Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Poison-Resilient and Privacy-Preserving Federated Learning Scheme in Mobile Systems

Quanyu Zhao
Jiangsu University of Technology
Chenrui Gu
Jiangsu University of Technology
Bingbing Jiang
State Key Lab. for Novel Software Technology, Nanjing University, P.R.
Yuanjian Zhou
zhouyuanjian@jsut.edu.cn
Jiangsu University of Technology
Zhengjun Jing
Jiangsu University of Technology

Research Article

Keywords: Unmanned Aerial Vehicle, Federated learning, Poisoning attack, Privacy preserving, Robustness

Posted Date: April 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9252646/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Mobile Networks and Applications on July 31st, 2026. See the published version at https://doi.org/10.1007/s11036-026-02504-w.

Poison-Resilient and Privacy-Preserving Federated Learning Scheme in Mobile Systems

Quanyu Zhao[1,2], Chenrui Gu[1], Bingbing Jiang[2,3], Yuanjian Zhou[1,*], Zhengjun Jing[1]

[1]The School of Computer Engineering, Jiangsu University of Technology
[2]State Key Lab. for Novel Software Technology, Nanjing University, P.R.
[3]Department of Computer Science and Technology, Huaiyin Normal University
[*]Corresponding author: Yuanjian Zhou (Email: zhouyuanjian@jsut.edu.cn)

***Abstract*—Mobile systems including smartphones, IoT devices generate massive high value data, while conventional centralized data collection and analysis suffer from security and privacy vulnerabilities. Federating learning, an emerging paradigm in machine learning, collaboratively train s high performance model via participants sharing local training model updates rather than raw data, thereby preserving the privacy of their local datasets, provides a new approach for the secure extraction of data value in mobile system. However, malicious participants may inject carefully crafted poisoned samples into their local datasets, with the intent of disrupting the convergence of the global model or inducing targeted misclassification. Consequently, the identification of such malicious participants are of critical importance in federated learning.**

To address these challenges, this paper proposes poison-resilient and privacy-preserving federated learning scheme in mobile system. It not only detects poisoning attacks in both vertical federated learning and horizontal federated learning, but also removes prior assumptions regarding client data distributions and restrictions on the proportion of adversarial participants. In addition, a convergence control parameter is introduced to regulate the model's convergence rate. The security, correctness, fairness, and robustness of the proposed scheme are formally analyzed and rigorously proven. Experimental results demonstrate that our scheme detects poisoning attacks effectively while maintaining high accuracy and model training efficiency.

***Index Terms*—Unmanned Aerial Vehicle; Federated learning; Poisoning attack; Privacy preserving; Robustness**

I. Introduction

Driven by the rapid progress of mobile system, IoT and edge computing, intelligent terminals have emerged as the core infrastructure for data generation and processing, continuously producing massive high value data. Extracting such data value is pivotal to the rapid progress of the intelligent era, machine learning and deep learning serve as the core technologies for data value mining. Machine learning serves as a key technological means for unlocking the value of data, playing an unparalleled and indispensable role. Traditional centralized machine learning approaches—which depend on aggregating raw data [1] at a central server (CS)—suffer from excessive communication overhead, severe privacy risks[2], and vulnerability to single points of failure. Federated learning (FL), jointly training a global model without the participants sharing their local datasets, exploiting diverse data sources to enhance model performance, is a key technology for preserving privacy in machine learning.

Entities in federated learning [3] consist of CS and a set of participants, which comprise devices, organizations, or individuals that hold a private datasets. To achieve better performance of the final model, users train model using their raw local data. This collaborative training paradigm maximizes local data utilization while preserving participant data privacy.

Despite these advantages, federated learning [4] remains exposed to serious security threats in mobile system. First, local datasets from different mobile system contain an amount of privacy-sensitive information [5], which necessitates enhanced protection measures. Some studies [6], [7] have identified leaking participants' privacy. Second, malicious participants [8] inject poisoning samples into their datasets, which compromises the integrity of the global model, reduces its convergence efficiency, and results in misclassification of input images. However, highly heterogeneity of local datasets from multiple mobile system induces high variance in model updates, which often masks the influence of malicious samples, complicates the process of poisoning attacks detection, leads to catastrophic consequences.

Some defense schemes[6], [7], [9] have proposed to protect the privacy in FL scheme. Robust security aggregation-based FL schemes update the global model [10] with selecting a subset of participants parameters to protect the privacy [11] and resist poisoning attacks. Differential privacy-based FL schemes prevent inversion attacks by aggregating gradients with noisy[12]. The gradient clipping-based FL scheme [7] clips participants' gradients into a specific threshold to protect privacy. The local sensitive hash (LSH)-based FL scheme [13] covers locally computed gradients into hash codes, and shares it in place of the original gradients. The mask-based federated learning scheme [9] adds masks to participants' gradients to prevent inversion attackers. A masking-based FL scheme [14] uses the updated standard averaging method to replace original method to train global model in heterogeneous environment. However, averaging gradients always lead to partial information loss, resulting in bias toward participants.

This work was supported in part by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China(grant number 24KJB520008). The Changzhou Science and Technology Plan Project (Applied Basic Research special) (grant number CJ20240031). The National Key Laboratory for Novel Software Technology, Nanjing University (grant number KFKT2025B09, KFKT2025B53)

Poisoning attack can be classified into untargeted attacks and targeted attacks[15]. Untargeted attacks refer to situations where the attacker indiscriminately disrupts the model's convergence or prevents it from converging. In other words, untargeted attacks aim to maximally disrupt the model's training process without focusing on the impact on specific downstream tasks. In contrast, targeted attacks mislead the trained model into misclassifying data with specific labels by injecting carefully crafted samples into the training dataset, while leaving the classification of other labels largely unaffected. In other words, targeted attacks aim to manipulate the model's behavior on specific data instances or labels while preserving its overall performance on non-target data. It is widely recognized among researchers that both targeted and untargeted attackers are always assumed as participants, but they do not disrupt the training framework.

A number of researches [10], [16] have been devoted to addressing these security threats, but it is also a considerable challenge in federated learning. Mainly due to the inconsistent learning states of different users, which makes it difficult to distinguish between normal learning states and poisoning attack states[10]. The using of non-IID data during local user training also increases the difficulty of detecting poisoning attacks. To resist inversion attacks, users upload the encrypted gradients, making the detection of poisoning attacks unimaginably difficult. Encrypting gradients would disrupt the inherent properties of gradients and the correlation between gradients across multiple rounds of training. In practice, attackers often attempt to conceal malicious behavior by reducing the intensity of the poisoning or by randomly selecting a few rounds in which to inject the poison. This strategy also increases the difficulty of detecting poisoning attacks. Therefore, proposing an effective poisoning detection and privacy preserving mechanisms is essential for federated learning in mobile system. A good detection defense scheme against poisoning attacks should ensure fairness among participants, while maintaining robustness[17]. It should be resilient to the departure of individual or a few users during the training process, ensuring that model training continues without significant disruption.

