Breaking the Illusion: A Critical Study of Backdoor Defense in Federated Learning With Non-IID Data

Pei Ye, Yuqing Li, *Member, IEEE*, Kun He, *Member, IEEE*, Qiao Li, Tianjie Qin, Xiong Wang, *Member, IEEE*, Kaige Yang, Chujun Zhang, and Jing Chen, *Senior Member, IEEE*

***Abstract*—Existing backdoor defense methods for federated learning (FL) usually try to distinguish between benign and malicious clients. The key insight is that benign clients are densely distributed, whereas malicious clients tend to be outliers outside this distribution. However, this only holds when data is independent and identically distributed (IID), and the effectiveness of these methods under non-IID data has not been systematically examined. In this paper, we present a comprehensive systematization of FL backdoor defense by breaking down its overall pipeline into three key components, i.e., metrics for evaluating clients, techniques for amplifying the difference between benign and malicious clients, and mechanisms for identifying malicious clients. We conduct an empirical study of FL backdoor defense methods under non-IID data settings to explore whether benign and malicious clients can be fully distinguished. Experimental results show that the defense performance degrades significantly when data is non-IID. Our results also reveal how evaluation metrics, amplification techniques and identification mechanisms perform under diverse settings. Contrary to the established belief, we further conclude that these defenses have inherent shortcomings, due to lack of stability and robustness in detecting malicious clients. We believe that our findings can better facilitate the development of FL backdoor defenses.**

***Index Terms*—Federated learning, backdoor attack and defense, non-IID data.**

I. Introduction

FEDERATED learning (FL) enables multiple clients, such as smartphones and IoT devices, to collaboratively train a machine learning model without sharing their raw data [1], [2], [3], [4]. Specifically, each client submits the model update trained on its local data to a central server, which aggregates these updates to refine the global model. Since the server cannot directly access the computations and data on the client side, an attacker can manipulate some clients to stealthily inject a backdoor into the global model by submitting malicious updates. This backdoor forces the model to exhibit the attacker-chosen behaviors on specific test samples, while behaving normally on others. Researchers showed that backdoor attacks can cause serious damages to real-world FL applications, such as autonomous driving [5] and healthcare [6].

A. Fl Defenses Against Backdoor Attacks

To mitigate backdoor attacks on FL, current defense methods are often performed on the server side, which can be broadly classified into two categories, i.e., post-aggregation defense [7], [8], [9], [10] and pre-aggregation detection [11], [12], [13], [14]. Post-aggregation defenses focus on repairing the backdoor model after federated model aggregation is completed, leveraging techniques such as fine-pruning and knowledge distillation, similar to those used for centralized learning [15], [16]. However, they cannot promptly detect attacks and remove backdoors, potentially leading to wasted computational resources. Moreover, identifying the attacker (i.e., manipulated clients) for accountability remains a challenge. In contrast, pre-aggregation defenses apply an anomaly detector to identify and eliminate backdoor updates from suspicious clients for aggregation. Such defenses enable real-time detection of malicious updates during training, thereby preventing attackers from injecting backdoors into the global model. In this work, we focus on **pre-aggregation detections** because, in addition to the aforementioned advantages, we find that their effectiveness is highly dependent on the local data distribution of clients, which aligns with our research objective of reevaluating the performance of FL backdoor defenses under diverse data distribution scenarios.

B. Unique Features of Pre-Aggregation Detections for Fl

In FL, client updates are essentially their renewed local models. In this way, the most straightforward method for detecting malicious updates is to conduct backdoor detection on each local model individually. Although many backdoor model detection methods have been established for centralized learning, they often rely on computationally intensive trigger

Received 8 April 2025; revised 4 September 2025; accepted 30 November 2025. Date of publication 11 December 2025; date of current version 18 December 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3102100, in part by the National Natural Science Foundation of China under Grant 62302343 and Grant 62441237, in part by National Natural Science Foundation of China-Hong Kong Research Grants Council (NSFC-RGC) under Grant 62461160333, in part by Wuhan Scientific and Technical Achievements Project under Grant 2024030803010172, in part by the Key Research and Development Program of Shandong Province under Grant 2022CXPT055, and in part by Wuhan Natural Science Foundation Exploratory Program (Chenguang Program) under Grant 2024040801020210. The associate editor coordinating the review of this article and approving it for publication was Prof. Edgar Weippl. *(Corresponding author: Yuqing Li.)*

Pei Ye, Yuqing Li, Kun He, Qiao Li, Tianjie Qin, Kaige Yang, Chujun Zhang, and Jing Chen are with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: yepeiii@whu.edu.cn; li.yuqing@whu.edu.cn; hekun@whu.edu.cn; liqiaoqiao233@whu.edu.cn; rinpostk@whu.edu.cn; yangkg@whu.edu.cn; 2024282210318@whu.edu.cn; chenjing@whu.edu.cn).

Xiong Wang is with the National Engineering Research Center for Big Data Technology and System, Services Computing Technology and System Laboratory/Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xiongwang@hust.edu.cn).

Digital Object Identifier 10.1109/TIFS.2025.3643155

reverse engineering [17], [18], making them impractical in the FL context, where detection is performed before each round of federated aggregation across massive clients. Fortunately, anomaly detection offers a promising alternative to identify malicious client in FL, by leveraging the assumption that there definitely exist benign clients which can serve as reference for comparison. Building on this premise, current pre-aggregation detection methods in FL typically start by identifying clients with anomalous updates as potentially malicious clients. The key insight is that due to the injection of a backdoor into the local update, malicious clients tend to significantly deviate from benign ones, thus becoming detectable outliers.

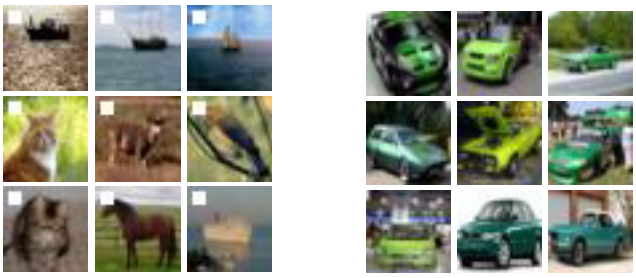
C. The Challenge Faced by Pre-Aggregation Detection-Based Fl Defenses

The fundamental insight behind pre-aggregation detections only holds when the data is independent and identically distributed (IID) across clients. However, as highlighted in the literature [19], [20], [21], FL often suffers from the challenge of data heterogeneity, that is, the local data across clients is non-IID. Accordingly, there may be larger variability among benign clients, further complicating the differentiation between benign and malicious clients. As a result, the effectiveness of pre-aggregation detections will be significantly reduced.

