Revisiting Defense Mechanisms in Federated Learning: Effective and Efficient Backdoor Attack via Trigger Pre-optimization

Abstract

Backdoor attacks and defenses in federated learning (FL) have attracted significant attention due to their implications for model security. Through reproducibility testing of current attacks and defenses, we found that existing attack methods often fail to deliver consistently high success rates. To address this gap, we analyzed the effects of poisoning rates, joint data-label distributions, and client-label distributions on defenses. We theoretically and experimentally investigate the relationship between data distribution differences and model update discrepancies and provide an upper bound for attack effectiveness.

Building on these insights, we propose PREFed, a novel backdoor attack method that PRE-optimizes and REfines triggers to enhance efficiency and effectiveness. PREFed leverages mid-training global models to simulate both normal and malicious updates, iteratively refining triggers by maximizing their similarity to optimize their initial state. This approach ensures higher attack efficiency early in training, while continuous optimization further improves attack performance in later stages.

We evaluated PREFed against six advanced defense methods and compared it with five attack methods using three benchmark datasets. Experimental results demonstrate that PREFed achieves superior attack success rates while minimizing its impact on main task performance. Notably, PREFed achieves over 80% attack accuracy within just five training rounds.

1. Introduction

The rise of deep learning has underscored the critical role of data in developing robust models (Xu et al., 2019). Federated learning (FL) has emerged as a privacy-preserving paradigm that enables multiple participants to collaboratively train high-quality models without sharing raw data (Konečný et al., 2016; Aono et al., 2017). This distributed training approach has been widely adopted across domains (Miao et al., 2023; Islam et al., 2022). However, FL is vulnerable to security threats, particularly stealthy and highly damaging targeted backdoor attacks (Nguyen et al., 2019; 2020). In such attacks, malicious participants inject backdoors into the global model by combining local backdoor training with central aggregation. While these attacks leave the model's primary task performance unaffected, inputs containing specific triggers yield attacker-defined outputs.

Current defenses primarily rely on detecting and filtering anomalous models or updates during aggregation (Nguyen et al.; Rieger et al.). To bypass these defenses, attackers have developed adaptive strategies, such as increasing the influence of malicious updates during aggregation (Li et al., 2023; Zhang et al., 2024). However, as shown in our repeated experiments (see Table 7), existing adaptive attack methods struggle to maintain high success rates and require frequent adjustments based on feedback from subsequent global model updates. In addition, these dynamic adjustments significantly reduce attack efficiency.

Motivated by these limitations, we revisited existing defense mechanisms, particularly anomaly detection algorithms, to understand their vulnerabilities. Our analysis revealed that differences in dataset distribution are reflected in model updates, making detection more likely under certain conditions. Specifically, when the poisoning rate is high (e.g., 1), backdoor updates are more distinguishable, while lower poisoning rates (e.g., 0) render backdoor updates nearly indistinguishable from normal ones. Through theoretical 3 and experimental analysis 5, we established a relationship between data-label distribution differences and model update patterns. This insight led us to develop a novel backdoor attack approach that optimizes trigger design before deployment.

We propose PREFed, a backdoor attack method that incorporates Pre-optimizing and Refining trigger [1]. By simulating both backdoor and normal training processes before the attack phase, PREFed refines the trigger to maximize the similarity of both model updates, enhancing attack stealth and efficiency. Furthermore, PREFed continuously refines the trigger during the attack phase, further improving effectiveness and adaptability.

Our contributions are summarized as follows:

1) We establish a connection between dataset distribution and model updates through theoretical and experimental

[1] we also introduce PreFed with only trigger Pre-optimization.

analysis. We provide boundary conditions for backdoor attacks under detection mechanisms, offering strong evidence for PREFed's feasibility.

2) We present PREFed, the first backdoor attack method to incorporate preemptive trigger optimization. PREFed significantly improves attack efficiency through optimized triggers and further enhances effectiveness via continuous fine-tuning during backdoor implantation.

3) PREFed is evaluated on three benchmark datasets, six state-of-the-art defense mechanisms, and three commonly used attack strategies. Our method achieves superior attack performance, with backdoor accuracy exceeding 90% while causing minimal degradation (e.g., a maximum primary task reduction of 5.58% on CIFAR-10). Additionally, PREFed demonstrates high efficiency, achieving over 80% backdoor accuracy within five training rounds and reducing per-client per-round time costs by 82.9% compared to 3DFed.

2. Related Work

2.1. FL Backdoor Attack

Backdoor attacks in federated learning (FL) can be broadly categorized into fixed trigger attacks and trigger optimization attacks.

Fixed trigger attacks use predefined trigger patterns that remain constant visually or in data. Xie et al. (2019) and Gong et al. (2022) exploit FL's distributed nature to design collaborative backdoor attacks. To counter evolving defense mechanisms, Li et al. (2023) introduced 3DFed, an advanced attack method integrating adaptive modules to bypass multiple defenses. Similarly, Zhuang et al. (2023) improve backdoor implantation by targeting critical model layers, replacing benign updates with compromised ones, thereby evading detection.

Trigger optimization attacks are often more effective, as optimized triggers can more reliably activate backdoors (Pang et al., 2020). A notable example, A3FL (Zhang et al., 2024), predicts dynamic changes in the global model, allowing triggers to adapt and extend the lifespan of backdoors.

However, fixed and optimized trigger attacks depend on feedback from global model updates or estimates of other external information. This reliance on complex calculations reduces attack efficiency and, in some cases, limits their overall effectiveness.

2.2. FL Backdoor Defense

Defense mechanisms in FL against backdoor attacks generally fall into three categories: filtering strategies, mitigation strategies, and hybrid approaches.

Filtering strategies aim to identify and exclude inconsistent updates based on anomaly detection. Foolsgold (Fung et al., 2020) uses historical update information to identify malicious contributions. To address the high-dimensional nature of large models, RFLBAT (Wang et al., 2022) employs Principal Component Analysis (PCA) (Maćkiewicz & Ratajczak, 1993) for dimensionality reduction. FreqFed (Fereidooni et al.) further advances this approach by applying Discrete Cosine Transform (DCT) for spectral analysis, focusing on low-frequency components to improve clustering accuracy.

Mitigation strategies, inspired by differential privacy (McMahan et al., 2017b), aim to disrupt backdoor effectiveness by modifying uploaded model updates. This includes limiting update weights and adding noise (Bagdasaryan et al., 2020; Naseri et al., 2020). While effective in mitigating backdoor attacks and enhancing client privacy, these methods can introduce efficiency challenges and degrade overall model performance.

Hybrid approaches combine elements of filtering and mitigation for more robust defense. For example, Deepsight (Rieger et al.) and FLAME (Nguyen et al.) integrate norm clipping, noise addition, and HDBSCAN clustering (Campello et al., 2013) to counter backdoor attacks. These methods are compatible with common aggregation rules like FedAvg and FedSGD (McMahan et al., 2017a), ensuring adaptability to various FL frameworks.

3. Design of PREFed

In this section, we first conduct an in-depth analysis of the mechanism of the existing detection algorithm, with a particular focus on capturing the difference between model updates (Rieger et al.; Wang et al., 2022; Fereidooni et al.). We establish the relationship between model update discrepancies and dataset distribution differences and demonstrate that even models that have not fully converged can capture these differences.

Additionally, we perform a theoretical analysis to explore the attack boundary under the detection algorithm and demonstrate the relationship between attack feasibility and the dataset distribution difference (which has also been verified through experiments in Section 5). Based on these findings, we propose the PREFed, which enhances attack effectiveness by optimizing triggers in advance.

3.1. Discrepancy Analysis of Client Updates

At first, differences in dataset distribution primarily stem from two key aspects:

- **The Data-Label Joint Distribution**. In federated learning, variations in the joint distribution of labels and data among client datasets refer to differences in the joint probability distribution of data features and their corresponding labels across clients. Let $P(X, Y)$ represent the joint distribution of the data feature X and label Y. For two clients i and j, a difference in the joint distribution exists if their respective distributions, $P_i(X, Y)$ and $P_j(X, Y)$, are not identical.

- **Label Distribution Difference**. Differences in label distribution refer to variations in the probability distribution of labels across clients. Let $P(Y)$ denote the distribution of label Y. For two clients i and j, a difference in label distribution exists if their respective distributions, $P_i(Y)$ and $P_j(Y)$, are not identical.