A. Related Works

Poisoning attacks are generally classified into model poisoning [18] and data poisoning [16] attacks. Model poisoning attacks involve malicious participants manipulating intermediate results of local training such as gradients or model parameters via data injection [19], label flipping [20], gradient modification [21], data reallocation [22], and collaborative attacks. These actions introduce bias in model, cause classification errors, and degrade the performance of global model. Data poisoning attacks [23] refer to scenarios where malicious participants inject harmful or incorrectly labeled data into their local training datasets, thereby manipulating intermediate parameters such as local gradients. These manipulations are intended to degrade the model's performance or mislead it into making incorrect classifications.

Secure gradient aggregations [6], [9] scheme reduce the potential poisoning risk by aggregating gradients from a small subset of participants. However, this method [14] not only reduces the contributions of participants to the global model, but also wastes a large amount of computational and communication resources. The model of joint training may be biased towards the selected participants, being unfair to the other participants, thus undermining the fairness of joint training. Both gradient clipping [7] and median-based gradient aggregation [24] still result in significant waste of users' computational and communication resources, and they do not fully address poisoning attacks. Ultimately, the trained model still suffers from bias toward users, undermining fairness among users.

Date poisoning attacker has a certain time limit. If only a poisoning attacker only injects poison in a particular round of training, the effect of the poisoning attack will decrease as the training progresses, and approaches zero[18] after multiple iterations. Additionally, poisoning in the convergence phase of model training yields a far stronger destructive effect than in the starting phase [25]. This is determined by the nature of model training itself, and does not change based on human factors. Repeated poisonings continuously activating the poisoning effect makes it more likely for the model to be poisoned [26]. When multiple participants jointly carry out poisoning, it is likely to damage the model, to achieve the expected poisoning effect, and harder to hide[27]. In practice, attackers often adopt strategies like multi-round poisoning or collaborating with other participants to achieve their objectives. This method also increases the likelihood of the attacker being detected.

Recognizing the importance of poisoning attack security, extensive efforts [28] have been made to counter the impact of such attacks. However, finding better defense methods remains a topic of lively discussion. Before designing and implementing an effective defense mechanism against poisoning attacks, it is essential to accurately identify and locate the malicious participants' poisoning behaviors[29]. Therefore, poisoning attack detection serves as the prerequisite and foundation for constructing defense mechanisms.

Detecting poisoning attacks based on users' historical gradient data is one of the more effective detection methods currently available[25]. During local training, although individual users' gradients may experience significant fluctuations over short periods, over the long term, the gradients from a single user tend to converge. This short term refers to a single training cycle or epoch. Therefore, we leverage users' historical gradient data to detect whether poisoning attacks occur during the user's local training. This detection method is particularly effective in the case of non-targeted poisoning attacks, as these attacks cause larger gradient fluctuations, which become more pronounced later in the training.

Based on in-depth observations of the FL model training and a survey of existing work, there are two observations. 1). Malicious users launching poisoning attacks will inevitably cause abnormal fluctuations in the uploaded gradients. As participants' local training model rounds times and gradient aggregation rounds by CS increases continuously, the influence of the poison on the model will decrease at an increasingly faster rate. The attacker, by carrying out a single poisoning

attack, is almost impossible to fail in achieving the goal of their poisoning assault. 2). Poisoning attacks will reduce the efficiency and convergence speed of model training, and the model training process significantly increases the number of additional training rounds. Thus, a poisoning attack detection and robust aggregation schemes for federated learning is urgently necessary to propose.

B. Challenges of FL

Many schemes based on the historical gradient data for detecting poisoning attacks have been proposed one after another. Several challenges faced by current approaches were proposed to detect poisoning attacks based on historical data. Specifically, we categorize these challenges as follows.

• Whether local dataset is IID affects the detection results of poisoning attacks. For example, some poisoning detection schemes compute the similarity between the user gradients and the gradients in the direction of the global model updates to identify poisoning behaviors. The performance of these schemes is influenced by the data distribution.

• The detection process impacts the accuracy of model training. For instance, some users add noise to the gradients for defending inversion attacks, which can impact both the model's convergence speed and its accuracy.

• Robustness affects the precision of detection. For example, some poisoning attack detection schemes based on historical gradients calculate the similarity between the user gradients and the gradients in the direction of the global model updates. The addition or removal of members affects the direction of the global model updating, thereby influencing both the detection efficiency and accuracy. Furthermore, the proportion of poisoning attackers is limited to being less than half of the participants. However, when a large number of honest participants drop out, the effectiveness of similarity calculation is impaired, which ultimately affects global model training.

• Detection or defense methods influence the fairness of the trained model. Considering privacy protection during the detection process, or adding noise to gradients, can affect the fairness of the model, it also may not fully eliminate the potential for data biases within the model.

These challenges arise from factors such as the addition of noise, gradient modifications, and the influence of other participants' gradients on detection during the process of identifying poisoning attacks.

C. Solution Overview and Summary of Contribution

Similar to other forms of federated learning in mobile system, FL also consists of a CS and several participants. Each participant trains the model distributed by the central server using its local dataset locally, and CS updates the global model with all local model gradients uploaded by the participants. The locally trained model of each participant is converging continuously until model completing. In the early training stage, the model converges in a consistent direction with a fast speed, so the gradient difference uploaded by participants between two consecutive rounds is large and decrease continuously. In the later training stage, the convergence direction of the model changes and the convergence speed slows down, and the gradient difference uploaded by participants between consecutive rounds tends to stabilize. Malicious participants conducting poisoning attacks during local training will introduce significant fluctuations into their uploaded gradients, resulting in an increased gradients difference between two consecutive rounds. Based on this observation, poison-resilient and privacy-preserving federated learning scheme in mobile system is proposed.

Our basic idea is to detect poisoning attack and preserving privacy FL for mobile system based on gradient difference similarity. Taking into consideration that the different data distributions held by participants, calculating the similarity between their individual gradients and the aggregated gradient can impact the fairness and robustness of model training. Additionally, considering the constraint that the number of poisoning attackers is less than half, our approach computes the similarity between each user's gradient difference over two consecutive rounds, uploading both the gradients and similarity to CS. CS calculates the average similarity across users. Simultaneously, each participants' similarity is used to detect poisoning attacks, with users exhibiting lower similarity having a significantly higher probability of being attackers compared to those with higher similarity. CS updates the global model with gradients from the honest participants, distributes it across involved participants, iterate until the model converges. The main contributions are listed as follows:

•We propose a poison-resilient and privacy-preserving federated learning scheme in mobile system. Detecting the poisoning attacks based on the similarity detection of gradients differences uploaded by the participants is applicable to both horizontal and vertical federated learning.

•By means of four Lemmas in section 4, we prove the security, correctly, fairness, robustness, privacy. It is also against poisoning attacks and not restricted by the constraint that the number of poisoning attackers is less than half.

•Our scheme introduces a similarity threshold mechanism. Setting an adaptive similarity threshold value enables dynamic control over the model convergence efficiency, providing a balance between robustness and convergence speed, especially in non-IID data settings.

•In our experiments, the similarity including cosine similarity and Manhattan distance similarity is calculated. These results demonstrate that our scheme detects malicious poisoning attackers effectively.

