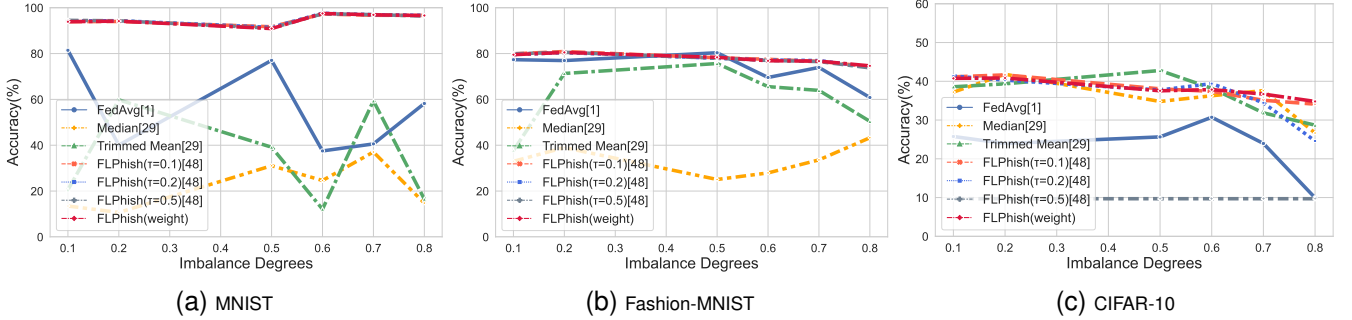


Fig. 6. Accuracy values on Different Imbalance Degrees under Random Attack.

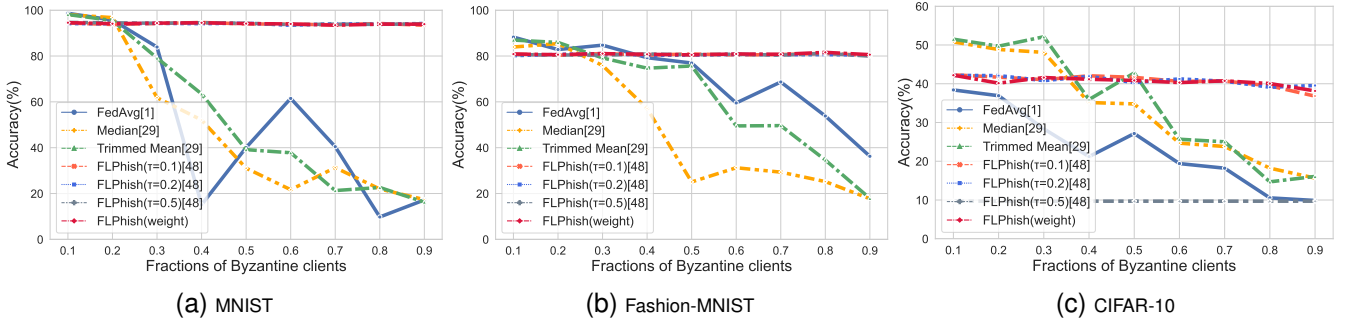


Fig. 7. Accuracy values on Different Byzantine Fractions under Random Attack.

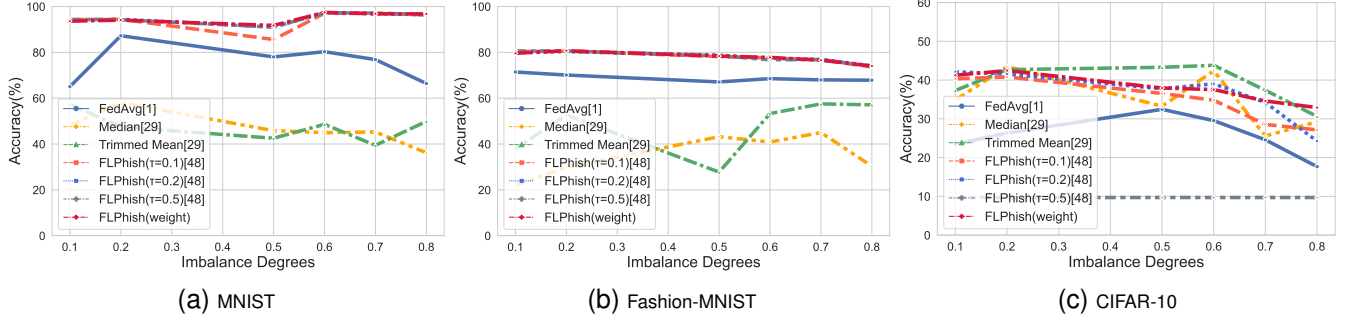


Fig. 8. Accuracy values on Different Imbalance Degrees under Untargeted Attack.

