Available Online: 19 May 2026 DOI: 10.54254/2977-3903/2026.33579

Experimental study of poisoning attacks in federated learning

Yanqin You

Scotland College, Wuxi Taihu University, Wuxi, China

yx3745yyq@126.com

Abstract. Federated Learning has become an important distributed machine learning paradigm, enabling several clients to work together to jointly train a global model without exchanging raw data without sharing raw data. Although it improves data privacy, it also creates security vulnerabilities, particularly poisoning attacks conducted by malicious clients. Understanding these risks has become crucial for guaranteeing system robustness and dependability as federated learning is used more frequently in privacy-sensitive situations. This study investigates the impact of label flipping attacks in federated learning systems through experimental analysis. A distributed learning environment is constructed using multiple clients, and malicious participants are introduced to manipulate training data. This study evaluates the performance of the global model under different attack ratios, and further uses gradient analysis to observe abnormal update patterns caused by poisoned clients. Results show that poisoning attacks greatly affect model stability and accuracy. As the number of malicious clients increases, training becomes less stable and the convergence process is more easily disrupted. The findings highlight the vulnerability of federated learning systems and emphasize the necessity for robust defense mechanisms.

Keywords: federated learning, poisoning attack, label flipping, machine learning security, distributed learning

1. Introduction

Federated Learning (FL), allowing several clients to jointly train a global model without exchanging raw data, has become a significant distributed machine learning paradigm. This decentralized approach helps protect user privacy and decreases the risks linked with centralized data storage. Thus, FL has been widely applied in areas such as healthcare, mobile applications, and Internet of Things (IoT) systems, where sensitive data cannot be easily shared among participants [1,2].

Despite its privacy benefits, federated learning also introduces new security challenges. Malicious or compromised participants may purposefully alter local data or model updates because the global model is updated based on model parameters submitted by multiple clients. Such behavior can lead to poisoning attacks, aiming to degrade global model performance or influence model predictions [3].

Prior studies have shown that poisoning attacks can significantly affect federated learning systems. In particular, label flipping attacks have been proven effective in reducing model accuracy [4,5]. Although

existing research has examined poisoning attacks from different perspectives, further experimental analysis is still needed to better understand their impact under different attack conditions.

This study conducts an experimental analysis of poisoning attacks in federated learning by reproducing label flipping attacks under different malicious client ratios. The global model's performance is assessed in various attack scenarios. The study emphasizes the significance of creating strong defense mechanisms and offers additional insight into the vulnerability of federated learning systems.

2. Literature review

2.1. Federated learning

A distributed machine learning framework called Federated Learning (FL) enables several clients to work together to train a common global model while maintaining the privacy of their local data. The Federated Averaging (FedAvg) algorithm was initially proposed by McMahan et al. to aggregate model updates from distributed clients [1]. With this method, each client uses its own data to train a local model, then transmits the updated parameters to a central server. These parameters are then combined by the server to update the global model.

Federated learning has drawn a lot of interest because it can lower communication overhead and protect data privacy. It has been extensively used in a number of fields, including IoT systems, healthcare data analysis, and mobile keyboard prediction [2]. However, since federated learning relies on the contributions of multiple clients, it introduces new security challenges. In particular, malicious participants may upload manipulated model updates that degrade the global model performance.

Enhancing the effectiveness and resilience of federated learning has also been the subject of numerous studies. By enabling clients to complete several local training steps prior to sending updates to the server, FedAvg, for instance, lowers communication costs [1]. Furthermore, to improve data security in federated learning systems, recent studies have investigated secure aggregation methods and differential privacy [2]. Despite these improvements, federated learning remains vulnerable to various attacks, particularly poisoning attacks.

Recent studies have also explored challenges in heterogeneous federated learning environments. Li et al. studied federated optimization under non-IID data distributions and demonstrated that data heterogeneity can significantly impact model performance [6]. These findings highlight additional vulnerabilities in federated learning systems, which may further amplify the impact of poisoning attacks.