The difference in dataset distribution is propagated to backdoor updates through model training and can subsequently be captured by detection algorithms (Fereidooni et al.; Wang et al., 2022; Rieger et al.). This intuition is supported by the following observations (discussed in Section 5.2):

1. When the data poisoning rate is zero, the update is indistinguishable from the clean dataset.

2. As the poisoning rate increases, the difference becomes larger and is easier to be detected.

3. Moreover, the detection mechanism can capture the difference between model updates in the middle stage of training.

Theorem 1. *In federated learning, an unconverged model's parameter updates can reflect the differences in dataset distributions. Specifically, let $D(\mathcal{D}_1, \mathcal{D}_2)$ denote a measure of difference between the dataset distributions $\mathcal{D}_1$ and $\mathcal{D}_2$. As $D(\mathcal{D}_1, \mathcal{D}_2) \to 0$, let $f(\theta_{1,t+1}, \theta_{2,t+1})$ represent the difference in model updates for datasets D_1 and D_2 at iteration $t+1$, respectively. Then $f(\theta_{1,t+1}, \theta_{2,t+1}) \to 0$, indicating that as the dataset distributions converge, the model updates also become increasingly similar.*

Remark 1. *We use the cosine similarity metrics to quantify the difference between the normal update and the backdoor update. By calculating the cosine similarity of the two model update vectors, we can measure both the update difference and the dataset distribution difference.*

3.2. Attack Boundary Analysis

Detection-based defense mechanisms can identify differences in model updates. However, the distributional differences between client datasets *inherently* create a potential vulnerability that can be exploited[2]. The following provides a theoretical analysis of the attack boundary within defense mechanisms that employ detection algorithms.

The influence of the joint distribution of data and labels is visualized and analyzed in Section 5.1. When the data have similar representations but different labels, it leads to significantly different model updates. In this context, we assume that the joint distribution of data and labels is consistent and focus on the scenario where only label distribution differences exist. We then analyze the attack boundary for attackers under defense mechanisms that rely on detection algorithms.

Assumption 1. *Consider a federated learning setup with n for a classification task, with clients denoted as $C = \{c_1, c_2, ..., c_n\}$. For all categories, sampling is performed according to the Dirichlet distribution with a parameter vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$, where k is the number of categories and each category has the same number of samples. This allows us to simulate label distribution differences among clients using the Dirichlet distribution, a common approach in federated learning simulations.*

Definition 1. *For each client $c_i \in C$, a probability vector*

$$\boldsymbol{p}_i = (p_{i1}, p_{i2}, \cdots, p_{iK})^\mathsf{T}$$

can represent the dataset distribution $\mathcal{D}_i$ of client c_i, which is sampled from the Dirichlet distribution with parameters $\boldsymbol{\alpha}$. The covariance matrix of the dataset distribution of all clients is defined as $\Sigma_{N\times N}$:

$$\Sigma_{ii} = \sum_{l=1}^{k} Var(p_{il}) = \sum_{l=1}^{k} \frac{\alpha_l(\sum_{l=1}^{k}\alpha_l - \alpha_l)}{(\sum_{l=1}^{k}\alpha_l)^2(\sum_{l=1}^{k}\alpha_l + 1)}, \tag{1}$$

$$Cov(p_{il}, p_{jm}) = -\frac{\alpha_l\alpha_m}{(\sum_{l=1}^{k}\alpha_l)^2(\sum_{l=1}^{k}\alpha_l + 1)}, \tag{2}$$

where $Cov(p_{il}, p_{jm})$ is the covariance between the l-th category of client c_i and the m-th category of client c_j. So that the elements of the covariance matrix

$$\Sigma_{ij} = \sum_{l=1}^{k}\sum_{m=1}^{k} Cov(p_{il}, p_{jm}) \tag{3}$$

can be considered the measure of the difference between the dataset distributions $D(\mathcal{D}_i, \mathcal{D}_j)$ of clients c_i and c_j.

Remark 2. *From Proposition 1, there is the same trend in the dataset distribution difference and the model update difference. Consequently, the boundary in dataset distribution differences can be interpreted as the boundary in model update differences.*

[2]This becomes more evident in settings with more pronounced non-IID (Independent and Identically Distributed) data.

Theorem 2. *Under the defense mechanism with a detection algorithm, there exists a space of differences between client dataset distributions, where the poisoned dataset can be concealed, allowing the attacker to evade detection. The upper bound and the lower bound of this difference space are defined as follows:*

$$0 \leq D(\mathcal{D}_i, \mathcal{D}_j) \leq N \sum_{l=1}^{k} \sum_{m=1}^{k} \frac{\alpha_l \alpha_m}{(\sum_{l=1}^{k} \alpha_l)^2 (\sum_{l=1}^{k} \alpha_l + 1)}. \tag{4}$$

These bounds are determined by the maximum and minimum eigenvalues of the covariance matrix $\Sigma_{N\times N}$ (The details of proof can be seen in Appendix A).

Remark 3. *When $\alpha_k = \alpha$ for all k, the attack interval is given by:*

$$0 \leq D(\mathcal{D}_i, \mathcal{D}_j) \leq \frac{N}{K\alpha + 1}. \tag{5}$$

This indicates that when the data distribution deviates from IID, it becomes easier for attackers to execute successful attacks (related experiments are presented in Section 5.3).

3.3. Methodology

We discard the traditional method of increasing the weight of malicious updates during aggregation, such as increasing the number of local training epochs, raising the local training learning rate, and using a scaling factor to adjust the value of malicious updates. Instead, we maintain the same training scheduler and hyperparameter setting as in normal training and do not scale the uploaded update. Consequently, our attack is confronted with the following challenges:

1. **Evading diverse detection algorithms (resisting filtering mechanisms):** The attacker is unaware of the specific detection algorithm employed by the server. Only when the backdoor update bypasses the detection mechanism and participates in aggregation can there be a chance to implant a backdoor in the global model.

2. **Enhancing the robustness of the backdoor (resisting mitigation mechanisms):** Due to the existence of norm-clip, it is impossible to increase the weight of backdoor updates during aggregation. Moreover, the perturbation caused by adding noise can also affect the expression of the backdoor. Therefore, the implanted backdoor needs to be sufficiently robust.

From the previous analysis, Therorem 2 offers proof regarding the theoretical boundary for the dataset distribution in backdoor attacks. In practical applications, given the continuous high-dimensional features of images and the one-hot discrete features of labels, it is challenging to directly calculate and measure the joint distribution. According to Theorem 1, we can optimize the trigger to ensure that the poisoned dataset remains within this boundary by minimizing the disparity between normal updates and backdoor updates. Moreover, Theorem 1 indicates that even models not fully converged can detect the difference in dataset distributions.

Therefore, we propose PREFed which utilizes a global model that has not converged in the middle of training to optimize the trigger in advance. For challenge 1, by simulating normal training and backdoors to obtain their respective update parameters θ_c and θ_b, we will maximize their similarity as one goal to optimize the trigger, thereby reducing the difference between the poisoned dataset and the original clean dataset, the loss function is defined as follows: $\mathcal{L}_{CS} = 1 - CS(\theta_c, \theta_b)$, here we use cosine similarity as the similarity measure. In addition, for challenge 2, inspired by Pang et al. (2020), we will optimize a generic trigger by adversarial training as one goal to enhance the robustness of backdoor attacks, the loss function is defined as follows: $\mathcal{L}_{CE}(y_{\text{pred}}, y_{\text{target}})$, where we use cross entropy loss function. So, the overall optimization goal is defined as follows:

$$\min \mathcal{L}_t = \alpha * \mathcal{L}_{CE}(y_{\text{pred}}, y_{\text{target}}) + (1-\alpha)\mathcal{L}_{CS}. \tag{6}$$

The overview of PREFed is shown in Figure 1, including the following three phases: The details of PREFed are as follows:

1. **Model Warm-up:** In the early stage of model training, the attacker normally participates in the training process, which prompts the global model to contact the dataset fully. Implementing backdoor attacks at this stage will not only have an adverse impact on the learning of the main task but also interfere with the implantation of the backdoor due to the large variation of global model parameters, thereby reducing the attack's efficiency.

2. **Trigger Initialization:** In the middle stage of model training, the model has fully contacted the dataset and can capture the difference in dataset distributions. At this time, the attacker uses the backup global model of an arbitrary round to simulate normal training and backdoor training respectively, and initializes the trigger as shown in Equation 6 to improve the efficiency of backdoor attacks.

3. **Malicious Updates Uploading and Trigger Refinement:** In the later model training stage, the attacker began to manipulate the client to launch backdoor attacks. At the same time, to adapt to the dynamic changes of the global model, the attacker continued to adjust the trigger to further improve the attack effect.