D. Our Contributions

To our knowledge, several studies [22], [23], [24] mention that pre-aggregation detections are not suitable for non-IID scenarios, but none have thoroughly examined the reasons behind this. In this work, we conduct a comprehensive study of the pre-aggregation detections under non-IID scenarios. Our ultimate goal is to reveal the inherent limitations of such defenses and inspire research toward more robust solutions to backdoor attacks in FL. Concretely, we make the following key contributions.

- We establish six non-IID scenarios to simulate three types of data heterogeneity, including label distribution skew, feature distribution skew, and quantity skew. With this setup, we extensively evaluate the performance of defense methods under various data heterogeneity conditions and explore the impact of different non-IID types.
- We summarize the FL backdoor defense paradigm for pre-aggregation detection methods, which encompasses the metrics for evaluating clients, techniques for amplifying the differences between benign and malicious clients, and mechanisms for identifying malicious clients. By decoupling the defense pipeline, this paradigm facilitates exploring the defense performance of evaluation metrics, amplification techniques, and identification mechanisms separately under various non-IID data settings.
- Based on extensive experimental results, we analyze the effectiveness of pre-aggregation detections under non-IID data settings from multiple perspectives, including the impact of learning task complexity, the distinctions among various types of data heterogeneity, and the robustness of defense methods when facing different poisoning hyperparameters. Considering more realistic federated learning scenarios, we reach a conclusion that contradicts the common beliefs in the existing literature: **pre-aggregation detections are not capable of resisting backdoor attacks in FL**. We summarize the inherent limitations of these defenses and offer some recommendations for practitioners.



(a) Artificial backdoor attacks (b) Semantic backdoor attacks

Fig. 1. Illustration of the trigger patterns for artificial backdoor attacks and semantic backdoor attacks.

II. Background

A. Federated Learning

Consider a typical FL system involving a central server and a set of M clients that jointly train a machine learning model while keeping their training data local. Each client i holds a local dataset D_i with $s_i = |D_i|$ data samples. In each training round t, the server randomly selects m clients and broadcasts the current global model G^t to them. Each selected client i performs local training using D_i and sends the model update W_i^t to the server. The server then aggregates these updates to build a new global model $G^{t+1} = G^t + \frac{1}{m}\sum_{i=1}^{m} W_i^t$.

B. Fl Backdoor Attacks

Despite its advantages, FL is vulnerable to backdoor attacks [25], [26], where an attacker can manipulate several clients to submit malicious updates, causing the global model to be poisoned. We denote the proportion of compromised clients among all participants as poisoned model rate (PMR). The poisoned model would perform well on normal inputs but behave maliciously on specific attacker-chosen inputs with certain trigger patterns. Specifically, the attacker manipulates the local data of compromised clients to insert a backdoor. We denote the fraction of poisoned data among the overall local data as poisoned data rate (PDR). Based on trigger pattern characteristics, FL backdoor attacks are broadly categorized into artificial backdoor attacks and semantic backdoor attacks, as illustrated in Fig. 1.

1) Artificial Backdoor Attacks: In artificial backdoor attacks, the trigger (e.g., square pattern) is attached to the original input and the target samples can belong to any class. To make the assault more convert, DBA [27] exploits the distributed nature of FL by decomposing the target trigger into multiple local triggers, with each compromised client injecting only one partial trigger into its local data. Beyong fixed triggers, some advanced attacks further optimize the trigger during training using regularization terms (e.g., CerP

[28]) or adversarial learning (e.g., A3FL [29]) to enhance the attack effectiveness and stealthiness.

The focus of our work is to investigate the defense performance under non-IID data settings, an intrinsic problem of FL [19], [20], [21], making it hard to deploy in practice. Due to space limitations, we only examine the most representative DBA attack. This does not affect our main conclusion: defense methods fail under non-IID data conditions even against the simplest attack. In fact, if a defense method cannot resist DBA, it will be even less effective against more advanced attacks.

2) Semantic Backdoor Attacks: Different from artificial backdoor attacks, semantic backdoor attacks [30] treat specific characteristics within a certain class of samples as triggers (e.g., green car and stripped pattern in the scene). The attacker flips their labels without modifying the features (e.g., labeling green car images as "frog"). To prevent the benign clients' updates from diluting the backdoor effect, ECBA [25] proposes to poison those out-of-distribution samples. Since these samples lie at the tail of the data distribution of benign clients, the impact of the backdoor is maintained.

In ECBA, malicious clients possess some out-of-distribution samples with labels flipped by the attacker. We observe that when each benign client only holds samples from a single class (i.e., an extreme non-IID scenario), there is no clear distinction between the behavior of malicious and benign clients, as both assign a specific class to their local samples. To investigate whether malicious and benign clients become indistinguishable in such highly heterogeneous settings, we conduct a case study based on ECBA in Section IV-F.

C. Simulating Non-IID Data

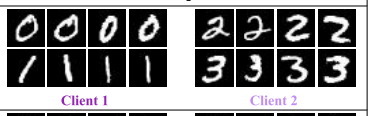
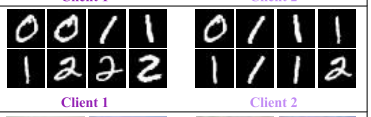
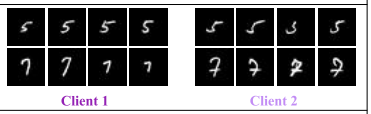
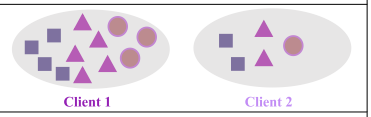
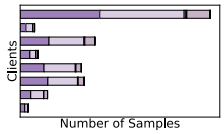
In FL, non-IID data settings can be categorized from a distribution perspective [31]. Let x_i and y_i denote the features and labels of local data D_i, respectively. Considering the local data distribution $P(x_i, y_i) = P(x_i|y_i)P(y_i) = P(y_i|x_i)P(x_i)$, data heterogeneity includes the following five types: (1) label distribution skew, i.e., $P(y_i)$ varies across clients; (2) feature distribution skew, i.e., $P(x_i)$ differs across clients; (3) identical labels but different features, i.e., $P(x_i|y_i)$ varies across clients; (4) identical features but different labels, i.e., $P(y_i|x_i)$ differs across clients; and (5) quantity skew, i.e., $P(x_i, y_i)$ is the same but data volume is different among clients. Among the above types, the third is typically observed in vertical FL [32], while the fourth is uncommon in most FL studies. Hence, we primarily focus on the other three types in this work. Below, we introduce them and present the corresponding simulation methods used in our experiments, as shown in Table I.