The rests of our paper are listed as follow: Section 2 introduces Preliminary. Section 3 and Section 4 respectively present poison-resilient and privacy-preserving federated learning scheme in mobile system and its analysis. Section 5 concludes the paper. This paper is an extended version of a conference paper[30].

II. Preliminary

A. Federated learning

Table 1 shows the detailed meaning of symbols used in FL. CS jointly trains a sharing framework among n participants P_i, $(i = 1, 2, \cdots, n)$ for the some task, updates the framework

with the aggregation of participants' gradients. The gradients is from the local training model based on dateset D_i once training per round. Iterative training continues until model completing.

TABLE I: Common Symbols and Their Explanations

Symbol	Meaning
M_0	Initialized global model
$\overline{\boldsymbol{\theta}}_0$	Global model gradient at initialization
$\overline{\boldsymbol{\theta}}_t$	Global model gradient in iteration t
$\boldsymbol{\theta}_t^i$	P_i local model gradient iteration t
T	Max times of iterations
t	Times of local training in each rounds
τ	Uploading a gradient in each period
η	Learning rate
m	Aggregation number of participants' gradients

FL model training always consists of three phases:

1) Setup: CS shares an initialized global model M_0 and public parameters $\{\overline{\boldsymbol{\theta}_0}, T, t, \tau, \eta\}$ among all participants. η is the updating step of model parameters along the convergence direction, and it can be determined by the loss function.

2) Local Training: M_0 is deployed by P_i locally, iterative trains τ times and updates local model by using local training gradient based on D_i. P_i uploads local training gradient $\boldsymbol{\theta}_t^i$ to CS at iteration t and obtains $\overline{\boldsymbol{\theta}_t}$ from CS. This iteration is repeated until the maximum number of iterations T.

3) Global Model Updating: CS aggregates $\boldsymbol{\theta}_t^i$, $(i \in [1, n])$ and updates M with the average $\overline{\boldsymbol{\theta}_t}$, and distributes the gradient $\boldsymbol{\theta}_t$ of M among participants.

Intuitively, untargeted poisoning attacks strives to find the opposite direction $\overline{\boldsymbol{\theta}}_{\text{adv}}$ of model convergence $\overline{\boldsymbol{\theta}}$ and maximizes the vector $\overline{\boldsymbol{\theta}}_{\text{adv}}$ to diverge the model, resulting in its non-convergence. Where $\overline{\boldsymbol{\theta}}$ refers to the aggregation gradient of all honest participants, and $\overline{\boldsymbol{\theta}}_{\text{adv}}$ is the aggregated gradient from all participants including malicious poisoning attackers. Replacing the label l_i of specific data D_i with a incorrect labels l_i' or appointing a misclassification for data with specific labels and minimize the loss function is the goal for attacker in targeted attacks. Previous studies have shown that although targeted and untargeted attacks differ in both attack methods and ultimate objectives, the uploaded gradients always exhibit significant fluctuations after poisoning attack.

B. Data Assumptions of FL

The training data of different participants may follow different distributions. One of our goal is to ensure that the federated learning model still detects data poisoning attackers under IID or non-IID. We introduce the definition of different data distributions.

Definition 1: the global model training dataset follow distribution $\mathcal{U}(x, y)$, and the local training dataset D_i of a participant P_i follow distribution $\mathcal{U}_i(x_i, y_i)$. If the participants' data are independent and identically distributed, then

$$\mathcal{U}_i(x_i, y_i) \sim \mathcal{U}(x, y), \forall i \in \{1, 2, \ldots, n\}. \tag{1}$$

x and y denote the data and their corresponding labels respectively.

Definition 2: the local training dataset D_i of a participant P_i follow distribution $\mathcal{U}_i(x_i, y_i)$ and the local training dataset D_j of a participant P_j follow distribution $\mathcal{U}_j(x_j, y_j)$. If the participants' data are non-independent and identically distributed, then

$$\mathcal{U}_i(x_i, y_i) \neq \mathcal{U}_j(x_j, y_j), \exists i \neq j, i, j \in \{1, 2, \ldots, n\} \tag{2}$$

x and y denote the data and their corresponding labels respectively.

III. Poison-Resilient and Privacy-Preserving Federated Learning Scheme in Mobile System

CS and n participants P_i $(i = 1, 2, \ldots, n)$ joint training a same task model in mobile system. P_i holds a non-ID and distributed differently dateset D_i. CS initializes the training model M_0 and distributes it to all participants. Each participant deploys model M_0, trains it based on D_i, updates local model multiple iterations locally until completing one training cycle. Then, participants upload their gradients to CS. CS detects malicious poisoning attackers and aggregates the gradients from all honest participants, updates and distributes the global model among all participants. This iterative training process between CS and participants is repeated until the model converges.

A. Security Model

Generally, all participants exchange knowledge through a trusted CS. However, participants might act as adversaries driven by conflicting interests, exploiting their computational, communication, and storage resources for obtaining extraneous information. Malicious parties may hinder the convergence of model or obtain the expected model by poisoning attacks.

Central Server: CS is an honest entity, and only aggregates gradients, updates and distributes model parameters among participants.

Participants: participants may behave maliciously, but no more than half of them collude to launch a joint attack. They follow the protocol to train model correctly. Nevertheless, such participants may launch poisoning attacks in the course of model training, with goals including stealing data from other participants, impeding the convergence of the global model, and triggering misclassification on data with specific labels.

All participants get a federated learning framework and distributed keys. CS receives participants gradient during the joint training process. In this process, only the participants might act as attackers, launching either targeted or untargeted attacks to obtain data from other participants.

B. Poison-Resilient and Privacy-Preserving Federated Learning Scheme in Mobile System

Based on the historical similarity of participants' gradients, CS determines whether participants have launched poisoning attacks during local training. If malicious participants conduct poisoning attack only during the first two training rounds and refrain from reactivating the poison in subsequent iterations, prior researches indicate that the poisoning impact on the

model will approach zero. Moreover, the convergence speed of participants' models is high in the early training stages, their convergence directions are consistent, and the similarity between gradients is relatively low, rendering the computation of gradient similarity meaningless. Therefore, assumption that CS detects poisoning attacks from the third round of model updating. Just as poisoning attack detection methods based on historical data do not start from the first cycle.

To address the four main challenges in poisoning detection: training efficiency, accuracy, fairness, and requirements on the training dataset, we propose a poison-resilient and privacy-preserving federated learning scheme in mobile system based on historical gradient. Our core idea is unlike prior approaches that determine poisoning attacks by measuring the similarity between a participant's gradient and the aggregated global model gradient.

Our core idea is described as follows: during long iterative training, a participant's local model gradually converges until training completes, and the upload gradients exhibit relatively small fluctuations. When a participant launches a poisoning attack, the uploaded gradients display abnormal fluctuations, leading to irregular convergence patterns. We identify such behavior by analyzing changes in the difference between gradients uploaded in two consecutive rounds. In the early stages of training, these inter-round gradient differences are large; as the model approaches convergence, the differences become smaller. Since post-convergence gradient differences may be either positive or negative, where the sign merely reflects the convergence direction and carries no substantive meaning, we take the absolute value of the differences. As the model nears convergence, these absolute differences tend toward zero. Therefore, we compute the similarity between consecutive gradient differences: a high similarity (values near zero) indicates that the model is converging, whereas a poisoning attack will cause large fluctuations in the gradients, producing significant deviations in the inter-round differences and resulting in low similarity. Hence, CS determines whether participants conduct poisoning attacks based on the historical similarity, iteratively updates the global model by aggregating gradients of honest participants until the completing training model. Our scheme includes three phases: setup, local training, and gradient aggregating. Algorithm 1 shows poison-resilient and privacy-preserving federated learning scheme in mobile system.