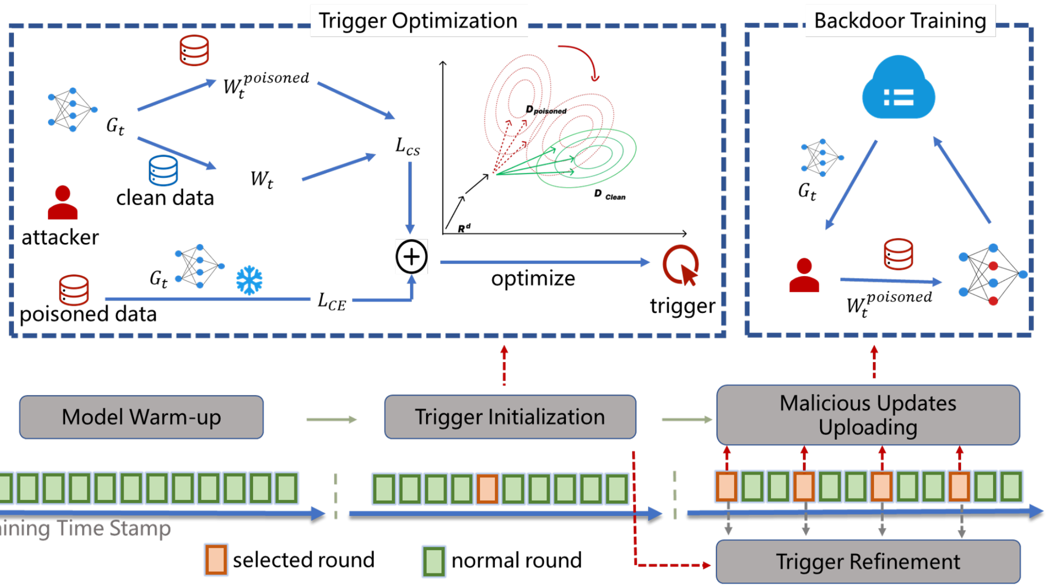


Figure 1: The overview of PREFed. The method includes three phases: model warm-up, trigger initialization, and malicious updates uploading and trigger refinement. We also introduce the PreFed, which only implements the malicious updates uploading without trigger refinement.

The detail of the PREFed algorithm is shown in Algorithm 1 in the Appendix. In addition, we also introduce the PreFed method, which only implements the backdoor attack in the later stage of model training without trigger refinement. Later experiments will show it also improves the attack success rate.

4. Experiments

In this section, we conduct experiments under six advanced existing defense mechanisms (RFLBAT, FoolsGold, FLDetector, DeepSight, FLAME, FreqFed) to highlight PREFed's outstanding effectiveness and efficiency on three benchmark datasets (Cifar10, Cifar100, and tiny-Imagenet (Krizhevsky et al., 2009; Le & Yang, 2015)). In addition, we use the backdoor accuracy (BA) and the attach success rate (ASR) to measure the performance of the backdoor attack, and the main task accuracy (MA) to measure the performance of the main task [3].

Following by Li et al. (2023), we kept the poisoning rate at 0.3 and the scaling factor 3, and set the concentration parameter of Dirichlet distribution to 0.9 to simulate the non-IID data distribution across client sides in real-world scenarios (Hsu et al., 2019; Sattler et al., 2019). The rate of compromised clients was set to 0.2, and the number of clients was set to 100.

[3]The code is developed based on Li et al. (2023). It corrects the errors in the DeepSight module and adds the implementation of the DBA attack algorithm and the FreqFed defense algorithm.

Table 1: Experimental statistics on the number of successful attack trials under different defense mechanisms. Each experiment consists of 10 trials, with an attack considered successful if the backdoor accuracy exceeds 50% in the final testing round.

F

Defense\Attack	3DFed	DBA	ModelReplace	PreFed	PREFed
Deepsight	8	10	6	9	9
FLAME	2	0	0	10	10
FLDetector	9	8	9	9	10
Foolsgold	1	10	9	7	10
FreqFed	1	10	6	10	10
RFLBAT	2	0	0	7	7
FedAvg	8	9	9	10	10

4.1. Experimental Results

4.1.1. ASR of Various Attacks

On the Cifar-10 dataset, we performed 10 arbitrary experiments under six existing defense methods [4]. Table 1 shows that our attack method has generally improved the attack success rate compared to previous methods, further refinement of the trigger can significantly improve the attack success rate by comparing PreFed with PREFed.

It is worth noting that even under the FedAvg aggregation rule without a defense mechanism, the attack method based on the scaled update value cannot achieve a 100% success rate. This is because while scaling for model updates can

[4]The final result can be seen in Table 7 for details in the appendix.

Table 2: The performance of different attack methods (ModelReplace (MR), Distributed Backdoor Attack (DBA), 3DFed, and PREFed) under various defense methods (RFLBAT, FoolsGold, FLDetector, DeepSight, FreqFed, and FLAME). The red data indicate the best effect and the underlined data indicate the second-best effect.

	Defense	RFLBAT		FoolsGold		FLDetector		DeepSight		FreqFed		FLAME		Avg.	
Dataset	Attack\Metric(%)	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA
Cifar-10	MR	10.83	79.28	98.65	71.28	97.31	74.28	99.25	65.96	98.47	60.17	10.5	73.02	69.17	70.67(-12.98%)
	DBA	45.54	79.10	99.80	59.74	99.94	73.96	98.45	77.11	99.99	48.88	11.45	72.80	75.86	68.60(-15.5%)
	3DFed	65.84	79.10	80.65	76.41	84.01	25.85	88.64	78.80	62.57	74.10	86.28	69.29	78.00	67.26(-17.18%)
	PREFed	97.4	79.12	96.97	76.25	99.12	80.53	99.08	79.33	97.6	76.18	99.74	68.62	98.32	76.67(-5.58%)
	w/o	10.44	81.54	10.26	80.82	10.33	81.68	10.22	81.64	10.53	80.49	10.50	81.06	10.38	81.21
Cifar-100	MR	61.43	24.26	4.53	51.87	99.24	30.47	1.00	52.03	99.90	49.13	0.96	51.20	44.51	43.16(-17.08%)
	DBA	0.76	51.54	0.85	51.90	16.46	25.46	0.73	52.11	37.19	27.40	0.78	51.12	9.46	43.26(-16.90%)
	3DFed	4.96	49.51	92.47	51.68	90.61	50.88	0.88	51.91	5.48	51.81	1.22	51.63	32.60	51.24(-1.57%)
	PREFed	74.34	50.55	99.05	51.60	97.76	51.70	93.29	52.00	89.04	51.46	99.48	50.42	92.16	51.29(-1.47%)
	w/o	1.00	52.44	0.82	51.93	0.94	52.07	0.92	52.22	0.72	52.23	0.80	51.42	0.87	52.05
Tiny - Imagenet	MR	7.77	68.12	0.52	70.96	0.53	70.75	0.54	71.14	0.67	70.85	0.50	70.90	1.76	70.45(-1.00%)
	DBA	29.22	59.09	0.52	71.13	100.00	69.70	0.55	71.24	0.62	70.97	0.54	70.83	21.91	68.83(-3.29%)
	3DFed	8.50	70.75	0.52	70.94	97.62	70.64	0.51	70.95	0.63	70.82	0.53	70.89	18.05	70.83(-0.47%)
	PREFed	99.99	70.92	99.99	71.14	99.99	70.75	99.46	71.46	100.00	70.74	99.99	70.83	99.90	70.97(-0.27%)
	w/o	0.54	71.22	0.53	71.21	0.54	71.25	0.54	71.21	0.52	71.05	0.54	71.06	0.54	71.17

greatly improve attack efficiency, it can lead to numerical stability issues. Specifically, over-scaling can cause numerical overflows, causing the model's computational results to become unstable during inference. F

4.1.2. Comparison with Baseline

From Table 2, it can be noticed that our method achieves more than 90% attack accuracy under different defense mechanisms. Even if we cannot achieve the highest BA in all scenarios, the overall attack effect is the best. In addition, our method also has the smallest loss on the main task (5.58% reduction on Cifar-10, 1.47% reduction on Cifar-100, and 0.27% reduction on Tiny-Imagenet).

It is worth noting that in some cases, BA cannot be suboptimal simultaneously as MA in other attack methods. This is because scaling-based backdoor attacks can compromise the performance of the main task during training. More seriously, as can be seen from Figure 9, the training method based on scaling updates becomes unstable. If the scaling factor selected is too large, the model update may lead to numerical overflow.

In addition, we also reproduced A3FL and Backdoor-Critical layer attacks (can been in the appendix C.2), which are the advanced trigger-optimization attacks and adaptive attacks respectively. From Figure 6 and 8, PREFed is superior in terms of attack effectiveness and efficiency.

4.1.3. Attack Efficiency

Our experiments evaluated attack efficiency by measuring both time cost and the number of attack rounds. The improvement of attack efficiency has more practical impacts:

Table 3: Comparison of time overhead per round for different attacks under FLAME framework on Cifar-10 dataset.

Attack	DBA	ModelReplace	3DFed	PREFed
Time(s)	5.35	7.07	26.75	4.53

In the training process, even if the training equipment is offline or fails to catch up with the timestamp due to some reasons, the attack task can be completed in a shorter time, thereby reducing the loss caused by attack interruption due to unexpected situations and improving the success rate and stability of attacks.

Figure 9 shows the performance of different attack methods under various defense mechanisms on the Cifar-10 dataset, which shows that the attack accuracy of PREFed rose to over 80% within 5 rounds. In addition, Table 3 shows the per-round time consumed attack on the Cifar-10 dataset. The result shows that PREFed outperforms other attack methods in terms of time cost, an approximately 14.56% improvement over DBA, about 36.03% over ModelReplace, and roughly 82.94% over 3DFed.