1) Label Distribution Skew: Label distribution skew refers to variations in label distributions among clients' local data. For example, the specialized hospitals that focus on treating specific diseases tend to collect data with a higher proportion of samples related to those diseases.

a) Quantity-based label imbalance: In this scenario, each client holds samples from a fixed set of labels. Specifically, each client is assigned k distinct labels, denoted as $\#C = k$, where C represents the set of unique labels held by a client. Labels are first uniformly distributed across all clients, and the data corresponding to each label is evenly shared among clients possessing that label.

TABLE I

CLASSES OF NON-IID DATA IN FL AND THE CORRESPONDING SIMULATION METHODS USED IN OUR EXPERIMENTS

Category	Simulation methods	Description
Label distribution skew	*Quantity-based label imbalance*	
	Distribution-based label imbalance	
Feature distribution skew	*Noise-based feature imbalance*	
	Real-world feature imbalance	
Quantity skew	*Distribution-based quantity imbalance*	
Mixed types of skew	*Real-world data imbalance*	

b) Distribution-based label imbalance: In this setup, each client is allocated a portion of samples from each label based on the Dirichlet distribution [33], a continuous probability distribution defined over multidimensional probability vectors. Specifically, for a class c in the dataset, a probability vector $[p_{c1}, p_{c2}, \cdots, p_{cM}]$ is generated using the parameter α, where M is client number and $\sum_{i=1}^{M} p_{ci} = 1$. Each client then samples data for class c according to the assigned probabilities. The parameter α controls the level of label skew and smaller α values result in more imbalanced data distributions. For ease of presentation, we denote this simulation method as Dir(α).

2) Feature Distribution Skew: Feature distribution skew refers to the situation where clients share the same conditional distribution $P(y_i|x_i)$, yet the feature distributions of their local data $P(x_i)$ differ significantly. For example, clients collecting handwriting data from different individuals encounter variations in writing styles.

a) Noise-based feature imbalance: This situation simulates the skewed feature distributions by introducing varying degrees of noise to the original data. Specifically, the data is first evenly distributed among clients. For client i, the Gaussian noise with a mean of 0 and a standard deviation of $\sigma \cdot i/M$ is added to its local data, where σ controls the level of feature dissimilarity among clients. We denote this method as Gau(σ).

b) Real-world feature imbalance: The FEMNIST dataset [34] is a collection of handwritten digits from various writers. In particular, the writers (and their corresponding digits) are randomly and equally assigned to each client. As each writer

has unique character features, the feature distributions are different among clients.

3) Quantity Skew: Quantity skew occurs when clients share the same distribution $P(x_i, y_i)$ but differ in the amount of samples they possess. For example, large banks typically have far more customers than small banks, resulting in a larger volume of training data.

a) Distribution-based quantity imbalance: This setup simulates the variations in sample quantities among clients using the Dirichlet distribution. Specifically, a probability vector $[p_1, p_2, \cdots, p_M]$ is generated based on the parameter α. This vector determines the number of samples allocated to each client. The samples are then randomly assigned to clients according to these proportions. We denote this simulation method as qDir(α).

4) Mixed Types of Skew: Mixed types of skew refers to a combination of multiple types of skew, making the situation more complex. For example, when training a prediction model using purchase information on users' mobile devices, each user's purchase activity level and pReferences vary, leading to differences in both data quantity and feature distribution.

a) Real-world data imbalance: The LOAN dataset [35] comprises data related to loan predictions, with each sample containing information about a loan application and the label indicating the application's outcome. In LOAN, the data is distributed among clients based on the states where loan applicants reside, with each client representing one state, resulting in a total of 51 clients. The data distribution across clients is visualized in Table I, where only eight clients are shown as examples and each color represents a different label. It is evident that this dataset exhibits quantity skew. Additionally, as each client's data comes from different states, it also demonstrates feature distribution skew.

III. Systematization of FL Backdoor Defenses

A. Overview

Existing server-side backdoor defenses for FL can be broadly categorized into two main types: post-aggregation defense and pre-aggregation detection. The first type is designed to directly repair the global model after aggregating all local updates, regardless of whether they contain malicious ones. It can effectively eliminate the impact of backdoor, by applying techniques such as perturbation [7], smoothing [8], knowledge distillation [9], and fine-pruning [10] to the aggregated model. In contrast, the second type employs anomaly detection to identify malicious clients and exclude them from aggregation, under the premise that malicious clients inevitably exhibit substantial deviations from benign ones due to the backdoor injected into their local updates [11], [12], [13], [14]. It is particularly attractive to FL, as it allows for real-time detection of malicious updates during training, thereby preventing attackers from injecting backdoors into the global model. Moreover, it also enables accountability for malicious clients when necessary.

Despite these benefits, the performance of pre-aggregation detection can be affected by the local data distributions of clients. This is because there may be larger variability among benign clients under non-IID data situations, further complicating the differentiation between benign and malicious clients. Our work focuses on systematically investigating the effectiveness of the pre-aggregation detection methods in distinguishing malicious clients from benign ones during training under non-IID data settings. Therefore, post-aggregation defenses that do not differentiate between benign and malicious clients are beyond the scope of this work. In the rest of this paper, "FL backdoor defense" refers specifically to the pre-aggregation detection, unless explicitly stated otherwise.

B. The Pipeline of Fl Backdoor Defense

We break down the FL backdoor defense pipeline into three steps: (1) select a metric to evaluate clients' local updates, (2) amplify the differences between benign and malicious clients based on this metric, and (3) identify malicious clients among all participants and exclude them from global model aggregation. Below, we introduce the efforts made by nine state-of-the-art (SOTA) FL backdoor defenses in distinguishing between benign and malicious clients at different steps, aiming to maintain defense effectiveness even under non-IID settings. Specifically, at each step of the defense pipeline, we categorize the employed methods into several prototypes. A summary of these prototypes is presented in Table II.