2.2. Poisoning attacks in federated learning

One of the main security risks in federated learning is poisoning attacks. Malicious clients alter their local data or model updates in these kinds of attacks in order to affect the global model. These attacks fall into two general categories: model poisoning attacks and data poisoning attacks.

Bhagoji et al. investigated adversarial attacks in federated learning and demonstrated that malicious clients could significantly degrade model performance by manipulating local training data [3]. Their work showed that federated learning systems are vulnerable to adversarial behavior, especially when malicious participants control a portion of the training data.

Bagdasaryan et al. proposed model poisoning attacks and demonstrated that attackers can inject backdoors into federated learning models [4]. Their results showed that attackers could manipulate model behavior without significantly affecting overall model accuracy, making such attacks difficult to detect.

Fang et al. further studied local model poisoning attacks and demonstrated that even Byzantine-robust aggregation methods may still be vulnerable to malicious clients [5]. Their findings highlighted the limitations of existing defense mechanisms and emphasized the need for more robust federated learning frameworks.

Another common poisoning attack is the label flipping attack, where malicious clients intentionally change the labels of training data. This attack is simple to implement but can significantly degrade model performance, especially when multiple malicious clients participate in training. Several studies have shown that label flipping attacks remain effective under different federated learning settings.

Baruch et al. revealed that even a small number of malicious participants can successfully bypass defense mechanisms and significantly degrade model performance [7]. In addition, Label flipping attacks can successfully lower global model accuracy under a variety of experimental conditions, according to Tolpegin et al.'s investigation of data poisoning attacks in federated learning [8]. These studies further emphasize the importance of evaluating poisoning attacks in federated learning systems.

2.3. Research gap

Although previous studies have investigated poisoning attacks in federated learning, many of them focus on theoretical analysis or specific attack scenarios. Experimental research that assesses the efficacy of poisoning attacks under various circumstances is still required. In particular, the impact of malicious client ratios, attack timing, and different datasets requires further investigation. Additionally, reproducibility of poisoning attacks is important for understanding the security risks of federated learning systems. Therefore, this study conducts an experimental analysis by reproducing poisoning attacks and evaluating their impact on federated learning performance. In addition to analyzing the vulnerability of federated learning models under various attack scenarios, this work attempts to provide a thorough experimental evaluation of poisoning attacks in federated learning systems.

3. Methodology

3.1. Experimental setup

The implementation and attack setting outlined by Tolpegin et al. served as the basis for all of the experiments in this study [8]. The PyTorch deep learning framework was used to implement the experiments in Python. Without exchanging local data, the federated learning environment mimics several dispersed clients working together to train a global model.

The Fashion-MNIST dataset was used in this study. It comprises 60,000 training samples and 10,000 testing samples, with 10 classes of clothing images [9]. The dataset was partitioned across 50 clients to simulate a federated learning environment.

A convolutional neural network (CNN) model was used as the global model, which is widely applied in image classification tasks [10]. The same model architecture was deployed across all clients to ensure consistency during training.

The federated learning settings were configured with specific parameters to ensure effective model training and aggregation. The system included 50 clients in total, with 5 clients selected to participate in each training round. The training process consisted of 200 rounds in total, and each client used a batch size of 64 for local model training. For the aggregation of local model updates, the FedAvg method was adopted [1]

3.2. Label flipping attack

To evaluate poisoning attacks, this study adopts the label flipping attack method [8]. In this attack, malicious clients modify their local dataset by flipping labels during training. Specifically, label "1" is replaced with label "9" to introduce incorrect training information.

The impact of attack strength is examined by varying the number of malicious clients. Zero malicious clients, five malicious clients, ten malicious clients, and fifteen malicious clients are the settings that are utilized.

All experiments were conducted for 200 communication rounds. The performance of the global model was evaluated using classification accuracy.

3.3. Evaluation metrics

Classification accuracy was used to assess the model's performance. A common metric in federated learning research is accuracy, which quantifies the percentage of correctly classified samples [2]. The accuracy of the global model was recorded after each communication round to analyze the impact of poisoning attacks.