Overall, PREFed not only proves to be enough effective in backdoor attacks but also displays significant advantages in terms of attack efficiency.

5. Further Analysis

In this section, we conduct a detailed analysis to unveil the relationship between dataset distribution differences and model update differences, further deepen our understanding of backdoor attacks.

Table 4: The performance of PREFed with different poison rates. The red data represents the bad cases where the attack was unsuccessful.

Poison Rate	0.1		0.2		0.3		0.4		0.5		0.8		1.0	
Defense\Accuracy(%)	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA
Deepsight	81.13	91.8	81.35	95.77	80.46	98.92	81.04	98.22	80.95	99.1	81.34	13.4	81.4	7.71
FLAME	78.02	98.02	80.35	98.04	80.92	99.51	80.0	98.81	78.78	99.53	81.03	17.62	81.39	10.78
FLDetector	81.34	94.55	80.48	98.11	81.03	99.34	80.19	97.78	79.48	98.36	81.45	85.09	81.33	12.86
Foolsgold	80.42	92.31	80.42	97.98	80.93	99.34	80.5	98.75	80.55	99.19	81.14	14.06	80.86	7.17
FreqFed	80.13	95.11	80.53	97.8	81.35	98.91	80.17	99.26	79.26	99.76	80.51	97.24	80.87	11.08
RFLBAT	81.12	94.05	80.77	98.72	80.93	98.39	81.24	62.74	80.75	99.0	80.55	15.1	81.22	7.64
Avg.	80.36	94.31	80.65	97.74	80.94	99.07	80.52	92.59	79.96	99.16	81.00	40.42	81.18	9.54

5.1. Visualization and Analysis

We analyze the impact of data-label joint distributions by visualizing data representations and model updates.

Figures 2 and 3 respectively represent the visualization of 3DFed and PREFED on Cifar-10 under the FLAME defense. This includes the t-SNE result of clean and poisoned datasets, as well as the PCA visualization graphs of model updates during normal and backdoor training. 3DFed uses a patch-size image as the trigger, while PREFED uses the global-size trigger and constructs poisoned data in blend form. In order to better distinguish the target label from the original label, we set the target label to 10 (a new label).

The observed results show that when 3DFed uses a patch as the trigger to poison the dataset, the representation of data is highly similar to the normal dataset. This is because the patch does not significantly affect the overall data characteristics. However, due to the label flip (from the source label to the target label), significant changes have occurred in model updates. Furthermore, to establish the association between the target label and the trigger, the malicious updates of the damaged client become more concentrated, as shown in Figure 2d.

In contrast, the data poisoned by PREFED is clearly distinguished from other clean data, but its malicious model updates are closer to normal updates and more concealed. As shown in Figure 3d, malicious updates can be better hidden among other model updates.

Takeaway 1: Attackers seeking to enhance the concealment of backdoor attacks should focus on the consistency of data-label joint distributions instead of data features.

5.2. Poison Rate

We tested the performance of our method under different poisoning rates, still using attack accuracy over 50% as the criterion for attack success.

Tabel 4 showed that with the increase in poisoning rate (from 0.1 to 0.5), the attack accuracy overall showed an upward trend. However, at the poisoning rate of 0.4, the overall attack accuracy decreased slightly due to the accuracy of only 62.74% under the RFLBAT defense mechanism. This shows that under the premise of breaking through the defense, increasing the poisoning rate usually enhances the attack effect. However, when the poisoning rate continues to increase to a higher level, the attack effect decreases.

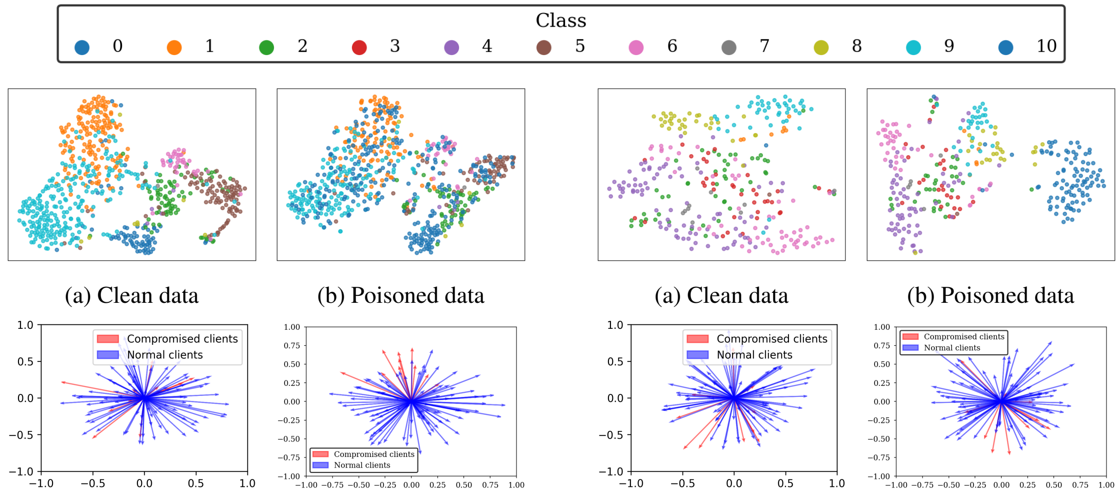
At the poisoning rate of 0.8 and 1.0, the average attack accuracy decreases to 40.42% and 9.54%, respectively. This phenomenon shows that when the poisoning rate is too high, the distribution difference of the dataset becomes too obvious, resulting in the malicious model updates being easily recognized by the detection mechanism.

Takeaway 2: PREFed achieves high attack accuracy with a lower poisoning rate. However, excessively high poisoning rates can reduce the attack's success, highlighting that dataset distribution differences are indeed reflected in model updates and can be detected by defense algorithms.

5.3. Non-IID Data

We delve into the impact of non-IID data on the PREFed attack strategy. The hyperparameter α controls Dirichlet distribution, when $\alpha \to \infty$, the data distribution of all clients is identical to the prior distribution and is completely in line with IID.

According to Table 5, when the data distribution is closer to being non-IID ($\alpha < 1$), we can observe that PREFed demonstrates exceptional stability, maintaining a backdoor accuracy rate above 90%. However, when the $\alpha < 0.5$, the accuracy of the primary task is affected to a certain extent, with a decline ranging from 10% to 30%. It is noteworthy that MA will also significantly decrease due to the non-IID data even without attacking. For instance, at $\alpha = 0.1$, the main task accuracy drops by 48.27% compared to $\alpha = 0.9$.



(a) Clean data (b) Poisoned data (a) Clean data (b) Poisoned data

(c) Normal training (d) Backdoor training (c) Normal training (d) Backdoor training

Figure 2: 3DFed under FLAME on Cifar-10

Figure 3: PREFed under FLAME on Cifar-10

Figure 4: Visualization of data distribution and model updates in normal vs. backdoor training. This figure shows the data distribution and model updates at two stages: before the attack begins and during the first round of the attack.

Table 5: The impact of non-IID data for PREFed under FLAME defense on the Cifar-10 dataset.

α	0.1		0.3		0.5		0.7		0.9		10		20		50		100	
Attack\Accuracy(%)	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA
PREFed	27.82	98.61	57.14	91.02	69.67	94.06	71.16	96.55	72.48	97.53	77.97	25.88	77.76	20.52	78.25	22.55	77.97	22.78
w/o	37.93	28	65.01	11.32	70.34	14.37	73.51	9.45	73.33	11.06	78.19	10.82	77.72	10.85	77.81	10.25	78.26	10.57

Compared to the impact of PREFed on the accuracy of the primary task, the non-IID data distribution has a more pronounced effect on the primary task. Moreover, when the data distribution is closer to being IID ($\alpha \geq 10$), the behavior of attack was detected and the attack effect is poor (the backdoor task accuracy is below 30%).

Takeaway 3: In practical scenarios, the non-IID data provide a larger attackable interval for attackers, which does not affect the attack effectiveness of PREFed and has a greater impact on the performance of the main task.

6. Limitaions and Future Work

Attacker Perspective: Our experiments focus primarily on image classification tasks, as PREFed is not yet well-suited for other domains. In fields like text and tabular data, their discrete nature poses challenges for designing triggers that are both effective and inconspicuous. Addressing these limitations and adapting PREFed will be a key focus of future research.

Defense Perspective. Although existing defense mechanisms increase the difficulty of executing backdoor attacks, they are not foolproof. Many require either a higher number of compromised clients or rely on elevated poisoning rates to mitigate attacks. PREFed's success with minimal resources highlights the inadequacy of existing defenses during training and the pressing need for more advanced.