1) Metrics for Evaluating Clients: Based on the characteristic that malicious clients embed backdoors into their local models, evaluation metrics are crafted to capture the unique traits of malicious updates as much as possible.

a) All parameters of local update: Using all parameters of the local update is the simplest and most straightforward method since it contains all information in the local update, including the backdoor in malicious update.

b) Local update of the last FC layer: Since the labels of the attacker-chosen samples are flipped in backdoor attacks, the impact of the backdoor on the last fully connected (FC) layer is more pronounced than on other layers, which motivates using local update of the Last FC Layer for evaluation.

c) DCT-transformed local update: Backdoor attacks associate specific patterns with particular outputs, shifting the energy distribution of updates towards certain frequencies. As a result, using DCT transformation [40], malicious updates become more distinguishable from benign ones in frequency domain.

d) Model-update consistency: For a benign client i employing gradient descent for local model training, its updates W_i^{t-1} and W_i^t in two consecutive rounds satisfy $W_i^t = W_i^{t-1} + H_i^t \cdot (G^t - G^{t-1})$ according to Cauchy's mean value theorem. Here, H_i^t is an integrated Hessian matrix for client i, and G^t and G^{t-1} are global models in rounds $t-1$ and t, respectively. This equation exhibits the consistency between two consecutive updates from a benign client, while the malicious updates, manipulated by the attacker, fail to achieve this.

e) Model activations of benign samples: Although the backdoor model can correctly predict the labels of benign samples, the activations of its hidden layers differ from those of a benign model. When benign samples are unavailable, the server can distribute the uploaded local models to the

TABLE II
THE PROTOTYPES OF EACH STEP IN THE PIPELINE OF NINE SOTA FL BACKDOOR DEFENSES

Method	Evaluation Metric	Amplification Technique	Identification Mechanism
FLAME [26]	All parameters of local update	Peer comparison$_1$	Unsupervised learning-based categorization$_1$
Multi-metrics [36]	All parameters of local update	Peer comparison$_2$	Predefined parameter-based selection$_1$
FLTrust [37]	All parameters of local update	Server benign reference	Predefined parameter-based selection$_2$
FL-Defender [38]	Local update of the last FC layer	Peer comparison$_1$, Median-based reference	Predefined parameter-based selection$_1$
FoolsGold [39]	Local update of the last FC layer	Within-group consistency-based measurement	Predefined parameter-based selection$_1$
FreqFed [13]	DCT-transformed local update	Peer comparison$_1$	Unsupervised learning-based categorization$_1$
FLDetector [11]	Model-update consistency	History-based enhancement	Unsupervised learning-based categorization$_2$
CrowdGuard [12]	Model activations of benign samples$_1$	N/A	Unsupervised learning-based categorization$_3$
FLShield [14]	Model activations of benign samples$_2$	N/A	Predefined parameter-based selection$_1$

[i] "Model activations of benign samples$_{1\text{-}2}$" use the activations from all layers and from only the final layer, respectively.
[ii] "Peer comparison$_{1\text{-}2}$" represent forming a vector of pairwise distances and summing the pairwise distances, respectively.
[iii] "Predefined parameter-based selection$_{1\text{-}2}$" denote the top-K selection and threshold comparison, respectively.
[iv] "Unsupervised learning-based categorization$_{1\text{–}3}$" refer to HDBSCAN, K-means, and agglomerative clustering, respectively.

validation clients (referred to as "validator" hereinafter) and construct evaluation metrics according to their feedback on local samples. This prototype is further categorized into two variants, one utilizing activations from all layers and the other relying solely on the final layer, which are denoted as subscript 1 and subscript 2, respectively, in Table II.

2) Techniques for Amplifying Difference: Amplification techniques can further enhance the distinction of evaluation metrics between benign and malicious clients, thereby aiding in the subsequent identification of malicious clients.

a) Peer comparison: This technique measures a client using the pairwise distance (e.g., cosine distance, Euclidean distance, and Manhattan distance) between evaluation metrics of this client and all other clients. Since the evaluation metrics of malicious clients deviate from those of benign ones, it enhances the distinction between malicious and benign clients. This prototype can be instantiated in two forms, one forming a feature vector from the pairwise distances and the other aggregating the pairwise distances by summation, which are marked with subscripts 1 and 2 in Table II.

b) Server benign reference: Assuming that the server holds a portion of clean data, it can use this clean data to train a benign update as a trusted reference in each round. The distance between the evaluation metrics of each client and this benign one is then calculated to amplify the differences between benign and malicious clients.

c) Median-based reference: If more than 50% of the clients are presumed benign, the median value of all clients' evaluation metrics can serve as a trusted reference. Calculating the distance between each client's evaluation metric and the median further distinguishes malicious clients from the benign ones.

d) Within-group consistency-based measurement: This technique considers that the evaluation metrics of malicious clients exhibit similarity because they have a shared attack objective while benign ones show diversity due to the stochastic nature of stochastic gradient descent (SGD). Therefore, when measuring a client by its minimum distance to all other clients, malicious clients often exhibit smaller values.

e) History-based enhancement: Since malicious clients upload malicious updates over multiple rounds, incorporating historical information within a recent time window can enhance the client's evaluation metrics.

3) Mechanisms for Identifying Malicious Clients: After measuring each client in the previous two steps and obtaining their amplified metrics, the identification mechanism determines which clients are malicious.

a) Predefined parameter-based selection: Based on empirically determined parameters, this type of identification mechanism directly selects malicious clients, such as through top-K selection or threshold comparison, which can be denoted with subscripts 1 and 2, respectively, in Table II.

b) Unsupervised learning-based categorization: As unsupervised learning can learn relative relationships of data, this type of identification mechanism deploys clustering to categorize malicious clients, such as HDBSCAN [41], K-means and agglomerative clustering [42] with a fixed number of clusters, which dynamically determines the number of clusters. These three mechanisms are sequentially marked with subscripts 1, 2 and 3 in Table II.

IV. ANALYSIS OF DEFENSE EFFECTIVENESS UNDER NON-IID SETTINGS

In this section, we analyze the effectiveness of FL backdoor defense methods under non-IID data settings. Specifically, we start by evaluating the performance of SOTA methods in Section IV-B, and then explore the impact of learning task complexity, non-IID type and poisoning hyperparameters on defense effectiveness under non-IID settings (Sections IV-C to IV-E). Finally, we present a case study in Section IV-F based on ECBA under the $\#C = 1$ setting, and further investigate the existence of defense boundaries in Section IV-G. For reproducibility of experimental results, we make our code available at https://github.com/Gloriatry/Back_to_non-IID_data

A. Experimental Setup

We implement all experiments in Python using PyTorch 1.10 on a server with Intel Xeon Gold 6133 CPU, 251GB RAM, 24 GB NVIDIA GeForce RTX 4090 GPU, and Ubuntu 20.04.