Setup

CS sets up and shares a initialization framework M_0 and public parameters $\{\overline{\boldsymbol{\theta}}_0, T, \tau, \eta\}$ among P_i $(i = 1, 2, \ldots, n)$.

Local Training

Participants deploy and train M_0 based on local dataset D_i, iterative update the local model τ times, upload the gradient $\boldsymbol{\theta}_t^i$ to CS.

$$\boldsymbol{\theta}_t^i \leftarrow \boldsymbol{\theta}_{t-1}^i - \eta \boldsymbol{\theta}_{t-1}^i \tag{3}$$

Starting from the third training round, the gradients' similarity Sim_t^i, $(t \geq 3)$ between the differences of adjacent gradient is computing as follow,

$$\mathrm{Sim}_t^i = \mathrm{Simfun}(|\boldsymbol{\theta}_t^i - \boldsymbol{\theta}_{t-1}^i|, |\boldsymbol{\theta}_{t-1}^i - \boldsymbol{\theta}_{t-2}^i)| \tag{4}$$

Algorithm 1 : Poison-resilient and privacy-preserving federated learning scheme in mobile system

Input: M_0, P_i, D_i, $i \in [1, n]$, η, T, τ, t, aggregated gradients $\overline{\boldsymbol{\theta}}_t$, m
Output: Global model M
CS shares $\overline{\boldsymbol{\theta}}_0$ and M_0 with participants P_i
for $i = 1$ to n **do**
  for $t = 1$ to T **do**
    $\boldsymbol{\theta}_t^i \leftarrow \boldsymbol{\theta}_{t-1}^i - \eta \boldsymbol{\theta}_{t-1}^i$
    if τ divides t **then**
      P_i uploads $\boldsymbol{\theta}_t^i$
      P_i downloads $\overline{\boldsymbol{\theta}}_t$ and updates local model
      if $T \geq t \geq 3$ **then**
        P_i computes and uploads Sim_t^i by computing equation (4)
      end if
    end if
  end for
end for
for $t = 1$ to T **do**
  if τ divides t **then**
    for $i = 1$ to n **do**
      S checks sim_t^i uploaded by P_i
      Computes $\overline{\boldsymbol{\theta}}_t \leftarrow \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\theta}_t^i$
    end for
  end if
  CS updates model M with $\overline{\boldsymbol{\theta}}_t$
end for

and participants upload Sim_t^i and $\boldsymbol{\theta}_t^i$ to CS together. Participants use the downloaded gradients $\overline{\boldsymbol{\theta}}_t$ to update and train local model until T or completed training model.

Gradient Aggregating

CS determines whether participants conduct poisoning attack based on Sim_t^i. The similarity is computing

$$\overline{\mathrm{Sim}_t} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Sim}_t^i \tag{5}$$

$$|\mathrm{Sim}_t^i - \overline{\mathrm{Sim}_t}| > \xi \tag{6}$$

$\overline{\mathrm{Sim}_t}$ is average similarity at training round t after excluding malicious participants' extreme outlier values, and denotes an intermediate estimate for poisoning detection. Real number ξ is similarity threshold. By adjusting its value, CS eliminates participant gradients that deviate significantly from the mean similarity, thereby filtering out gradients from potential poisoning attackers in local training. CS aggregates local gradients contributed by benign participants, excluding those potentially affected by poisoning attacks.

$$\overline{\boldsymbol{\theta}}_t \leftarrow \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\theta}_t^i \tag{7}$$

CS updates global model with $\overline{\boldsymbol{\theta}}_t$ and shares the updated global model among P_i $(i = 1, 2, \ldots, n)$, iterating until the completed training.

IV. Analysis

In federated learning, a common assumption is that participants' local training datasets in mobile system follow an independent and identically distributed (IID) distribution. Nevertheless, in real-world scenarios, enforcing such an assumption on participants' data is always difficult. Subsection 2.2 has outlined the definitions of both IID and non-IID settings. In our scheme, we remove this restriction and allow participants' datasets to follow arbitrary distributions. Within this generalized setting, we conduct a comprehensive analysis of the proposed scheme in terms of security, correctness, fairness, and robustness.

A. Security Analysis

Under the scheme's security assumptions, this scheme only assumes malicious participants as adversaries, excluding both CS as an adversary and collusion attacks between CS and participants.

The training paradigm of the federated learning model determines that participants only communicates and exchanges data with CS, rather than directly with each other. As a result, malicious participants only obtain other participants' private data from the CS-downloaded gradients $\overline{\boldsymbol{\theta}}_t$. $\overline{\boldsymbol{\theta}}_t$ aggregated from the honest participants' gradients $\boldsymbol{\theta}_t^i$, $(i \in [1, m])$ is always used to update the global model M by CS in each cycle. When an adversary P_i obtains the gradient $\boldsymbol{\theta}_t^j$ or local data D_j of other participants P_j $(j \in [1, n], j \neq i)$ from the aggregated gradients $\overline{\boldsymbol{\theta}}_t$, the attack is successful. The impact of gradients $\boldsymbol{\theta}_t^i$ uploaded by participants P_i on the global model M depends on the distribution of participants' local dateset D_i $(i \in [1, n])$. If the distributions of local data D_i $(i \in [1, n])$ are similar, the influence of individual gradients $\boldsymbol{\theta}_t^i$ on the model M will decrease, and the probability of extracting other participants' gradients $\boldsymbol{\theta}_t^j$ from the aggregated gradients $\overline{\boldsymbol{\theta}}_t$ will drop significantly. Then, the probability of a successful attack is negligible. Thus, no federated learning can fully guarantee participants' data privacy and security in this scenarios, this attack method is outside our consideration, and we only consider attacks where participants poison their local datasets.

Experiments show that gradients uploaded by honest participants fluctuate greatly in the early training stage and stabilize in the later stage. After poisoning attacks by malicious participants, their uploaded gradients fluctuate significantly throughout the training process and are uncontrollable. If the malicious participants randomly control the gradients within a very small range, their poisoning behavior is equivalent to train local model normally, making targeted or untargeted attacks impossible. Thus, detecting the similarity of uploaded gradient differences to determine the convergence of participants' local models and monitoring changes in similarity can resist poisoning attacks by malicious participants in federated learning.

B. Correctly Analysis

This subsection focuses on the correctness analysis of our scheme. Specifically, our scheme considers correctness from two perspectives: (i) accurately identifying malicious participants who launch poisoning attacks and (ii) ensuring the correctness of the aggregated global model in the presence of poisoning attacks.