7. Conclusion

In this paper, we revisited existing defense mechanisms based on anomaly detection in model updates, and explored the relationship between data distribution differences and the resulting model update discrepancies, offering a theoretical basis to understand the vulnerabilities in current defense strategies. Our analysis shows that larger gaps in client dataset distributions create broader attackable intervals, making it easier for attackers to implant backdoors.

Building on this understanding, we introduced PREFed, a novel backdoor attack method that pre-optimizes the attack trigger, significantly enhancing attack efficiency. Our experimental results demonstrate that it outperforms existing advanced attack methods in terms of both effectiveness and efficiency under current defenses.

References

Aono, Y., Hayashi, T., Wang, L., Moriai, S., et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE transactions on information forensics and security*, 13(5):1333–1345, 2017.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pp. 2938–2948. PMLR, 2020.

Campello, R. J., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer, 2013.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Fereidooni, H., Pegoraro, A., Rieger, P., Dmitrienko, A., and Sadeghi, A.-R. FreqFed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning. In *Proceedings 2024 Network and Distributed System Security Symposium*. Internet Society. ISBN 978-1-891562-93-8. doi: 10.14722/ndss.2024.24620. URL `https://www.ndss-symposium.org/wp-content/uploads/2024-620-paper.pdf`.

Fu, C., Zhang, X., Ji, S., Wang, T., Lin, P., Feng, Y., and Yin, J. {FreeEagle}: Detecting complex neural trojans in {Data-Free} cases. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 6399–6416, 2023.

Fung, C., Yoon, C. J., and Beschastnikh, I. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, pp. 301–316, 2020.

Gong, X., Chen, Y., Huang, H., Liao, Y., Wang, S., and Wang, Q. Coordinated backdoor attacks against federated learning with model-dependent triggers. *IEEE network*, 36(1):84–90, 2022.

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

Islam, T. U., Ghasemi, R., and Mohammed, N. Privacy-preserving federated learning model for healthcare data. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0281–0287. IEEE, 2022.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Li, H., Ye, Q., Hu, H., Li, J., Wang, L., Fang, C., and Shi, J. 3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1893–1907. IEEE, 2023.

Maćkiewicz, A. and Ratajczak, W. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017a.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017b.

Miao, Y., Zheng, W., Li, X., Li, H., Choo, K.-K. R., and Deng, R. H. Secure model-contrastive federated learning with improved compressive sensing. *IEEE Transactions on Information Forensics and Security*, 18:3430–3444, 2023.

Naseri, M., Hayes, J., and De Cristofaro, E. Local and central differential privacy for robustness and privacy in federated learning. *arXiv preprint arXiv:2009.03561*, 2020.

Nguyen, T. D., Rieger, P., Chen, H., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., Mirhoseini, A., Zeitouni, S., Koushanfar, F., Sadeghi, A.-R., and Schneider, T. FLAME: Taming backdoors in federated learning (extended version 1). URL `http://arxiv.org/abs/2101.02281`.

Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., and Sadeghi, A.-R. Dïot: A federated self-learning anomaly detection system for iot. In *2019 IEEE 39th International conference on distributed computing systems (ICDCS)*, pp. 756–767. IEEE, 2019.

Nguyen, T. D., Rieger, P., Miettinen, M., and Sadeghi, A.-R. Poisoning attacks on federated learning-based iot intrusion detection system. In *Proc. Workshop Decentralized IoT Syst. Secur.(DISS)*, volume 79, 2020.

Pang, R., Shen, H., Zhang, X., Ji, S., Vorobeychik, Y., Luo, X., Liu, A., and Wang, T. A tale of evil twins: Adversarial inputs versus poisoned models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 85–99, 2020.

Rieger, P., Nguyen, T. D., Miettinen, M., and Sadeghi, A.-R. DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection. In *Proceedings 2022 Network and Distributed System Security Symposium*. Internet Society. ISBN 978-1-891562-74-7. doi: 10.14722/ndss.2022.23156. URL `https://www.ndss-symposium.org/wp-content/uploads/2022-156-paper.pdf`.

Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

Wang, Y., Zhai, D., Zhan, Y., and Xia, Y. Rflbat: A robust federated learning algorithm against backdoor attack. *arXiv preprint arXiv:2201.03772*, 2022.

Xie, C., Huang, K., Chen, P.-Y., and Li, B. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*, 2019.

Xu, G., Li, H., Ren, H., Yang, K., and Deng, R. H. Data security issues in deep learning: Attacks, countermeasures, and opportunities. *IEEE Communications Magazine*, 57 (11):116–122, 2019.

Zhang, H., Jia, J., Chen, J., Lin, L., and Wu, D. A3fl: Adversarially adaptive backdoor attacks to federated learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhuang, H., Yu, M., Wang, H., Hua, Y., Li, J., and Yuan, X. Backdoor federated learning by poisoning backdoor-critical layers. *arXiv preprint arXiv:2308.04466*, 2023.

A. The Proof of Theorem

The following is the detailed proof of Theorem 1

Proof. Let the model be $f(x;\theta)$, where x is the input data and θ represents the model parameters. For two different datasets D_1 and D_2, the loss corresponding functions are $L_1(\theta)$ and $L_2(\theta)$, respectively.

The update formula for the model parameters is $\theta_{t+1} = \theta_t - \alpha\frac{\partial L}{\partial \theta_t}$, where α is the learning rate. For dataset D_1, the parameter update is $\theta_{1,t+1} = \theta_t - \alpha\frac{\partial L_1}{\partial \theta_t}$; for dataset D_2, the parameter update is $\theta_{2,t+1} = \theta_t - \alpha\frac{\partial L_2}{\partial \theta_t}$.The difference between the two updates is: [5]

$$\begin{aligned} f(\theta_{1,t+1},\theta_{2,t+1}) &= |(\theta_t - \alpha\frac{\partial L_2}{\partial \theta}) - (\theta_t - \alpha\frac{\partial L_1}{\partial \theta})| \\ &= |\alpha(\frac{\partial L_2}{\partial \theta} - \frac{\partial L_1}{\partial \theta})|. \end{aligned} \tag{7}$$

As the dataset distributions become more consistent, it can be assumed that $L_1(\theta)$ and $L_2(\theta)$ approach each other, i.e., $|L_1(\theta) - L_2(\theta)| \to 0$. The dataset distribution can be reflected in the model's parameter updates, it follows that $\frac{\partial L_2}{\partial \theta} - \frac{\partial L_1}{\partial \theta} \to 0$. Therefore, the difference between the model updates, $f(\theta_{1,t+1},\theta_{2,t+1})$, will also tend to 0. □

The following is the complete proof for Theorem 1. We use the Gershgorin circle theorem to approximate the maximum eigenvalue of $\Sigma_{N\times N}$ which is the upper bound of the attack interval.

Proof. From the Definition 1, the process of proving process is as follows:

Step 1: Since Σ is a non-negative matrix, the minimal eigenvalue of $\Sigma_{N\times N}$ is close to zero. We can approximate the minimum eigenvalue as:

$$\lambda_{\text{min}}(\Sigma_{N\times N}) \geq 0.$$

Step 2: The maximum eigenvalue of $\Sigma_{N\times N}$ can be approximated by the Gershgorin circle theorem. The theorem states that the eigenvalues of a matrix are located in the union of the Gershgorin circles:

$$\lambda \in \bigcup_{i=1}^{N} \left\{ z \in \mathbb{C} : |z - \Sigma_{ii}| \leq \sum_{j=1, j\neq i}^{N} |\Sigma_{ij}| \right\}, \tag{8}$$

where each Gershgorin circle is centered at Σ_{ii} with radius $\sum_{j\neq i} |\Sigma_{ij}|$.

$$\lambda_{\text{max}}(\Sigma_{N\times N}) \leq \max_i \left(\Sigma_{ii} + \sum_{j\neq i} |\Sigma_{ij}| \right), \tag{9}$$

Step 3: Since the absolute value of $|\Sigma_{ij}|$:

$$|\Sigma_{ij}| \leq \sum_{l=1}^{k} \sum_{m=1}^{k} \frac{\alpha_l \alpha_m}{(\sum_{l=1}^{k} \alpha_l)^2 (\sum_{l=1}^{k} \alpha_l + 1)}. \tag{10}$$

[5]Here we simplify the expression of the differences between updates.

the maximum eigenvalue can be approximated as:

$$\lambda_{\text{max}} \leq N \sum_{l=1}^{k} \sum_{m=1}^{k} \frac{\alpha_l \alpha_m}{(\sum_{l=1}^{k} \alpha_l)^2 (\sum_{l=1}^{k} \alpha_l + 1)}. \tag{11}$$

Consequently, the lower and upper bounds for the differences in client data distributions can be expressed as equation 4. Then the proposition is proved. □

B. PREFed Algorithm

The following is the complete algorithm of PREFed. The algorithm is designed to optimize the triggers in advance to improve the attack efficiency.