TABLE III
DATASETS USED IN OUR EXPERIMENTS

Datasets	#Records	#Features	#Classes	Default Model	#Params
EMNIST	280K	784	10	CNN	431K
CIFAR-10	60K	1024	10	ResNet-18	2.7M
CIFAR-100	60K	1024	100	ResNet-18	2.7M
CINIC-10	270K	1024	10	ResNet-18	2.7M
FEMNIST	380K	784	10	CNN	431K
LOAN	2.26M	91	9	MLP	5.53K

1) Datasets and Models: Following recent studies on FL backdoor attacks and defenses [14], [26], [30], [36], we conduct systematic experiments on multiple widely used datasets, including five image datasets (i.e., EMNIST [43], CIFAR-10 [44], CIFAR-100 [44], CINIC-10 [45], and FEMNIST [34]) and one tabular dataset (i.e., LOAN [35]), to validate the robustness and generalizability of our results. For the employed models, we use a CNN [46] for EMNIST and FEMNIST, a ResNet-18 [47] for CIFAR-10, CIFAR-100, and CINIC-10, and an MLP [48] for LOAN. The statistics of these datasets and models are summarized in Table III.

2) Assessment Metrics: We use the following metrics to assess the effectiveness of FL backdoor defenses. *Backdoor accuracy (BA)* measures the proportion of trigger samples for which the model outputs the attacker-chosen label, reflecting model performance on the backdoor task. Attackers aim to maximize BA, whereas a lower BA indicates a more effective defense. *Main task accuracy (MA)* refers to the proportion of test samples for which the model predicts the correct label, measuring model accuracy on the main task. Attackers need to maintain high MA for stealthiness, while the defense techniques should avoid adversely affecting MA. *True positive rate (TPR)* represents the ratio of correctly identified malicious clients among all participating malicious clients (i.e., $TPR = \frac{TP}{TP+FN}$), while *true negative rate (TNR)* measures the ratio of correctly identified benign clients among all participating benign clients (i.e., $TNR = \frac{TN}{TN+FP}$). We present both TPR and TNR using the average value across all rounds. And higher TPR and TNR reflect more accurate detections. All metric values are reported in percentages.

3) Default Configurations: By default, we assume 100 clients, with 25 selected per training round. The SGD optimizer is used with a learning rate of 0.01 and a momentum of 0.9. The batch size and local epochs are set to 64 and 3, respectively. Attackers use a PDR of 0.2 and a PMR of 0.3.

To simulate various non-IID settings in our experiments, we meticulously choose a set of representative parameters. Specifically, to contrast with the IID setting, we set $\#C = 2$ and $\#C = 3$, which not only exhibit a high degree of heterogeneity but also better reflect realistic conditions. For distribution-based label imbalance, we adopt Dir(0.1) and Dir(0.5), following numerous studies on data heterogeneity in FL [49], [50], [51], where Dir(0.1) corresponds to a highly heterogeneous case in practice, while Dir(0.5) represents a more moderate level. For noise-based feature imbalance, we follow [52] in choosing the Gau(0.1) setting, where the injected noise effectively differentiates feature distributions across clients. For quantity skew, we also follow [52] and use the qDir(0.5) setting, wherein the variation in local sample sizes among clients is sufficiently large to induce heterogeneity.

B. Evaluating SOTA Defenses

In this section, we summarize the defense performance of each SOTA method under various non-IID data settings and analyze the effectiveness of the three steps in its defense pipeline using visualization results on CIFAR-10.

Table IV presents the defensive results of different defense methods against DBA on different datasets under various non-IID data settings, with each defense's effectiveness assessed compared with FedAvg. We consider a defense successful if it reduces BA to below 20% while decreasing MA by no more than 3%. The successful defense cases are marked out.

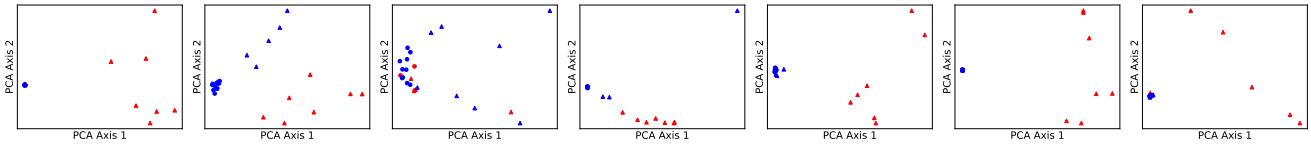
In FL, non-IID data often degrades MA [52] and BA may fluctuate throughout the training process [30]. Moreover, some defense methods considered in this paper not only detect and remove malicious updates but also subsequently apply post-aggregation defense to the global model, which further reduces BA, albeit with potential impacts on MA. For example, FLAME adds noise to the aggregated model in each round. To highlight the defense methods' ability to distinguish between benign and malicious clients, we present the TPR and TNR metrics on CIFAR-10 in Table V, with successful defense cases also marked out. There is an obvious correlation between TPR and BA. In general, a TPR above 95% ensures successful defense; otherwise, many malicious updates infiltrate the global model, resulting in a high BA. In conclusion, *BA and MA capture the overall performance of FL backdoor defense methods, while TPR and TNR evaluate their efficacy in detecting malicious clients.*

1) Flame Evaluation: From Tables IV and V, FLAME successfully defends in most cases, but under the Dir(0.1) setting, the post-defense BA still reaches 91.87%. To visualize the distribution of clients, we use principal component analysis (PCA) [53] to reduce the dimension of amplified metric in FLAME (i.e., the vector composed of pairwise distances to clients) for each client to 2. The results on CIFAR-10 are presented in Fig. 2. Under the IID, $\#C = 2$, and Gau(0.1) settings, the amplified metrics effectively distinguish between benign and malicious clients, as benign clients (blue points) are clustered together while malicious clients (red points) are farther apart. However, under other settings, the differences among benign updates become larger, leading to lower consistency in the amplified metrics of benign clients (e.g., Dir(0.5) and $\#C = 3$) or even bringing them closer to those of malicious clients (e.g., Dir(0.1) and qDir(0.5)). The HDBSCAN clustering algorithm employed in FLAME, which dynamically determines the cluster number, identifies all clients that are difficult to group into the largest cluster as malicious. This avoids mistakenly grouping malicious clients into the benign cluster just because there exist more different benign ones. Consequently, despite ambiguous boundaries in the amplified metric distributions of malicious and benign clients, HDBSCAN can successfully identify malicious clients under the Dir(0.5), $\#C = 3$, and qDir(0.5) settings with