Two practical core assumptions are introduced to prove the correctly of our scheme. First, the similarity of the inter-round gradient differences uploaded by honest participants is concentrated within a certain range. Second, the similarity of gradients uploaded by malicious participants is dispersed. In other words, for honest participants, the core assumption is formally described as the concentration of similarity, which satisfies the condition given in Equation 8.

$$\forall i \in \mathcal{H}, Pr(|\mathrm{Sim}_t^i - \mathrm{Sim}_t| > \gamma_t) \leq \epsilon_t. \tag{8}$$

Where $\mathcal{H}$ denotes the set of honest participants, $Pr(\cdot)$ represents the probability symbol. Sim_t denotes the average gradient similarity uploaded by honest participants in the $t-$the training round. $\gamma_t = \frac{\xi_t}{|1-2\sigma|}, (\sigma \neq \frac{1}{2})$ and $\gamma_t = \xi_t, (\sigma = \frac{1}{2})$. Real number ξ_t is similarity threshold in the $t-$the training round. When the proportion of malicious participants is unknown, CS can infer whether this proportion exceeds one-half by examining the aggregated level of gradient similarities uploaded by participants, and thereby determine the size of γ_t. σ represents the proportion of dishonest participants. Since we remove the constraint in other schemes that the proportion of dishonest participants must be less than one half, the absolute value of the denominator is taken in the computation of γ_t. ϵ_t is an arbitrary positive number. For the similarity of the malicious participants, it would be described as follow

$$\forall i \in \mathcal{B}, |\mathrm{Sim}_t^i - \mathrm{Sim}_t| \geq \triangle_t. \tag{9}$$

Where $\mathcal{B}$ denotes the set of malicious participants and $\triangle_t$ is a real number much larger than γ_t.

Lemma 1: Regardless of whether the participants' datasets follow an IID or Non-IID distribution, if a malicious participant launches a poisoning attack during local training, the probability of being detected by the historical-gradient-based detection mechanism is non-negligible, our scheme satisfies the correctness of poisoning attack detection.

Proof: According to Equation (4), regardless of whether the local datasets of participants are IID or non-IID, each participant computes the similarity between the gradient differences of the local model over two consecutive rounds. As long as the local model training satisfies the convergence conditions of federated learning, the similarity of gradient differences will gradually increase and eventually fluctuate within a small range. Since the proposed scheme employs the SGD algorithm, which ensures local model convergence, differences in data distribution do not affect the computation of similarity. In other words, the evaluation of participants' gradient similarity is unaffected by the IID or Non-IID distribution assumption.

Assumption there is an undetected malicious participant whose poisoning attack occurs during local training. Then $\exists i \in \mathcal{B}$

$$Pr(\exists i \in \mathcal{B}) = Pr(|\mathrm{Sim}_t^i - \mathrm{Sim}_t| \leq \gamma_t) \tag{10}$$

If $\overline{\text{Sim}_t}$ is sufficiently close to Sim_t, the properties of inequalities imply that

$$|\overline{\text{Sim}_t} - \text{Sim}_t| \leq \epsilon_t. \tag{11}$$

Then, we can imply $\exists i \in \mathcal{B}$ that

$$\begin{aligned} |\text{Sim}_t^i - \overline{\text{Sim}_t}| &\geq |\text{Sim}_t^i - \text{Sim}_t| - |\text{Sim}_t - \overline{\text{Sim}_t}| \\ &= |\text{Sim}_t^i - \text{Sim}_t| - |\overline{\text{Sim}_t} - \text{Sim}_t| \\ &\geq \triangle_t - \epsilon_t \end{aligned} \tag{12}$$

Since $\triangle_t$ is a real number much larger than γ_t, then $\triangle_t - \epsilon_t > \gamma_t$, then $|\text{Sim}_t^i - \overline{\text{Sim}_t}| > \gamma_t$,

$$Pr(\exists i \in \mathcal{B}) = Pr(|\text{Sim}_t^i - \overline{\text{Sim}_t}| \leq \gamma_t) \to 0. \tag{13}$$

Therefore, if a malicious poisoning attacker exists, the probability that it remains undetected during local training is negligible. Then our scheme accurately identifies malicious participants who launch poisoning attacks, and our scheme satisfies the correctness of poisoning attack detection. This completes the proof of Lemma 1 $\perp$.

Lemma 2: Regardless of whether the participants' datasets follow an IID or Non-IID distribution, our scheme satisfies the correctness of the aggregation result.

Proof: In our scheme, each participant computes the difference between the gradients uploaded in two consecutive local training rounds and hen computes the similarity of these differences. As a result, the reported gradient similarity of a participant is independent of other participants, which further implies that the gradient similarity is not affected by the IID or Non-IID assumption. Therefore, utilizing the similarity among participants' gradients to determine whether a poisoning attack exists is essentially independent of the specific characteristics of their local data distributions. In other words, regardless of whether the participants' data are IID or Non-IID, as long as the local training process satisfies the convergence requirements of federated learning, the gradient update directions of honest participants will gradually converge during the iterations. Consequently, the similarity metric would effectively distinguish honest participants from malicious ones.

In each round of training model, CS aggregates the gradients $\boldsymbol{\theta}_t^i$ (correspond to the true gradients trained on the local data distribution) from all honest participants $\mathcal{H}$. The aggregation resulting $\overline{\boldsymbol{\theta}}_t$ is consistent with the ideal aggregation result $\mathbb{E}(\boldsymbol{\theta}_t^i)$, then

$$\overline{\boldsymbol{\theta}}_t = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\theta}_t^i \approx \mathbb{E}_{i \in \mathcal{H}}(\boldsymbol{\theta}_t^i) \tag{14}$$

By combining the convergence theory of federated learning (assuming that the local objective functions of all honest participants are $\mathcal{L}-$smooth and satisfy convexity conditions), it can be shown that after multiple iterations:

$$\boldsymbol{\theta}_t^i \leftarrow \boldsymbol{\theta}_{t-1}^i - \eta \boldsymbol{\theta}_{t-1}^i \Rightarrow \mathcal{L}(\boldsymbol{\theta}_t^i) \leq \mathcal{L}(\boldsymbol{\theta}_{t-1}^i) - \eta \mathcal{L}(\boldsymbol{\theta}_{t-1}^i) \tag{15}$$

In other words, the global model still converges to the optimal solution. When malicious participants' gradients are excluded, the aggregation of the remaining honest participants' gradients still guarantees correctness. Our scheme satisfies the correctness of the aggregation result. This completes the proof of Lemma 2 $\perp$.

C. Fairness Analysis

This subsection analyzes the fairness of FL in our scheme from two perspectives: individual fairness and group fairness. Specifically, individual fairness requires that each participant's contribution be fairly incorporated during model training, thereby ensuring that all participants obtain comparable benefits. Group fairness, on the other hand, requires that all honest participants eventually receive consistent global models, leading to uniformity in model predictions across the group. We introduce the following formal definitions of individual fairness and group fairness.

$$\Theta_{Ind} = d(h_M(x_i) - h_M(x_j)) \leq \epsilon,\ if \quad d(x_i, x_j) \leq \mu \tag{16}$$

$$\Theta_{Gro} = |[h_M(x)|A = 1] - [h_M(x)|A = 0]| \to 0 \tag{17}$$

Where Θ_{Ind} and Θ_{Gro} denote the symbol of individual and group fairness, and $d(\bullet)$ is a distance-measuring function. $h_M(x_i)$ represents the prediction result of model M with inputting x_i. ϵ and μ denote a threshold value. $A = 1$ and $A = 0$ represent different participants in the group, who may come from diverse genders, races, geographic regions, etc.