Algorithm 1 PREFed on Client

Input: Model architecture G; Model parameters θ_r; Client dataset D_c; Trigger from last round T_{r-1}
Output: Optimized triggers T_{r+1}; Poisoned model parameter updates Θ_{r+1}

Initialize $\Theta \leftarrow \varnothing, T \leftarrow \varnothing$ **for** $A_i \in$ *Attackers* **do**
  $\theta' \leftarrow \theta_r$
  $\theta^* \leftarrow \theta'$
  $D_p \leftarrow \text{Poison}(D_c, T)$
  Initialize $t \leftarrow T$
  $\theta_c \leftarrow \text{Training}(G, \theta', D_c)$
  $\theta'_p \leftarrow \text{Training}(G, \theta_c, D_p)$
  $t_{\text{opt}} \leftarrow \text{TriggerOptimize}(G, \theta'_p, \theta_c, D_p, T)$
  Add $(\theta'_p - \theta_r)$ to Θ_{r+1}
  Add t_{opt} to T_{r+1}
end

Function `Training` (G, θ, D) **:**
  for $i \in E$ **do**
    for $(x, y) \in D$ **do**
      $y_{\text{pred}} \leftarrow G_\theta(x)$
      $\theta \leftarrow \theta - \eta \nabla L_{CE}(y_{\text{pred}}, y)$
    end
  end
  return θ

Function `TriggerOptimize` $(G, \theta_p, \theta_c, D_c, T)$ **:**
  for $i \in E$ **do**
    for $((x_c, y_c), (x_p, y_p)) \in (D_c, D_p)$ **do**
      $(y_c^{\text{pred}}, y_p^{\text{pred}}) \leftarrow (G_{\theta_c}(x_c), G_{\theta_p}(x_p))$
      $L_{CS} = 1 - \text{CosineSimilarity}(\Delta\theta_c, \Delta\theta_p)$
      $L_1 \leftarrow \alpha \cdot L_{CE}(y_p^{\text{pred}}, y_p) + (1 - \alpha) \cdot L_{CS}$
      $T \leftarrow T - \nabla C_t$
    end
  end
  return t_{opt}

C. The Implementation of Experiments

Federated Learning Setup. The global model uses the ResNet-18 architecture, with a pre-trained model for the Tiny-Imagenet dataset task. To simulate non-IID data distribution in the real world, we use Dirichlet distribution, setting the concentration parameter to 0.9 in the main experiment (consistent with previous studies (Li et al., 2023)). In subsequent sensitivity experiments, we analyze the impact of data non-IID by adjusting this parameter. In each communication round, each client trains the local model for 2 epochs using the SGD optimizer. The entire global training process lasts for 220 communication rounds.

Attacker Setup. Referring to the mainstream experimental settings (Zhang et al., 2024; Li et al., 2023; Zhuang et al., 2023), in the main experiment, we set 20% of the clients to be controlled by attackers and set the poisoning rate of each damaged client data set to 30%. Here, we use a global trigger and carry a blend strategy with a parameter setting of 0.1:

$$\text{Blended Image} = 0.1 \times \text{Trigger} + 0.9 \times \text{Image},$$

which is a small value and does not affect the visual appearance of the images (Chen et al., 2017; Fu et al., 2023).

C.1. The Number of Initial Epochs

We investigate the impact of initial trigger optimization on PREFed using the Cifar-10 dataset and the FLAME defense mechanism. Our analysis highlights how pre-optimizing the trigger can significantly enhance the efficiency of backdoor attacks.

Table 6 shows the effect of the number of compromised clients on the performance of PREFed, based on experiments conducted with the Cifar-10 dataset and the FLAME method. In the experiment, a total of 100 clients participated in the training. The results indicate that when the number of compromised clients is small, the attack's effectiveness is limited. For example, when only one client is compromised, BA is only 32.87%, and with two compromised clients, BA further drops to 13.75%.

Further analysis reveals that when the number of compromised clients exceeds 8, or more than 8% of the total clients, PREFed's effectiveness increases significantly, achieving a success rate above 90%. This demonstrates that PREFed can achieve high attack performance when a sufficient proportion of clients are compromised, highlighting the method's reliance on the number of controlled clients during execution.

C.2. A3FL and Backdoor-Critical Layer Attack

Figure 6 shows the performance of A3FL (Zhang et al., 2024) under FLAME and FreqFed defense mechanism, and

Figure 7 and Figure 8 show the performance of Backdoor-Critical Layer attack (BC) (Zhuang et al., 2023) in different parameter settings. From these figures, both attacks need more attack rounds.

A3FL attack enhances effectiveness through carefully designed triggers, focusing particularly on the dynamic changes of the global model to strengthen the persistence of backdoor attacks. The Backdoor-Critical Layer Attack adopts a more analytical approach by assessing the contribution of each layer in the backdoor model to its effect and sorting them based on their influence. In this method, specific layers of a well-trained benign model are replaced with corresponding layers in the backdoor model, where the rank n of replacement layers is a parameter that can be dynamically adjusted as needed. These complex computations need more time to attack.

Through these experiments, we further confirm the outstanding performance of PREFed in terms of attack efficiency and effectiveness.

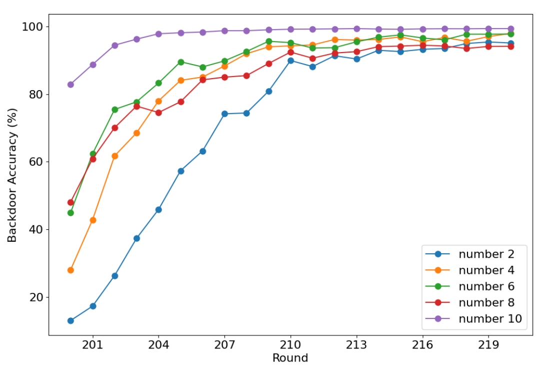


Figure 5: The impact of the number of initial epochs for PREFed under FLAME on Cifar-10.

C.3. The Number of Compromised Clients

Table 6 shows the effect of the number of compromised clients on the performance of PREFed, based on experiments conducted with the Cifar-10 dataset and the FLAME method. In the experiment, a total of 100 clients participated in the training. The results indicate that when the number of compromised clients is small, the attack's effectiveness is limited. For example, when only one client is compromised, BA is only 32.87%, and with two compromised clients, BA further drops to 13.75%.

Further analysis reveals that when the number of compromised clients exceeds 8, or more than 8% of the total clients, PREFed's effectiveness increases significantly, achieving a success rate above 90%. This demonstrates that PREFed can achieve high attack performance when a sufficient proportion of clients are compromised, highlighting the method's reliance on the number of controlled clients during execution.

C.4. Compared with Baseline

The following figures 9, 10 and 11 represent the performance of PREFed compared with other attacks under six advanced defenses on Cifar10, Cifar100 and Tint-Imagenet datasets. The attacker begins to upload malicious updates at the start of the 201st round of training. PREFed uses the global model of the 100th round to optimize the trigger in advance.

C.5. The Results of Ten Experiments

Table 7 is the result of 10 experiments, which shows the performance of 3DFed, DBA, ModelReplace (MR), PreFed, and PREFed under different defense mechanisms on the Cifar-10 dataset.

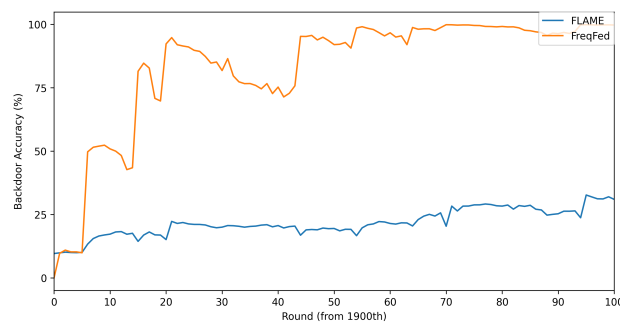
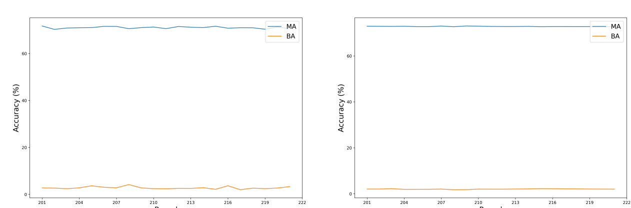


Figure 6: The backdoor accuracy of A3FL with FLAME and FreqFed defense mechanism. The backdoor attack starts from the 1900th round.