In this study, accuracy was selected as the primary evaluation metric because the task is a multi-class image classification problem. It provides a direct and intuitive measure of global model performance under different attack ratios. In addition, changes in accuracy over communication rounds can clearly reflect convergence stability and the influence of malicious updates on the training process.

4. Results and analysis

4.1. Label flipping attack results

Experiments were carried out under four different malicious client ratios to assess the effect of label flipping attacks in federated learning. A popular data poisoning technique is label flipping, in which malevolent clients purposefully alter training labels to impair model performance [8]. In this experiment, malicious clients change the label "1" to "9" in their local datasets. During training, the global model then incorporates these tainted updates.

The federated learning framework in this experiment follows the standard FedAvg aggregation algorithm, where model updates from multiple clients are averaged to obtain the global model [1]. Although this aggregation strategy is effective under normal conditions, it is vulnerable to malicious client updates [3].

The experiment considers four scenarios: 0 malicious clients (0%), 5 malicious clients (10%), 10 malicious clients (20%), 15 malicious clients (30%).

The total number of clients is fixed at 50, and five clients are randomly selected in each communication round. The model is trained for 200 communication rounds using the Fashion-MNIST dataset [9].

Table 1 summarizes the performance of the global model under different malicious client ratios.

Table 1. Global model accuracy under different malicious client ratios

Malicious Clients	Ratio	Initial Accuracy (%)	Final Accuracy (%)	Max Accuracy (%)	Average Accuracy (%)
0	0%	67.41	90.30	90.35	88.53
5	10%	73.18	90.02	90.33	88.36
10	20%	61.60	80.21	90.23	87.36
15	30%	63.24	90.10	90.26	86.61

From Table 1, several important observations can be made.

First, in the baseline scenario without malicious clients, the global model converges smoothly and achieves a final accuracy of 90.30%, with an average accuracy of 88.53%. This indicates that the federated learning system performs effectively under normal conditions.

When 10% of clients are malicious, the model still maintains relatively stable performance. The final accuracy only slightly decreases to 90.02%, and the average accuracy remains at 88.36%. This suggests that the FedAvg aggregation mechanism can tolerate a small number of malicious participants without significant degradation. Similar observations have also been reported in previous studies on federated learning robustness [2].

However, when the malicious client ratio increases to 20%, the performance begins to degrade more noticeably. The final accuracy drops to 80.21%, which is significantly lower than the baseline. This degradation occurs because malicious updates begin to dominate the aggregation process, causing the global model to learn incorrect patterns [5].

In the 30% malicious client scenario, the average accuracy further decreases to 86.61%, which is the lowest among all scenarios. The training process becomes highly unstable, with frequent fluctuations in model accuracy. This instability indicates that the global model struggles to converge consistently under significant attack pressure.

To further illustrate the training behavior, Figure 1 shows the global model accuracy curves under different malicious client ratios.