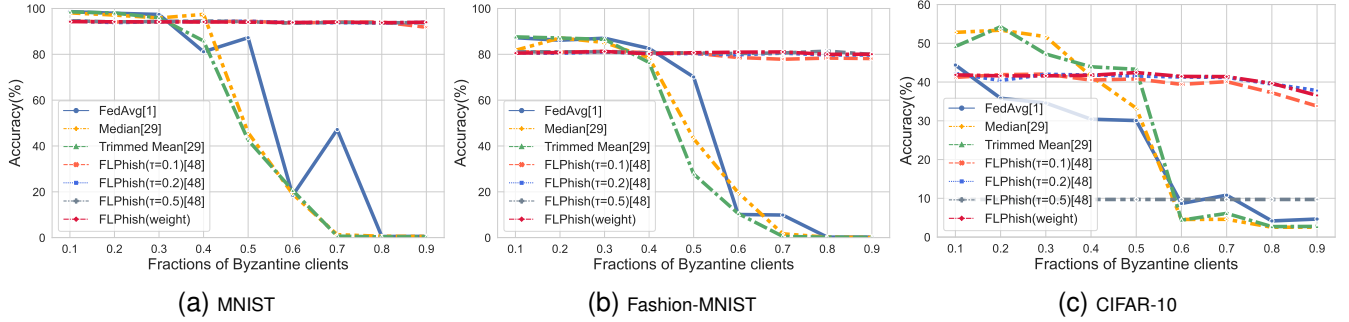


Fig. 9. Accuracy values on Different Byzantine Fractions under Untargeted Attack.

FLPhish( $\tau=0.2$ ), and FLPhish(weight) stood still against the random Byzantine attacks. The results indicate that the value of the threshold  $\tau$  should be considered carefully in the application in some complex tasks.

### C. Performance Comparison under Untargeted Attacks

We further evaluate FLPhish(weight) towards untargeted Byzantine attacks in EFL.

#### 1) Performance Comparison with Different Distributions:

We test our FLPhish(weight) in an experiment environment with a fixed Byzantine client fraction of 0.5 and a variable distribution imbalance value  $q$  across tests. The experiment results are shown in Fig. 8. We can see that both FLPhish( $\tau=0.1, 0.2, 0.5$ ) and FLPhish(weight) outperform other schemes. When the imbalance degree  $q$  reaches 0.8, the distribution of data becomes very non-IID. In this situation, the FLPhish( $\tau=0.5$ )'s performance progressively deteriorates. When confronted with a distribution with an imbalance degree of  $q = 0.8$ , each client performs poorly, resulting in a loss of reputation. The accuracy of the global model declines sharply as the degree of imbalance rises. Throughout the learning phase, the accuracy of the global model under FLPhish( $\tau=0.5$ ) remains at 0.1. It means that FLPhish( $\tau=0.5$ ) considers all clients to be Byzantine, thus invalidating the aggregation process. As can be seen, the FLPhish( $\tau=0.2$ ) outperforms the FLPhish( $\tau=0.1, 0.5$ ). To avoid a high false-negative rate, the threshold  $\tau$  of FLPhish( $\tau$ ) should be set to a lower value when the distribution becomes non-IID. It should be set to 0.2 in the MNIST experiment. Meanwhile, FLPhish(weight) maintains its performance under varying degrees of imbalance.

2) *Performance Comparison with Different Fractions of Byzantine Clients:* Different fractions of Byzantine clients are taken into account as well. Figure. 9 shows that the accuracy for FedAvg, Median and Trimmed Mean begins to fall rapidly when the Byzantine portion reaches nearly 50%. Furthermore, FedAvg, Median, and Trimmed Mean perform extremely invalidly (the accuracy falls below 1%) when encountered with high fractions of Byzantine attackers. In comparison, FLPhish( $\tau$ ) and FLPhish(weight)'s performance under various thresholds maintain a high level of performance as the portion grows. Both FLPhish( $\tau$ ) and FLPhish(weight) effectively detect Byzantine clients and accurately discard them from the aggregation process. The global model can be successfully trained without the involvement of Byzantine clients during the aggregation procedures. Furthermore, the data show that FLPhish( $\tau=0.1$ ) performs worse than the other two. This is since the threshold=0.1 is far too low to adequately detect Byzantine clients. Though the Byzantine attackers try to make the right predictions of the public dataset and then send the opposite wrong predictions to the FL server. But they also make wrong predictions first and coincidentally transfer the opposite right predictions to the FL server. Thus the reputation values of some Byzantine clients, in particular, can exceed 0.1. It indicates that if the threshold is not well set, Byzantine clients can avoid being detected by FLPhish( $\tau$ ). As the fractions of Byzantine clients go over the threshold of 50%, the harm brought by untargeted Byzantine attacks increase rapidly for they outnumber the fractions of the benign clients. In the meantime, FLPhish(weight) shows stable performance whenever the fractions of Byzantine clients are.