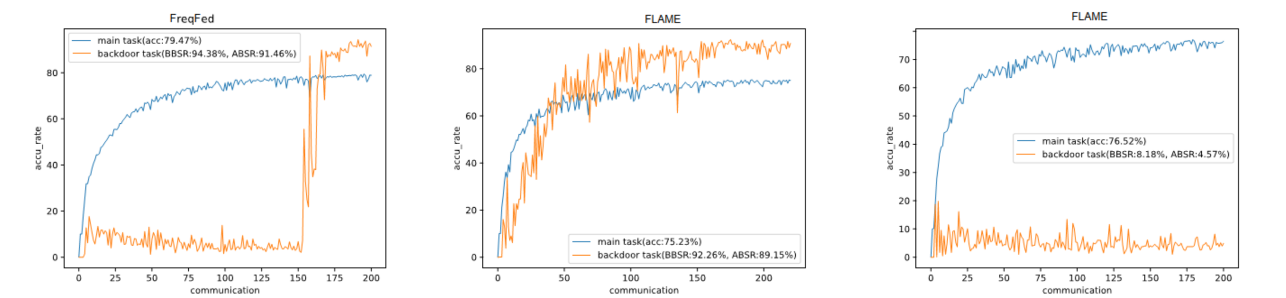


(a) FLAME (b) FreqFed

Figure 7: The performance of BC attack under FLAME and FreqFed defenses on Cifar-10. The setting is aligned with PREFed, and implementing the backdoor attack starts from the 201st round and ends at the 200th round.

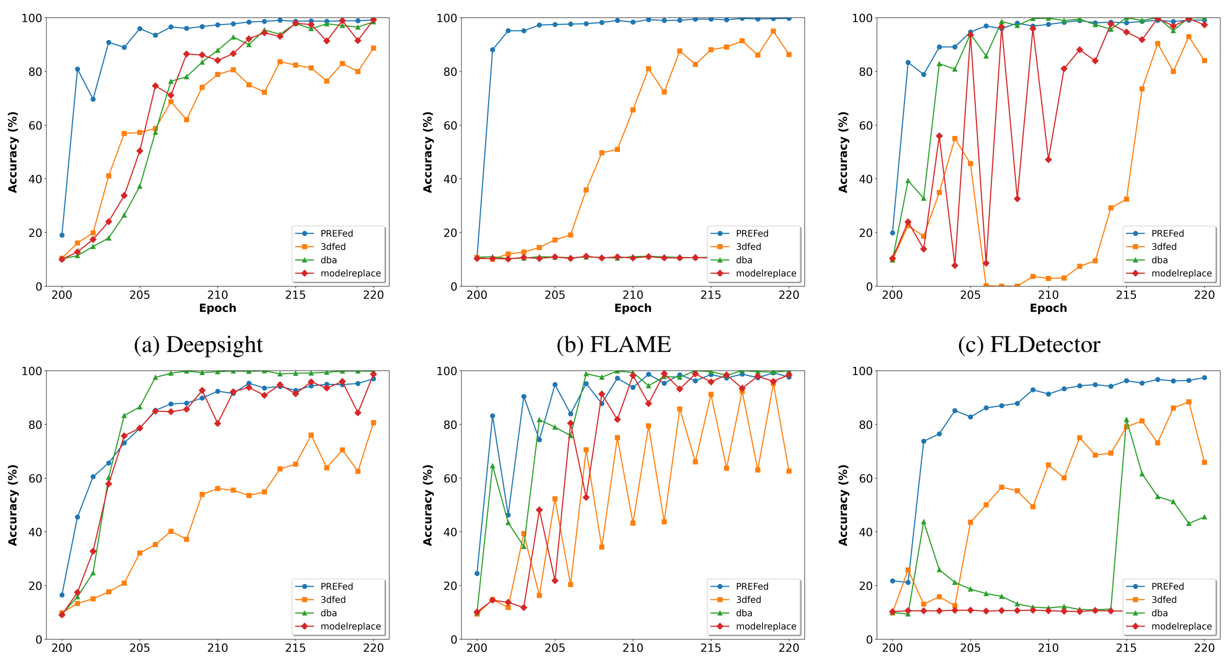
Table 6: The impact of the numbers of compromised clients for PREFed under FLAME on Cifar-10.

Number	1		2		4		5		6		8		10	
Defense\Accuracy(%)	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA
FLAME	81.44	32.87	80.95	13.75	81.56	53.83	81.36	89.46	81.44	68.08	80.65	97.79	80.79	98.90



(a) (b) (c)

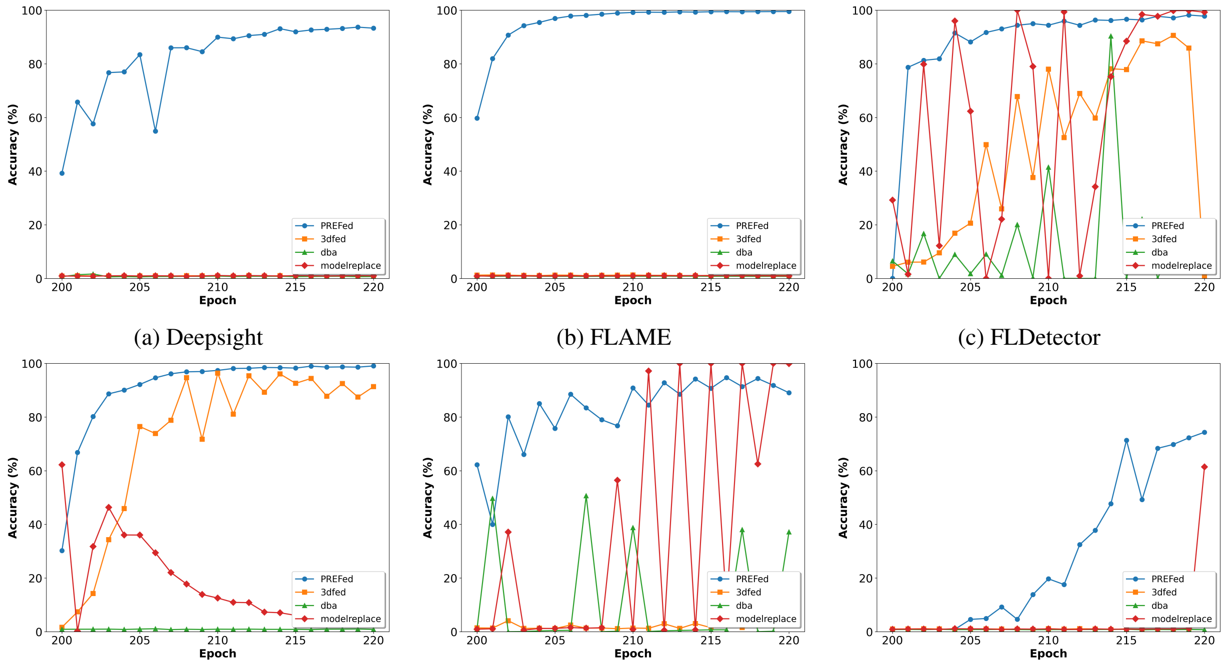
Figure 8: The performance of the Backdoor-Criticial layer attack on the Cifar-10 dataset as the number of rounds increases. (a) Adversary attacks from round 0 and end at round 200 with FreqFed; (b) Adversary attacks from round 0 and end at round 200 with FLAME; (c) Adversary attacks from round 180 and end at round 200 with FLAME.



(a) Deepsight (b) FLAME (c) FLDetector

(d) Foolsgold (e) FreqFed (f) RFLBAT

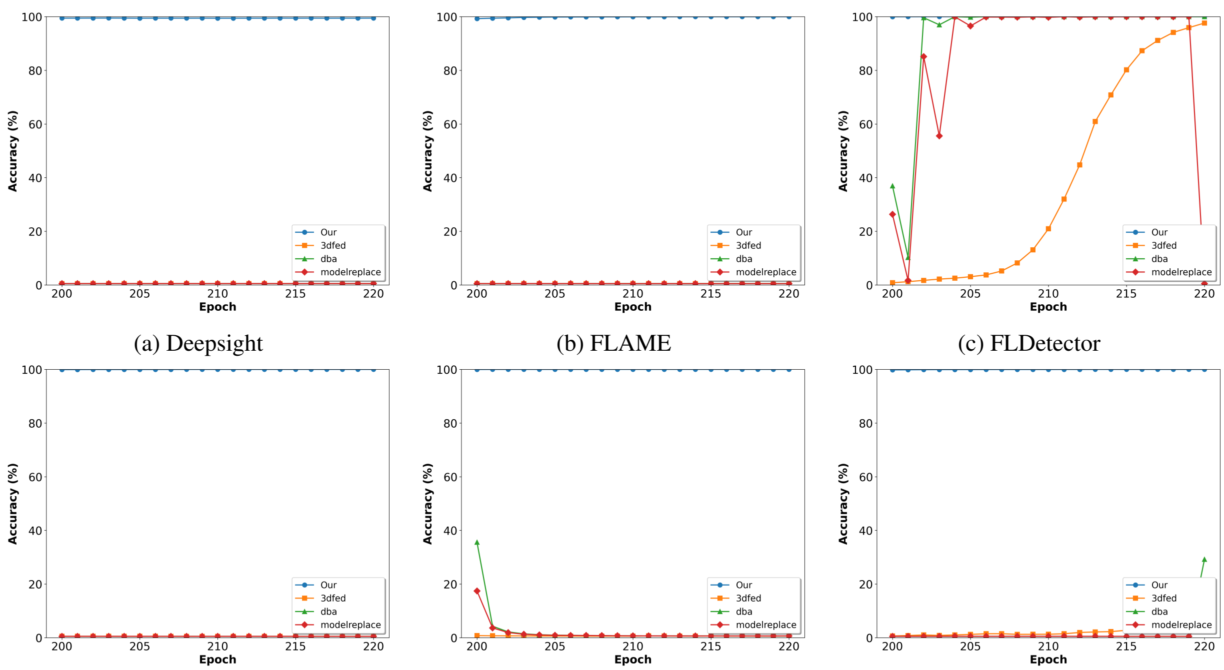
Figure 9: The performance of different attack methods under various defenses on Cifar-10.