Lemma 3: Our scheme satisfies fairness iff the final global model satisfies both individual fairness and group fairness.

Proof: CS aggregates gradients $\boldsymbol{\theta}_t^i$ from all honest participants P_i,$(\forall i \in \mathcal{H})$. During the poisoning detection process, outlier gradients $\boldsymbol{\theta}_t^i$ from potentially malicious participants P_i,$(\forall i \in \mathcal{B})$ are excluded, ensuring that all honest participants' gradients are aggregated for updating global model. Lemma 1 has been proven that if a malicious participant launches a poisoning attack, it cannot bypass the poisoning detection, and its uploaded gradients will be excluded. In contrast, the gradient similarity values uploaded by honest participants will lie within the range $[\text{Sim}_t - \gamma_t, \text{Sim}_t + \gamma_t]$, enabling them to successfully pass the poisoning detection; thus, their gradients will be aggregated for updating the global model. Therefore, our scheme fairly excludes the gradients uploaded by malicious participants and aggregates the gradients of all honest participants. Hence, the poisoning attack detection is fair to each honest participant.

Subsequently, during the gradient aggregation process, our scheme employs mean aggregation of honest participants' gradients $\overline{\boldsymbol{\theta}}_t = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\theta}_t^i$ to update the global model. Thereby guaranteeing that each honest participant makes an equal contribution to the global model. Upon completion of the global model training, this model is distributed to all participants. If CS repeatedly detects poisoning attacks in a participant's local training, this participant will be excluded from subsequent federated learning and will no longer be entitled to share its benefits. Then, when x_i and x_j, $(x_i \neq x_j)$ and $d(x_i, x_j) \leq \mu$, are input into model M, the model M will produce identical or similar predictions $|(h_M(x_i) - h_M(x_j)| \leq \epsilon$, even though x_i and x_j may originate from the dataset of the same participant or from different participants $x_i \in D_i$, $x_j \in D_i$ or $x_i \in D_i$, $x_j \in D_j$, $(i \neq j)$. Hence, each participant receives identical benefits, updating the global model using the mean value of gradients from honest participants ensures individual fairness.

The gradients $\boldsymbol{\theta}_t^i$,$(i \in [1, n])$ uploaded by malicious poisoning attackers P_i, $(\forall i \in \mathcal{B})$ will be excluded. If poisoning

is detected repeatedly, such participants P_i will be removed from the federated learning process. Therefore, it is reasonable to assume that after each gradient $\boldsymbol{\theta}_t^i$, $(\forall i \in \mathcal{H})$ aggregation and global model updating M, the updated global model is distributed to all honest participants P_i, $(\forall i \in \mathcal{H})$. CS distributes the same initialization model M_0, and in each round the global model M_t is updated using same gradients $\boldsymbol{\theta}_t$. Hence, all honest participants will ultimately obtain same global model. Although different participants P_i, $(i \in [1, n])$ possess highly diverse attributes, as long as the input x_i is same, model M will inevitably produce identical predictions.

$$\Theta_{Gro} = |[h_M(x_i)|A = 1] - [h_M(x_j)|A = 0]| \to 0 \quad (18)$$

Hence, this approach of updating the global model based on the mean aggregation of honest participants' gradients ensures group fairness, then our scheme satisfies fairness. This completes the proof of Lemma 3 $\perp$.

D. Robustness Analysis

In order to address the vulnerability of federated learning model to attack from malicious participants, Byzantine-robust federated learning[27] has been extensively studied, which can be broadly categorized into four types: perturbation-based robustness, robustness based on the proportion of malicious participants, optimization-based robustness, and aggregation-based robustness. In perturbation-based robustness, even if some participants upload tampered gradients, the degradation in the performance of the aggregated global model must remain controlled. Under a certain proportion of malicious attacks, the training process should still converge or maintain stable performance. Optimization-based robustness focuses on the global model still achieving an optimal solution in the worst-case scenario. Aggregation-based robustness requires that, despite deviations in the aggregated values, the updated global model can still converge and approximate the performance of a benign model. Our scheme introduces the definition of model robustness from two aspects: detection robustness and aggregation robustness.

Lemma 4: If our scheme satisfies both detection robustness and aggregation robustness, then our scheme satisfies robustness.

Proof: We prove the robustness of detection firstly, followed by the robustness of aggregation process. The detection robustness is described as follows. When the proportion of malicious participants is σ_t, CS can with probability as least $1 - \delta_t$, correctly identify all malicious participants without misclassifying any honest participants, namely,

$$Pr[\widehat{\mathcal{H}} \subseteq \mathcal{H} \wedge \mathcal{B} \subseteq \widehat{\mathcal{B}}] \geq 1 - \delta_t$$

Where $\widehat{\mathcal{H}}$ denotes participants retained post-detection, while $\widehat{\mathcal{B}}$ represents participants that have been excluded. We use notion of neighborhoods (r_t, δ_t) to prove the robustness of poisoning attack detection, the meaning of r_t is explained in equation 8 and δ_t is a very small real number between 0 and 1. For honest participants, $|\mathrm{Sim}_t^i - \mathrm{Sim}_t| \leq \gamma_t$, then

$$|\mathrm{Sim}_t^i - \overline{\mathrm{Sim}_t}| \leq |\mathrm{Sim}_t^i - \mathrm{Sim}_t| + |\mathrm{Sim}_t - \overline{\mathrm{Sim}_t}| \quad \leq \gamma_t + \epsilon_t < \zeta_t \quad (19)$$

Where ζ_t is the similarity detection threshold, satisfying $\zeta_t \in (\gamma_t + \epsilon_t, \triangle_t - \epsilon_t)$. Hence, the probability of an honest participant $i \in \mathcal{H}$ being misclassified in the $t-$th training round as malicious $i \in \mathcal{B}$ is given by

$$Pr_{i \in \mathcal{H}}(i \in \mathcal{B}) \leq Pr(|\overline{\mathrm{Sim}_t} - \mathrm{Sim}_t| > \epsilon_t) \leq \delta_t \quad (20)$$

For malicious participants, then,

$$|\mathrm{Sim}_t^i - \mathrm{Sim}_t| \geq \triangle_t \Rightarrow |\mathrm{Sim}_t^i - \overline{\mathrm{Sim}_t}| \geq \triangle_t - \epsilon_t > \zeta_t \quad (21)$$

Hence, the probability of an honest participant $i \in \mathcal{B}$ being misclassified in the $t-$th training round as malicious $i \in \mathcal{H}$ is given by

$$Pr_{i \in \mathcal{B}}(i \in \mathcal{H}) \leq \delta_t \quad (22)$$

Therefore, the probability of misclassifying both honest and malicious participants are negligible, and our scheme satisfies the robustness of poisoning attack detection.