TABLE IV

THE BA/MA OF SOTA DEFENSES AGAINST DBA UNDER VARIOUS NON-IID SETTINGS, WITH SUCCESSFUL DEFENSE CASES MARKED OUT

Category	Dataset	Setting	FedAvg	FLAME	Multi-metrics	FLTrust	FL-Defender	FoolsGold	FreqFed	FLDetector	CrowdGuard	FLShield
Homogeneous distribution	EMNIST	N/A	100/99.6	**0.05**/99.56	100/99.57	**0.07**/99.55	**0.07**/99.57	**0.06**/99.54	**0.06**/99.57	**27.66**/99.56	**0.07**/99.55	**0.06**/99.54
	CIFAR-10		97.71/69.28	**2**/69.46	96.04/70.42	98.4/71.07	**2.07**/70.68	**2.05**/69.34	**2.09**/69.48	**48.71**/70.11	**2.15**/69.6	**2.14**/69.41
	CIFAR-100		97.94/67.61	**1.03**/67.22	97.3/67.91	95.7/68.12	**2.35**/68.14	**4.42**/67.77	**1.1**/67.12	**82.04**/68.42	**1.07**/67.94	**1.1**/67.83
	CINIC-10		98.96/81.49	**4.13**/78.46	99.9/81.21	98.56/80.39	99.51/81.07	**3.89**/80.8	**4.12**/80.6	99.94/80.72	99.79/81.22	99.41/80.89
Label distribution skew	EMNIST	Dir(0.5)	100/99.53	**0.06**/99.52	99.78/99.51	**0.41**/99.5	100/99.53	100/99.53	**0.05**/99.52	**7.61**/99.53	**0.05**/99.51	**0.06**/99.53
		Dir(0.1)	100/99.4	**0.08**/99.09	**0.05**/99.35	**0.19**/99.18	100/99.39	100/99.39	**0.08**/99.11	**5.65**/99.36	100/99.34	**0.11**/99.25
		$\#C=3$	100/99.52	**0.05**/99.44	**0.09**/99.46	**0.07**/99.36	100/99.47	100/99.51	**0.05**/99.47	**13.45**/99.48	**0.06**/99.44	**0.05**/99.47
		$\#C=2$	100/99.39	**0.36**/98.33	**0.07**/99.43	**0.07**/99.14	100/99.36	100/99.4	**0.33**/98.74	87.24/99.43	**0.56**/99.35	**0.29**/99.23
	CIFAR-10	Dir(0.5)	94.81/68.02	**9.15**/66.22	92.85/69.43	94.38/69.01	96.1/67.83	94.76/69.06	**5.59**/69.24	**12.2**/70.02	92.92/69.21	**5.63**/69.2
		Dir(0.1)	95.86/56.42	91.87/58.66	94.9/55.76	95.44/52.49	94.14/58.24	92.52/59.82	93.45/ 57.12	96.3/60.54	95.2/56.96	93.59/59.37
		$\#C=3$	97.63/58.41	**7.4**/64.4	92.37/59.79	96.72/62.06	97.15/60.29	96.09/63.09	**6.64**/64.23	**14.58**/61.63	95.38/67.69	**6.12**/67.78
		$\#C=2$	99.08/28.73	**6.42**/31.94	96.86/28.7	97.74/23.9	96.5/24.69	98.31/26.62	**12.37**/30.36	**15.77**/30.7	98.62/35.77	99.66/31.99
	CIFAR-100	Dir(0.5)	97.69/65.35	**1.62**/65.67	97.68/65.13	96.34/66.38	90.61/65.97	97.17/65.21	**0.89**/65.37	99.29/65.88	88/65.45	**0.32**/66.23
		Dir(0.1)	98.05/61.94	93.09/61.08	97.74/62.34	97.45/62.12	94.63/61.83	98.78/61.49	97.44/62.08	99.39/62.03	88.4/61.85	91.34/61.92
		$\#C=30$	97.18/65.11	**0.72**/65.21	97.68/65.38	97.19/65.39	94.09/65.12	96.53/64.92	**0.65**/64.91	**57.13**/64.88	82.37/65.03	**0.78**/65.27
		$\#C=20$	97.83/63.74	**1.03**/63.26	97.25/63.87	97.53/64.92	83.51/62.77	99.26/63.58	**0.65**/63.41	99.31/63.76	81.35/63.28	92.57/64.15
	CINIC-10	Dir(0.5)	99.56/76.26	99.76/75.29	99.96/76.13	99.74/76.65	99.92/75.03	99.86/74.98	97.55/74.75	99.91/77.54	99.92/76.97	98.88/76.32
		Dir(0.1)	99.68/63.75	99.96/62.29	99.82/61.52	99.88/67.04	99.92/60.95	99.65/59.83	98.73/63.27	99.96/68.72	99.09/64.15	99.57/64.01
		$\#C=3$	99.93/49.46	100/43.92	100/48.24	99.98/53.27	99.99/47.22	99.85/46.19	99.98/44.42	99.99/55.48	99.99/47.61	99.92/49.45
		$\#C=2$	99.98/42.54	100/29.4	100/38.29	99.99/40.09	99.98/36.65	99.73/31.39	99.89/28.9	100/38.94	99.83/37.54	99.69/34.94
Feature distribution skew	EMNIST	Gau(0.1)	100/99.57	**0.06**/99.57	100/99.58	**0.06**/99.55	**0.09**/99.57	**0.07**/99.54	**0.06**/99.57	**28.46**/99.55	**0.07**/99.55	**0.07**/99.57
	CIFAR-10		94.83/70.8	**3.52**/70.88	89.68/72.01	97.52/71.66	80.78/72.08	**3.08**/71.55	**3.32**/70.77	**50.38**/71.42	**3.33**/70.79	**3.55**/70.95
	CIFAR-100		97.96/66.44	**0.75**/65.48	95.69/67.49	96.33/66.28	81.99/65.71	**0.85**/67.02	**0.73**/66.59	**78.89**/66.18	89.06/66.79	**0.27**/67.23
	CINIC-10		99.23/81.47	**4.13**/78.46	99.9/81.21	98.56/0.39	99.51/81.07	**3.89**/80.8	**4.29**/80.66	99.94/80.41	99.51/81.23	99.56/80.78
	FEMNIST	N/A	100/99.45	**0.04**/99.4	100/99.45	**0.04**/99.33	**0.04**/99.4	**0.08**/99.4	**0.04**/99.4	100/99.4	**0.03**/99.36	**0.03**/99.37
Quantity skew	EMNIST	qDir(0.5)	100/99.59	**0.06**/99.59	**0.09**/99.57	**0.07**/99.55	100/99.57	99.91/99.6	**0.07**/99.54	**7.68**/99.58	**0.06**/99.56	**0.06**/99.57
	CIFAR-10		95.07/74.56	**6.85**/73.96	92.37/74.51	92.11/74.51	72.16/74.73	**3.74**/74.51	34.03/73.31	**69.27**/74.85	90.15/75.17	**3.91**/74.25
	CIFAR-100		95.14/68.11	75.42/68.21	95.07/69.21	95.25/68.17	87.67/68.37	31.31/68.19	80.03/69.91	94.54/68.1	95.03/67.88	**0.65**/68.59
	CINIC-10		99.21/81.18	98.86/77.08	99.71/81.06	98.95/80.2	97.38/80.97	**3.49**/80.86	99.21/79.95	99.73/81.3	97.04/81.14	98.03/80.38
Mixed types	LOAN	N/A	99.96/97.36	99.96/97.29	99.96/97.28	99.83/97.32	**0.16**/97.34	**0**/97.19	99.97/97.31	99.95/97.31	99.93/97.07	99.97/97.05