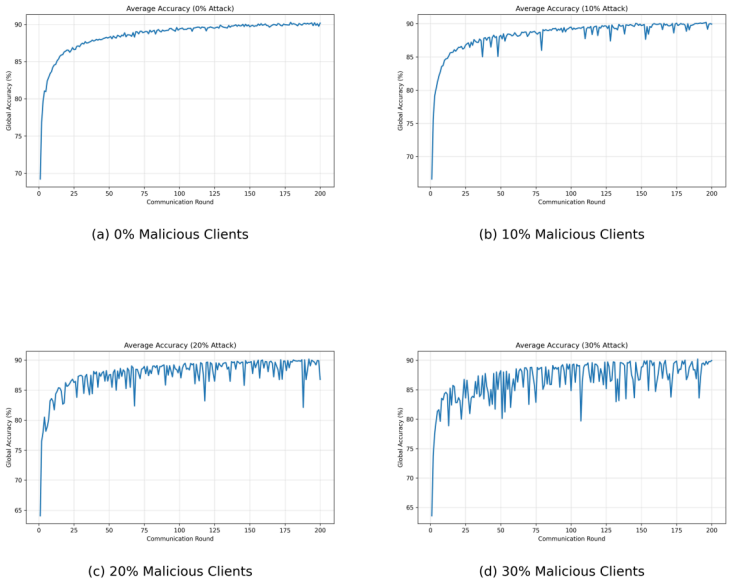


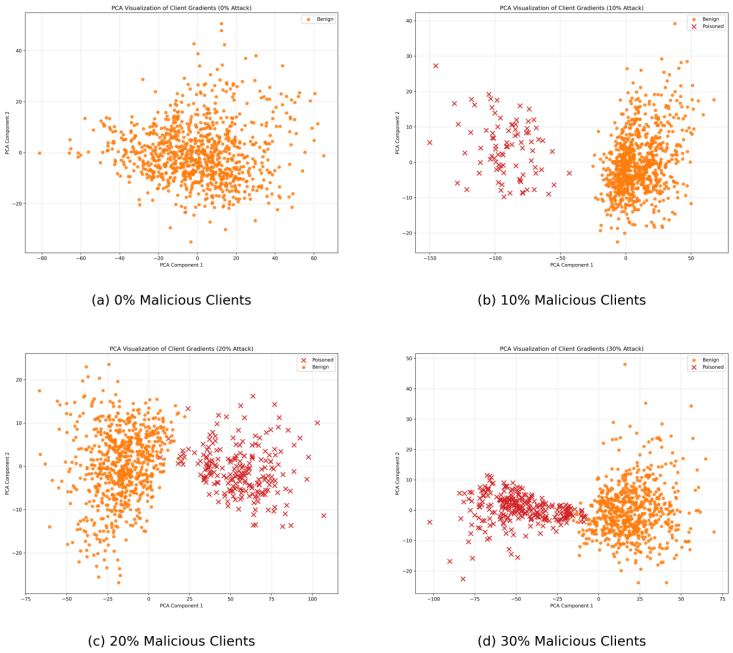
Figure 1. Average accuracy under different malicious client ratios

As shown in Figure 1, the model without attack converges smoothly throughout training. In contrast, the accuracy curves under malicious attacks exhibit increasing instability as the number of malicious clients grows. The fluctuations become more pronounced in the 20% and 30% attack scenarios, indicating that malicious updates introduce conflicting learning directions that hinder convergence.

These findings are consistent with previous studies showing that data poisoning attacks can significantly disrupt federated learning convergence [7,8].

4.2. Gradient analysis

To gain a deeper insight into how poisoning attacks influence the training process, gradient distributions from different clients were analyzed using Principal Component Analysis (PCA). PCA is commonly used to visualize high-dimensional gradient updates in federated learning environments [3]. Figure 2 presents the PCA visualization of client gradients under different malicious client ratios.



(a) 0% Malicious Clients (b) 10% Malicious Clients

(c) 20% Malicious Clients (d) 30% Malicious Clients

Figure 2. PCA visualization of client gradients under different attack ratios

It can be observed that, in the baseline scenario (0% malicious clients), gradients from different clients form a compact cluster. This indicates that clients share similar update directions and the global model converges consistently.

When 10% of clients are malicious, a small number of outliers begin to appear in the PCA space. These outliers correspond to malicious clients performing label flipping attacks. Although most clients still cluster together, malicious updates begin to introduce noise into the aggregation process.

In the 20% malicious client scenario, the separation between benign and malicious clients becomes more obvious. Distinct clusters emerge, indicating that malicious clients produce gradients that differ significantly from normal participants. This behavior has also been observed in adversarial federated learning studies [3,7].

In the 30% malicious client scenario, the gradient distribution becomes more dispersed. The malicious cluster grows larger, and overlap between benign and malicious gradients increases. This dispersion reflects the severe disruption caused by malicious updates.

These results demonstrate that poisoning attacks introduce abnormal gradient directions, which interfere with global model convergence. Gradient-based visualization therefore provides a useful tool for detecting malicious clients in federated learning systems [5].

5. Discussion

The outcomes of the experiment show that label flipping attacks can affect federated learning systems. Accuracy fluctuations become more noticeable and the global model becomes less stable as the number of malicious clients rises. Although the final accuracy may remain relatively high, the training process becomes increasingly unstable, especially under higher attack ratios.