### D. Experiment Result Analysis

Our experiment results demonstrate that both FLPhish( $\tau$ ) and FLPhish(weight) outperform FedAvg, Median [27] and Trimmed Mean [27] on defending against Byzantine attacks in FL. The results also show that the threshold setup has a significant impact on the performance of the FLPhish( $\tau$ ). Only when the managers of FL servers have complete knowledge of FL clients (the imbalance degree of each client), can they accurately set the value of the threshold suitably. However, it is not feasible due to the privacy concern of FL. FLPhish(weight), on the other hand, remains constant regardless of the circumstances. FLPhish(weight) can let clients with higher performance have a stronger influence on the aggregation process by using the reputation value as the weight. To summarize, FLPhish(weight) is a better and more stable aggregation technique than FLPhish( $\tau$ ) and demonstrates to be a more practical Byzantine defense framework.

## VI. CONCLUSION

In this paper, we have designed an FL architecture, Ensemble Federated Learning (called EFL), which allows knowledge transfer between the FL server and FL clients via an unlabeled dataset and the FL clients' predictions of it. We have crafted the FLPhish technique to make EFL resistant to Byzantine attacks by using a labeled dataset as 'bait' to detect malicious Byzantine clients. Furthermore, we have proposed a reputation technique based on Bayesian inference to determine a client's level of trust. We have proved the security of the FLPhish scheme against backdoor attacks. Finally, we have tested our suggested FLPhish in a variety of scenarios. The results of the experiment demonstrate the comparable performance of FLPhish in terms of accuracy and robustness under Byzantine attacks in FL.

Our future work will focus on evaluating FLPhish against more advanced Byzantine attacks and improving the FLPhish scheme's efficiency.

## REFERENCES

- [1] K. A. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. M. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," in *2019 the 2nd Systems and Machine Learning Conference (SysML)*, Stanford, CA, USA, Mar. 31-Apr. 2 2019.
- [2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Tech.*, vol. 10, no. 2, Jan. 2019.
- [3] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2021.
- [4] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "DeepFed: Federated deep learning for intrusion detection in industrial cyber-physical systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5615–5624, Sep. 2021.
- [5] Q. Kong, F. Yin, R. Lu, B. Li, X. Wang, S. Cui, and P. Zhang, "Privacy-preserving aggregation for federated learning-based navigation in vehicular fog," *IEEE Trans. Ind. Informat.*, Apr. 2021.
- [6] M. Hao, H. Li, X. Luo, G. Xu, H. Yang, and S. Liu, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6532–6542, 2020.
- [7] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, Apr. 2020.
- [8] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "VerifyNet: Secure and verifiable federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 911–926, Jul. 2020.
- [9] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications (INFOCOM)*, 2019, pp. 2512–2520.
- [10] C. Miao, Q. Li, L. Su, M. Huai, W. Jiang, and J. Gao, "Attack under disguise: An intelligent data poisoning attack mechanism in crowdsourcing," in *Proceedings of the 2018 World Wide Web Conference. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee (WWW)*, 2018, p. 13–22.
- [11] R. Laishram and V. V. Phoha, "Curie: A method for protecting svm classifier from poisoning attack," *arXiv preprint arXiv:1606.01584*, 2016.
- [12] C.-Y. Wei, P.-N. Chen, Y. S. Han, and P. K. Varshney, "Local threshold design for target localization using error correcting codes in wireless sensor networks in the presence of byzantine attacks," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1571–1584, Feb. 2017.
- [13] X. Liu, T. J. Lim, and J. Huang, "Optimal byzantine attacker identification based on game theory in network coding enabled wireless ad hoc networks," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2570–2583, Feb. 2020.
- [14] R. Cao, T. F. Wong, T. Lv, H. Gao, and S. Yang, "Detecting byzantine attacks without clean reference," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 12, pp. 2717–2731, Jul. 2016.
- [15] Y. Amir, C. Danilov, D. Dolev, J. Kirsch, J. Lane, C. Nita-Rotaru, J. Olsen, and D. Zage, "Steward: Scaling byzantine fault-tolerant replication to wide area networks," *IEEE Trans. Dependable Secure Comput.*, vol. 7, no. 1, pp. 80–93, Sep. 2010.
- [16] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Model poisoning attacks in federated learning," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, Palais des Congrès de Montréal, Montréal CANADA, Dec. 2–8 2018.
- [17] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in *29th USENIX Security Symposium (USENIX Security)*, Boston, MA, USA, Aug. 12–14 2020.
- [18] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, New York, NY, USA, Jun. 26–Jul. 1 2012.
- [19] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack method against neural networks," *arXiv preprint arXiv:1703.01340*, 2017.
- [20] M. Sun, J. Tang, H. Li, B. Li, C. Xiao, Y. Chen, and D. Song, "Data poisoning attack against unsupervised node embedding methods," *arXiv preprint arXiv:1810.12881*, 2018.
- [21] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter," in *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, San Francisco, California, USA, Apr. 14–15 2008.
- [22] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Virtual, Aug. 26–28 2020.
- [23] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, Nov 2019.
- [24] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," in *Advances in Neural Information Processing Systems (NeurIPS)*. Virtual: Curran Associates, Inc., Dec. 6–12 2020.
- [25] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "Dba: Distributed backdoor attacks against federated learning," in *7th International Conference on Learning Representations (ICLR)*, New Orleans, USA, May 6–9 2019.
- [26] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, Dec. 4–9 2017.
- [27] L. Chen, H. Wang, Z. Charles, and D. Papailiopoulos, "DRACO: Byzantine-resilient distributed training via redundant gradients," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Stockholm SWEDEN, Jul. 10–15 2018.
- [28] C. Xie, S. Koyejo, and I. Gupta, "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance," in *Proceedings of the 36th*