Next, we prove the robustness of federated learning aggregation. Similar to other schemes, our scheme may have a misjudgment probability in poisoning attack detection. The misclassification may include in two cases: (i) an honest participant is misidentified as malicious, and (ii) a malicious participant is misidentified as honest. For the first case, as long as an honest participant is not misclassified consecutively in multiple training rounds, the impact remains negligible since participants continue to contribute in subsequent rounds. Therefore, our discussion primarily focuses on the second case, namely when a malicious participant is misclassified as honest. We assume that the misclassification rate is denoted by ρ_t, then $\rho_t = \frac{\widehat{\mathcal{H}} - \mathcal{H}}{n}$. Therefore the mean similarity among honest participants changes by no more than υ_t. Where h_t and $\widehat{h}_t$ are the cardinality of $\mathcal{H}_t$ and $\widehat{\mathcal{H}}_t$, respectively. If the global model is updated using the mean of the gradients uploaded by the honest participants, then under the same aggregator, the followed equation holds.

$$\upsilon_t = |\mathbb{E}_{i \in \mathcal{H}_t}(Sim_t^i) - \mathbb{E}_{i \in \widehat{\mathcal{H}}_t}(Sim_t^i)| \leq \frac{\rho_t}{1 - \rho_t}\xi_t + \gamma_t \quad (23)$$

Where $\mathbb{E}_{i \in \mathcal{H}_t}(Sim_t^i) = \frac{1}{h_t}\sum_{i=1}^{m} Sim_t^i$, $\mathbb{E}_{i \in \widehat{\mathcal{H}}_t}(Sim_t^i) = \frac{1}{\widehat{h}_t}\sum_{i=1}^{m} Sim_t^i$. This demonstrates that although some malicious participants may be misclassified as honest, the deviation of their uploaded gradients has a upper bound. By substituting these gradients into the participants' local SGD algorithm, the model still converges linearly to a neighborhood. As a result, the impact on the global model updated by the CS through gradient aggregation is negligible, thereby satisfying the robustness of model aggregation. This completes the proof of Lemma 4 $\perp$.

E. Experimental Analysis

Assumption that CS and 10 participants jointly train a FL model, where each participant holds an independent and identically distributed dataset in our experiments. The dataset used in our experiments is MNIST[31], and the experimental equipment parameters are a personal computer with an Intel(R) Core(TM) i7-12700F @ 2.10 GHz CPU, 16.0 GB RAM, running Ubuntu 16.04.1 LTS 64-bit, implemented in Python, and the network environment was considered as a local area

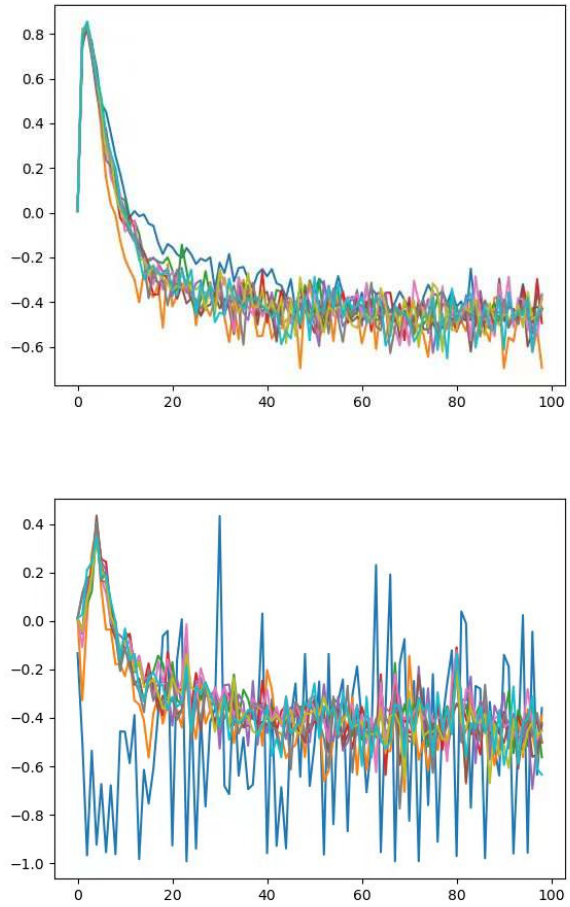


Fig. 1: Poison-resilient and privacy-preserving FL based on cosine similarity; (a) All honest participants; (b) Participants contain poisoning attackers

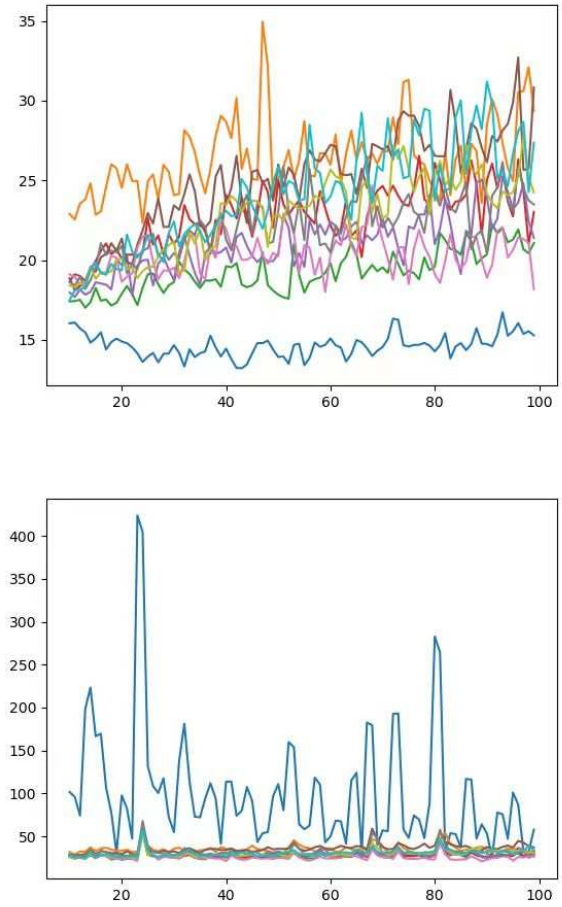


Fig. 2: Poison-resilient and privacy-preserving FL based on Manhattan distance similarity; (a) Honest Participants; (b) Participants contain poisoning attackers

networks. The maximum number of local training iterations for weak participants is set to $T = 100$, the learning rate $\eta = 0.001$, $\tau = 64$, $t = 93$.

We set up two control groups in experiments: one with all honest participants and the other containing one malicious data-poisoning attacker. Poisoning attackers are detected based on the similarity of gradient differences between two rounds, which is computed using cosine function and Manhattan distance. To better demonstrate the attack detection performance, the detected attacker is not excluded during the experiment.

The local model training gradients of the participants are recorded to calculate the similarity between the gradient differences. We plot the Figure 1 based on cosine similarity, and plot Figure 2 based on Manhattan distance similarity. Subfigure a and b of Figure 1 respectively the changes in cosine similarity when all participants are honest and when there is a malicious poisoning attacker. Subfigure a and b of Figure 2 respectively the changes in Manhattan distance similarity under the corresponding scenarios. In order to better demonstrate the detection effect of poisoning attacks, we started from the 10th round to calculate the Manhattan distance based on the gradient difference, and obtained Figure 2. The calculation starting from which round does not affect the actual experimental results.