(a) Deepsight (b) FLAME (c) FLDetector

(d) Foolsgold (e) FreqFed (f) RFLBAT

Figure 10: The performance of different attack methods under various defenses on Cifar-100.



(a) Deepsight (b) FLAME (c) FLDetector

(d) Foolsgold (e) FreqFed (f) RFLBAT

Figure 11: The performance of different attack methods under various defenses on Tiny-Imagenet.

Table 7: The results of 10 experiments of different attack methods under various defenses on Cifar-10.

		1		2		3		4		5		6		7		8		9		10	
Defense	Attack	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA	MA	BA
Deesight	3DFed	78.26	64.91	78.66	73.21	78.46	80.94	78.07	55.10	78.12	67.97	78.80	88.64	79.83	20.75	79.43	64.79	79.71	75.07	79.75	31.71
	DBA	79.23	89.29	74.67	94.39	77.11	98.45	79.49	52.59	68.04	93.41	76.24	96.93	79.01	65.68	69.07	92.87	75.78	96.41	76.94	86.08
	MR	65.96	99.25	79.36	10.33	75.81	96.71	76.31	97.87	79.13	10.23	80.09	33.93	79.73	9.64	72.72	94.34	80.43	10.48	73.37	97.82
	PreFed	79.32	92.91	79.97	88.63	79.47	91.93	79.58	89.89	79.02	23.86	79.47	91.91	78.84	77.91	79.60	92.65	79.91	90.11	79.35	78.16
	PREFed	79.84	89.16	79.03	96.84	79.03	98.09	79.15	94.39	79.46	98.22	79.71	96.07	79.33	99.08	79.07	97.57	79.15	97.30	79.81	18.87
FLAME	3DFed	68.65	9.25	70.30	47.01	72.13	10.45	69.29	86.28	71.35	10.56	73.17	84.94	71.32	11.04	71.71	10.92	71.64	10.57	71.00	10.91
	DBA	72.64	8.90	71.79	10.23	72.01	10.75	72.72	10.69	72.80	11.45	74.03	10.08	72.10	11.45	73.04	10.77	73.17	9.46	72.32	9.39
	MR	71.52	9.94	73.31	10.17	73.02	10.50	73.24	10.22	72.82	9.67	73.71	10.28	72.45	10.23	73.27	10.15	72.73	9.13	73.13	8.28
	PreFed	71.26	82.64	71.62	94.33	70.96	90.82	70.47	95.17	69.92	94.25	71.16	94.44	70.59	93.84	72.45	79.96	72.04	82.15	69.73	84.99
	PREFed	68.20	99.29	70.48	98.85	68.62	99.74	71.61	98.64	70.68	98.17	71.64	97.90	70.86	96.89	71.07	97.60	71.71	98.38	71.13	97.53
FLDetector	3DFed	72.24	58.09	78.88	64.43	25.85	84.01	79.53	80.18	78.80	70.89	34.79	66.90	73.62	77.45	77.91	67.40	28.49	38.22	72.61	76.53
	DBA	31.74	74.59	80.47	10.37	73.21	97.11	73.96	99.94	69.29	99.87	70.81	92.43	72.51	8.22	75.56	99.51	71.89	97.81	58.31	79.46
	MR	62.08	87.44	80.76	31.49	73.43	96.36	74.07	96.77	74.28	97.31	75.25	96.58	61.72	92.67	72.79	97.05	74.80	90.82	65.01	87.49
	PreFed	73.19	78.43	80.69	84.90	80.64	87.95	73.84	91.00	80.42	68.30	79.31	70.92	80.56	82.06	75.11	94.15	80.84	93.20	79.99	17.37
	PREFed	74.10	93.78	81.31	95.80	81.10	97.25	80.55	97.98	80.53	99.12	80.61	98.75	74.86	94.88	80.83	96.60	80.21	98.22	74.29	96.14
Foolsgold	3DFed	74.76	45.78	76.31	45.75	77.08	38.97	71.85	36.99	75.43	24.82	74.86	12.41	72.27	17.01	76.41	80.65	74.24	42.66	75.93	48.34
	DBA	59.74	99.80	73.23	96.60	70.81	87.56	70.49	96.82	62.91	99.19	69.12	91.04	70.40	97.94	74.88	95.09	74.70	86.98	74.43	97.08
	MR	72.33	61.31	74.74	59.37	71.28	98.65	73.11	92.03	71.35	63.49	75.60	63.82	68.99	46.78	71.93	51.02	76.92	57.39	72.28	79.72
	PreFed	78.00	81.95	75.35	87.29	76.48	79.91	78.01	70.41	71.71	73.26	71.69	36.42	75.47	72.34	72.36	43.55	74.81	72.98	75.76	49.66
	PREFed	75.42	96.11	76.25	96.97	76.30	93.97	77.16	92.63	74.87	93.14	76.53	90.75	73.62	92.62	77.08	92.04	76.12	95.08	77.12	91.91
FreqFed	3DFed	75.31	13.55	75.34	27.28	74.10	62.57	74.27	9.78	75.05	10.13	72.03	13.58	73.38	39.38	73.88	34.72	75.29	15.26	74.94	14.67
	DBA	64.83	99.90	72.26	93.79	48.88	99.99	73.67	88.00	73.48	93.52	69.88	98.18	72.67	69.20	65.41	99.78	73.20	86.30	69.75	80.81
	MR	76.15	10.59	73.04	88.36	74.11	60.47	60.17	98.47	74.31	9.50	61.37	97.10	72.54	30.75	75.17	69.27	73.35	79.03	74.14	39.68
	PreFed	75.99	82.82	75.38	58.26	75.13	53.89	74.86	78.27	74.75	63.44	74.58	76.60	74.09	71.46	74.70	59.20	75.44	91.92	75.38	77.67
	PREFed	76.18	97.60	75.88	95.70	75.17	69.10	74.43	96.86	75.49	87.51	73.03	92.84	74.18	97.44	74.32	89.79	75.84	87.75	75.21	94.42
RFLBAT	3DFed	78.34	10.89	78.04	56.13	79.98	9.80	79.80	12.61	79.32	28.22	79.34	21.35	79.44	9.52	79.03	9.82	78.39	36.60	79.10	65.84
	DBA	78.74	10.62	79.48	9.97	78.10	9.91	80.25	10.28	80.28	10.48	79.10	45.54	79.74	10.10	79.54	10.02	79.19	10.23	79.46	10.32
	MR	78.98	10.03	78.74	10.11	79.82	9.93	79.28	10.83	80.45	10.58	79.80	10.13	79.33	9.57	78.98	9.84	77.27	10.24	78.46	10.22
	PreFed	78.74	87.22	78.53	93.12	79.10	91.20	80.14	20.12	79.95	79.52	79.68	20.52	79.73	17.87	78.19	94.34	78.97	75.67	79.22	89.39
	PREFed	78.70	96.77	79.12	97.40	80.06	96.81	79.66	16.65	80.43	13.63	79.97	95.00	79.08	95.34	79.24	36.61	79.31	95.05	78.92	58.18
FedAvg	3DFed	79.16	46.47	80.45	81.10	80.08	65.23	80.52	51.69	79.25	69.54	76.17	41.50	80.63	55.88	78.77	59.05	78.98	73.55	79.26	66.97
	DBA	72.22	98.62	72.42	95.51	70.45	96.69	69.26	29.86	68.29	98.02	67.03	99.20	72.49	98.91	73.36	99.68	73.23	90.15	69.49	84.90
	MR	71.57	87.03	71.73	92.67	75.71	78.29	74.26	96.76	72.86	99.33	69.80	95.04	66.12	37.49	69.02	98.02	74.19	91.52	72.38	99.82
	PreFed	80.31	88.66	81.17	90.95	80.93	92.06	80.63	79.07	80.08	80.87	81.05	81.04	79.65	80.88	80.10	86.31	80.89	90.97	80.75	85.96
	PREFed	80.14	99.03	80.44	99.00	80.64	95.75	80.68	97.02	80.83	93.95	80.21	97.40	80.75	97.13	79.43	98.47	80.91	97.03	80.46	99.41