(a) IID (b) Dir(0.5) (c) Dir(0.1) (d) #C=3 (e) #C=2 (f) Gau(0.1) (g) qDir(0.5)

Fig. 2. Visualization of each participating client's amplified metric on CIFAR-10 under various non-IID settings in FLAME, where red and blue represent malicious and benign clients, triangles and circles represent clients identified as malicious and benign by FLAME, respectively.

TABLE V

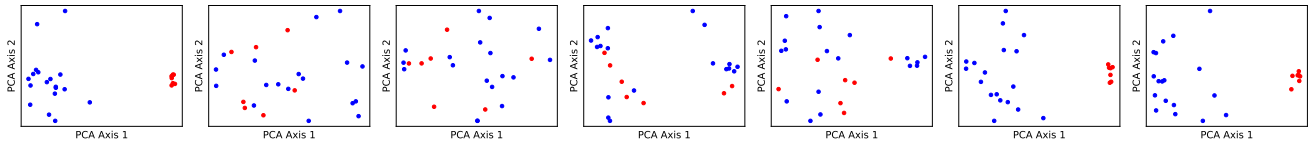
TPR AND TNR OF SOTA DEFENSES ON CIFAR-10 UNDER VARIOUS NON-IID SETTINGS, WITH SUCCESSFUL DEFENSE CASES MARKED OUT

Defense	Metric	IID	Dir(0.5)	Dir(0.1)	$\#C$=3	$\#C$=2	Gau(0.1)	qDir(0.5)
FLAME	TPR	**100**	**98.8**	49.81	**100**	**100**	**100**	**97.01**
	TNR	83.5	73.29	54.87	77.99	80.52	83.42	75.77
Multi-metrics	TPR	63.45	60.39	49.81	64.15	66.69	66.97	61.38
	TNR	85.79	84.58	80.48	86.03	87.04	87.17	84.98
FLTrust	TPR	44.57	46.59	46.09	47.61	52.6	47.67	58.56
	TNR	53.61	61.08	59.38	62.04	53.28	55.49	56.49
FL-Defender	TPR	**100**	35.44	29.93	27.8	41.57	95.99	90.6
	TNR	100	74.89	72.75	71.92	77.28	98.44	96.34
FoolsGold	TPR	**100**	64.06	67.03	43.32	37.62	**99.97**	**99.95**
	TNR	84.88	58.08	28.82	57.71	41	98.66	87.36
FreqFed	TPR	**100**	**100**	77.21	**100**	**100**	**100**	92.7
	TNR	82.6	75.71	64.9	79.18	78.81	84.21	75.59
FLDetector	TPR	**100**	**99.99**	27.78	**100**	**99.98**	**100**	**97.2**
	TNR	100	99.42	76.64	99.57	99.64	100	99.9
CrowdGuard	TPR	**100**	36.91	0	12.79	7.46	**100**	9.2
	TNR	100	93.43	93.2	93.01	93.82	99.72	98.02
FLShield	TPR	**100**	**100**	80.73	**100**	36.87	**100**	**99.96**
	TNR	72.22	72.22	60.77	72.22	47.67	72.22	72.21

high TPR. However, it still fails when the degree of data heterogeneity is too high (e.g., Dir(0.1)).

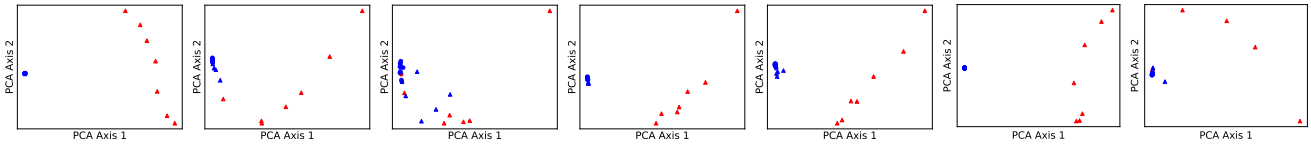
2) Multi-Metrics Evaluation: We observe that Multi-metrics performs poorly overall and exhibits instability under non-IID settings, with the post-defense BA generally exceeding 90% and a low TPR. To reduce dimensionality, Multi-metrics sums across the client dimension after computing pairwise distances between local updates. This method may lead to the loss of important information as it does not account for the distinctions between a client and other clients separately. Moreover, the top-K selection method used in Multi-metrics is overly simplistic, requiring the evaluation metrics and amplification techniques to possess strong discriminative ability. How to determine the optimal value for K also poses a big challenge.

3) FLTrust Evaluation: FLTrust performs poorly overall under non-IID settings, with the post-defense BA still above 92% and the TPR typically ranging between 45% and 55%. For amplification, FLTrust uses the server's trusted benign update as reference. However, the effectiveness of this technique heavily depends on the distribution of the clean dataset on the server. If there is a significant difference between the server's clean data and a benign client's local data, the distance will be large, leading to misclassification of benign updates. Furthermore, setting a distance threshold in FLTrust is not robust because it is difficult to determine



(a) IID (b) Dir(0.5) (c) Dir(0.1) (d) #C=3 (e) #C=2 (f) Gau(0.1) (g) qDir(0.5)

Fig. 3. Visualization of each participating client's feature vector on CIFAR-10 under various non-IID settings in FL-Defender, where red and blue represent malicious and benign clients, respectively.



(a) IID (b) Dir(0.5) (c) Dir(0.1) (d) #C=3 (e) #C=2 (f) Gau(0.1) (g) qDir(0.5)

Fig. 4. Visualization of each participating client's feature vector on CIFAR-10 under various non-IID settings in FreqFed, where red and blue represent malicious and benign clients, and triangles and circles represent clients identified as malicious and benign by FreqFed, respectively.

an appropriate threshold that performs well across diverse scenarios.