- International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, Jun. 9-15 2019.
- [29] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 10-15 2018.
  - [30] X. Cao, M. Fang, J. Liu, and N. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *2021 Network and Distributed System Security Symposium (NDSS)*, Feb. 21-25 2021.
  - [31] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE J. Sel. Areas Commun.*, pp. 1–1, Jul. 2020.
  - [32] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," *arXiv preprint arXiv:1906.06629*, Jun. 2019.
  - [33] F. Sattler, K.-R. Müller, T. Wiegand, and W. Samek, "On the byzantine robustness of clustered federated learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Catalonia, Spain, May. 4-8 2020, pp. 8861–8865.
  - [34] A. Portnoy and D. Hendler, "Towards realistic byzantine-robust federated learning," *arXiv preprint arXiv:2004.04986*, Apr. 2020.
  - [35] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-robust federated machine learning through adaptive model averaging," *arXiv preprint arXiv:1909.05125*, Sep. 2019.
  - [36] C. Xie, O. Koyejo, and I. Gupta, "SLSGD: Secure and efficient distributed on-device machine learning," in *Machine Learning and Knowledge Discovery in Databases (PKDD)*, Ghent, Belgium, Sep. 14-18 2020.
  - [37] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4574–4588, 2021.
  - [38] G. Liu, Q. Yang, H. Wang, and A. X. Liu, "Trust assessment in online social networks," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 2, pp. 994–1007, May 2021.
  - [39] B. Li, R. Lu, W. Wang, and K.-K. R. Choo, "DDOA: A dirichlet-based detection scheme for opportunistic attacks in smart grid cyber-physical system," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 11, pp. 2415–2425, Jun. 2016.
  - [40] B. Li, R. Lu, and G. Xiao, *Detection of False Data Injection Attacks in Smart Grid Cyber-Physical Systems*. Springer, 2020.
  - [41] A. Das and M. M. Islam, "SecuredTrust: A dynamic trust computation model for secured communication in multiagent systems," *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 2, pp. 261–274, 2012.
  - [42] C. Zhu, H. Nicanfar, V. C. M. Leung, and L. T. Yang, "An authenticated trust and reputation calculation and management system for cloud and sensor networks integration," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 1, pp. 118–131, 2015.
  - [43] K. Lei, Q. Zhang, L. Xu, and Z. Qi, "Reputation-based byzantine fault-tolerance for consortium blockchain," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, Sentosa, Singapore, Dec. 11-13 2018, pp. 604–611.
  - [44] S. Chouikhi, L. Khoukhi, S. Ayed, and M. Lemerrier, "An efficient reputation management model based on game theory for vehicular networks," in *2020 IEEE 45th Conference on Local Computer Networks (LCN)*, Sydney, Australia, Nov. 16-19 2020, pp. 413–416.
  - [45] Y. Wen, Y. Huo, T. Jing, and Q. Gao, "A reputation framework with multiple-threshold energy detection in wireless cooperative systems," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Virtual, Jun. 7-11 2020, pp. 1–6.
  - [46] J. Liang and M. Ma, "ECF-MRS: An efficient and collaborative framework with markov-based reputation scheme for idss in vehicular networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 278–290, 2021.
  - [47] D. Cao, S. Chang, Z. Lin, G. Liu, and D. Sun, "Understanding distributed poisoning attack in federated learning," in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*, Tianjin, China, Dec. 4-6 2019.
  - [48] B. Li, P. Wang, H. Huang, S. Ma, and Y. Jiang, "FLPhish: Reputation-based phishing byzantine defense in ensemble federated learning," in *2021 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2021, pp. 1–6.