These findings are consistent with prior studies on poisoning attacks in federated learning, which show that malicious clients can significantly influence global model performance [3,8]. Since the FedAvg algorithm aggregates client updates through simple averaging, malicious updates can slow down convergence and distort the global model [1].

The PCA analysis further supports this observation. As the proportion of malicious clients increases, gradient distributions become more dispersed, and abnormal clusters appear. These results indicate that poisoned clients produce gradients that differ significantly from benign clients, which disrupts the training process.

However, this study has several limitations. First, only label flipping attacks were considered. Other attack strategies such as backdoor attacks or model poisoning attacks may have different impacts. Second, the experiments were conducted using only the Fashion-MNIST dataset, which may not fully represent real-world scenarios. Finally, the number of clients and communication rounds were fixed, which may limit the generality.

By assessing additional attack tactics, testing various datasets, and investigating defense mechanisms like robust aggregation and anomaly detection, future research can expand on this study. These methods could strengthen federated learning systems' security and resilience.

6. Conclusion

This paper investigates the impact of label flipping attacks on federated learning systems through experimental analysis. A federated learning environment is constructed with multiple distributed clients, and malicious participants are introduced under different attack ratios to evaluate how poisoning behavior affects global model performance. The study mainly focuses on classification accuracy and gradient distribution in order to examine both performance degradation and training instability.

The experimental findings demonstrate the susceptibility of federated learning systems to poisoning attacks. When the number of malicious clients increases, the global model becomes less stable and the convergence process is more easily disrupted. Although the final accuracy does not always decrease dramatically in every setting, the average performance and training stability are clearly affected. In particular, higher malicious client ratios introduce stronger fluctuations during training, which indicates that poisoned updates interfere with normal model aggregation. In addition, the PCA-based gradient analysis shows that malicious clients tend to produce abnormal gradient patterns that are distinguishable from those of benign participants.

These findings suggest that traditional aggregation methods like FedAvg are not sufficiently robust against malicious behavior in adversarial federated learning environments. Therefore, improving the security and robustness of federated learning remains an important research issue. However, this study also has several limitations. Only label flipping attacks were considered, and the experiments were conducted on the Fashion-MNIST dataset under fixed settings. Future research can extend this work by investigating more attack strategies, testing additional datasets, and evaluating defense mechanisms.

References

[1] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54*, 1273–1282. PMLR. https: //proceedings.mlr.press/v54/mcmahan17a.html

[2] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). *Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14*(1–2), 1–210. https: //doi.org/10.1561/2200000083

[3] Bhagoji, A. N., Chakraborty, S., Mittal, P., & Calo, S. (2019). Analyzing federated learning through an adversarial lens. *Proceedings of the 36th International Conference on Machine Learning, 97,* 634–643. PMLR. https: //proceedings.mlr.press/v97/bhagoji19a.html

[4] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 108*, 2938–2948. PMLR. https: //proceedings.mlr.press/v108/bagdasaryan20a.html

[5] Fang, M., Cao, X., Jia, J., & Gong, N. (2020). Local model poisoning attacks to Byzantine-robust federated learning. In 29th USENIX Security Symposium (USENIX Security 20) (pp. 1605–1622). USENIX Association. https: //www.usenix.org/conference/usenixsecurity20/presentation/fang

[6] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *In Proceedings of Machine Learning and Systems 2020*. https: //proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html

[7] Baruch, M., Baruch, G., & Goldberg, Y. (2019). A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems, 32,* 8632–8642. https: //proceedings.neurips.cc/paper/2019/hash/ec1c59141046cd1866bbbcdfb6ae31d4-Abstract.html

[8] Tolpegin, V., Truex, S., Gursoy, M. E., & Liu, L. (2020). Data poisoning attacks against federated learning systems. In Computer Security – ESORICS 2020 (pp. 480–501). Springer. https: //doi.org/10.1007/978-3-030-58951-6_24

[9] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv. https: //doi.org/10.48550/arXiv.1708.07747

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. Nature, 521*(7553), 436–444. https: //doi.org/10.1038/nature14539