Subfigure a of both figures shows that regardless of the similarity calculation method used, the change trend of the similarity of gradient differences uploaded by honest participants is consistent: in the initial stage of local training, the gradient difference is large but the change direction is consistent; in the later stage, the gradient difference decreases and the direction is different With participants' local iterative training, resulting in the similarity fluctuating within a small range and the direction is not completely unchanged. Subfigure b of both figures shows that there is a difference in the similarity of gradient differences between poisoning attackers and honest participants. In the initial stage of local training, the changes in similarity uploaded by the two are different: the similarity uploaded by honest participants initially rises and then tends to be stable With participants' local iterative training, while the similarity of gradient differences uploaded by attackers fluctuates significantly from the initial stage to the final convergence stage, indicating large fluctuations

and low similarity of the uploaded gradients after malicious participants launch poisoning attacks.

Regardless of the similarity calculation method adopted, we draw the same conclusion — our scheme can accurately detect poisoning attackers. Therefore, we believe that whether a participant launches a poisoning attack can be detected by calculating the similarity of the gradients uploaded by the participant.

V. Conclusion

In this paper, we propose a poison-resilient and privacy-preserving federated learning scheme in mobile system based on the similarities of gradient differences. The scheme is applicable to both vertical and horizontal federated learning. Detection the poisoning attack based on each participant's historical gradient, and does not depend on the local data distribution of participants, thus offering broader applicability. We prove security, correctness, fairness, and robustness through several lemmas. The experiments uses cosine and Manhattan distances to compute participants' gradient similarities, and the results clearly identify the presence of poisoning attackers, demonstrating the effectiveness of the poisoning detection. Future work will focus on evaluating the effectiveness and security of the scheme in scenarios where the central server behaves maliciously.

References

[1] Huang Y, Gupta S, Song Z, et al. Evaluating gradient inversion attacks and defenses in federated learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 7232-7241.

[2] Li J, Rakin A S, Chen X, et al. Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 10194-10202.

[3] Hao M, Li H, Luo X, et al. Efficient and privacy-enhanced federated learning for industrial artificial intelligence[J]. IEEE Transactions on Industrial Informatics, 2019, 16(10): 6532-6542.

[4] Wei K, Li J, Ding M, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.

[5] Stevens T, Skalka C, Vincent C, et al. Efficient differentially private secure aggregation for federated learning via hardness of learning with errors[C]. The 31st USENIX Security Symposium (USENIX Security 22). 2022: 1379-1395.

[6] Zhu H, Li Z, Cheah M, et al. Privacy-preserving weighted federated learning within oracle-aided MPC framework[J]. arXiv preprint arXiv:2003.07630, 2020.

[7] Guo S, Wang X, Long S, et al. A federated learning scheme meets dynamic differential privacy[J]. CAAI Transactions on Intelligence Technology, 2023.

[8] Tolpegin V, Truex S, Gursoy M E, et al. Data poisoning attacks against federated learning systems[C]. The 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14 - 18, 2020, Proceedings, Part I 25. Springer International Publishing, 2020: 480-501.

[9] Yang L, Zhang J, Chai D, et al. Practical and Secure Federated Recommendation with Personalized Mask[C]. International Workshop on Trustworthy Federated Learning. Cham: Springer International Publishing, 2022: 33-45.

[10] Shejwalkar V, Houmansadr A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning[C]. NDSS. 2021.

[11] Zhou Y, Zhao T, Jing Z, Zhao Q, Zhu Y. A blockchain-based privacy-preserving data aggregation scheme with robustness in smart grids[J]. The Journal of Supercomputing, 2025, 81(5):675.

[12] Hu R, Guo Y, Li H, et al. Personalized federated learning with differential privacy[J]. IEEE Internet of Things Journal, 2020, 7(10): 9530-9539.

[13] Singhal K, Sidahmed H, Garrett Z, et al. Federated reconstruction: Partially local federated learning[J]. Advances in Neural Information Processing Systems, 2021, 34: 11220-11232.

[14] Gu T, Dolan-Gavitt B, BadNets S. Identifying vulnerabilities in the machine learning model supply chain[C]. Proceedings of the Neural Information Processing Symposium Workshop Mach. Learning Security (MLSec). 2017: 1-5.

[15] Shejwalkar V, Houmansadr A, Kairouz P, et al. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning[C]. 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022: 1354-1371.

[16] Chen J, Zhang X, Zhang R, et al. De-pois: An attack-agnostic defense against data poisoning attacks[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 3412-3425.

[17] Shejwalkar V, Houmansadr A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning[C].NDSS. 2021.

[18] Fang M, Cao X, Jia J, et al. Local model poisoning attacks to Byzantine-Robust federated learning[C]. 29th USENIX security symposium (USENIX Security 20). 2020: 1605-1622.

[19] Li Y, Wei X, Li Y, et al. Detection of false data injection attacks in smart grid: A secure federated deep learning approach[J]. IEEE Transactions on Smart Grid, 2022, 13(6): 4862-4872.

[20] Jiang Y, Zhang W, Chen Y. Data quality detection mechanism against label flipping attacks in federated learning[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1625-1637.

[21] Geiping J, Bauermeister H, Dr?ge H, et al. Inverting gradients-how easy is it to break privacy in federated learning?[J]. Advances in Neural Information Processing Systems, 2020, 33: 16937-16947.

[22] Duan M, Liu D, Chen X, et al. Self-balancing federated learning with global imbalanced data in mobile systems[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(1): 59-71.

[23] Gupta P, Yadav K, Gupta B B, et al. A Novel Data Poisoning Attack in Federated Learning based on Inverted Loss Function[J]. Computers and Security, 2023, 130: 103270.

[24] Li Q, Wen Z, He B. Practical federated gradient boosting decision trees[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(04): 4642-4649.

[25] Mao Y, Yuan X, Zhao X, et al. Romoa: Ro bust mo del a ggregation for the resistance of federated learning to model poisoning attacks[C]. The 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4 - 8, 2021, Proceedings, Part I 26. Springer International Publishing, 2021: 476-496.

[26] Wang N, Xiao Y, Chen Y, et al. FLARE: defending federated learning against model poisoning attacks via latent space representations[C].Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. 2022: 946-958.

[27] Fang M, Zhang Z, Hairi, et al. Byzantine-robust decentralized federated learning[C]. Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2024: 2874-2888.

[28] Truex S, Baracaldo N, Anwar A, et al. A hybrid approach to privacy-preserving federated learning[C]. Proceedings of the 12th ACM workshop on artificial intelligence and security. 2019: 1-11.

[29] Seif M, Tandon R, Li M. Wireless federated learning with local differential privacy[C]. 2020 IEEE International Symposium on Information Theory (ISIT). IEEE, 2020: 2604-2609.

[30] Zhao Q, Zhang Y, Jing Z, et al. Federated Learning Poison Attack Detection Scheme Based on Gradient Similarity[C]. International Symposium on Emerging Information Security and Applications. Cham: Springer Nature Switzerland, 2024: 20-36.

[31] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.