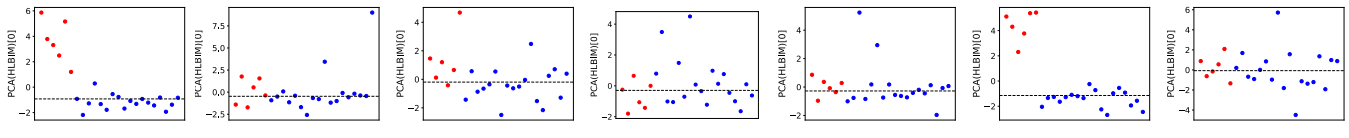
4) Fl-Defender Evaluation: FL-Defender performs poorly overall under label distribution skew settings, with the post-defense BA above 95%. Under the Gau(0.1) and qDir(0.5) settings, though BA decreases by 14%-23% compared to before defense, the defense remains unsuccessful. For each client, FL-Defender calculates the distances between its last FC layer's update and those of other clients to form a feature vector. We perform PCA on the feature vectors and visualize the results on CIFAR-10 in Fig. 3. Under the IID, Gau(0.1), and qDir(0.5) settings, the distribution of benign and malicious clients has a clear distinction; Yet, under the label distribution skew case (e.g., $\#C = k$ and Dir(α)), it becomes harder to distinguish them. This is because the last FC layer's parameters are influenced not only by backdoor attack but also by label distribution of training data. As a result, malicious clients become dissimilar due to varied label distributions of local data, and the differences among benign clients increase.

After obtaining the feature vectors, FL-Defender chooses the coordinate-wise median vector as a trustworthy benign reference to further amplify the differences between benign and malicious clients. Yet, this technique is not robust because the distribution of benign and malicious clients on a certain dimension of feature vectors may not be regular, making the median value not necessarily originating from a benign client. For example, in Fig. 3(f), the median value on Axis 1 clearly originates from a benign client, whereas the median value on Axis 2 may originate from a malicious client. Due to incorrect reference selection, FL-Defender still fails to defend effectively, despite the fact that malicious and benign updates are already well distinguished (cf., Fig. 3(f) and Fig. 3(g)).

5) FoolsGold Evaluation: FoolsGold performs well under the Gau(0.1) and qDir(0.5) settings, but fails in the case of label distribution skew with post-defense BA exceeding 92%. To amplify the differences, FoolsGold calculates, for each client, the minimum distance to all other clients, considering that malicious clients exhibit high consistency in their last FC layer parameters. However, this does not apply under the label distribution skew settings, because malicious updates in the last FC layer become diverse due to the differing label distributions of local data on malicious clients.

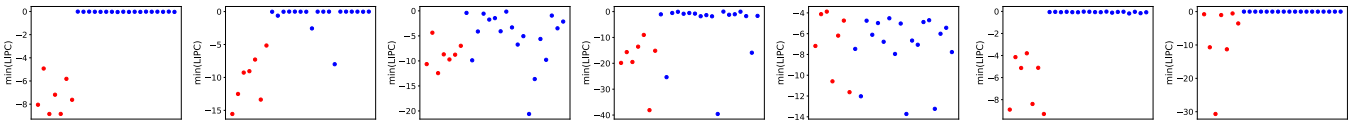
6) FreqFed Evaluation: FreqFed successfully defends in most cases, but under the Dir(0.1) setting, the post-defense BA still reaches 93.45%. We conduct PCA on the amplified metric in FreqFed (i.e., the vector formed by the pairwise distances of DCT-transformed local updates) for each client and visualize the results on CIFAR-10 in Fig. 4. Under all settings except Dir(0.1) and qDir(0.5), the distinction between benign and malicious clients is evident. However, under the Dir(0.1) and qDir(0.5) settings, the amplified metrics of benign and malicious clients become indistinguishable. This suggests that label distribution skew and quantity skew can induce significant differences in model updates in the frequency domain, which are even greater than those caused by backdoor.

7) FLDetector Evaluation: FLDetector successfully defends under most non-IID cases but fails under the Dir(0.1) setting, with a post-defense BA of 96.3%. It is worth noting that since FLDetector relies on historical data from N rounds to make detection decisions, there is no defense applied in the initial rounds, causing a rapid rise in BA. Once the defense is activated, BA gradually declines, but it does not drop below 20% quickly. Therefore, we use TPR instead of BA to assess the defense's success in FLDetector. Using model-update consistency as the evaluation metric, FLDetector employs L-BFGS algorithm [54] to estimate a single Hessian matrix $\hat{H}^t$ for all clients based on the differences in the global model over several past rounds. It then predicts the update W_i^t as $\hat{W}_i^t = W_i^{t-1} + \hat{H}^t \cdot (G^t - G^{t-1})$, and computes the difference between the predicted update $\hat{W}_i^t$ and the actual update W_i^t. However, under non-IID settings, the estimated Hessian matrix computed based on the global model is no longer accurate since the inconsistency in local model update directions leads to a decline in the global model's generalization ability [52]. Therefore, the difference between the predicted update and actual update for benign clients increases, which further complicates the distinction between benign and malicious clients.



(a) IID (b) Dir(0.5) (c) Dir(0.1) (d) #C=3 (e) #C=2 (f) Gau(0.1) (g) qDir(0.5)

Fig. 5. Visualization of each participating client's HLBIM on CIFAR-10 under various non-IID settings in CrowdGuard, where red and blue represent malicious and benign clients, respectively.



(a) IID (b) Dir(0.5) (c) Dir(0.1) (d) #C=3 (e) #C=2 (f) Gau(0.1) (g) qDir(0.5)

Fig. 6. Visualization of each participating client's LIPC metric on CIFAR-10 under various non-IID settings in FLShield, where red and blue represent malicious and benign clients, respectively.

8) CrowdGuard Evaluation: CrowdGuard successfully defends only under the Gau(0.1) setting but fails in the case of label distribution skew and quantity skew, with a post-defense BA exceeding 90%. CrowdGuard defines a hidden layer backdoor inspection metric (HLBIM), which measures the distance between the activation matrix of each local model and that of the global model. We visualize the distribution of the first PCA dimension of HLBIM matrix for participating clients and show the results in Fig. 5. Under the IID and Gau(0.1) settings, malicious clients are clearly distinguished from benign clients, suggesting that the hidden layers' outputs of backdoor and benign models differ greatly, even on benign samples. However, under the label distribution skew settings, we find that malicious clients are distributed closely to benign clients. When the label distributions of the validator's local data and the data used to train the benign models differ, a benign model tends to predict samples as the majority label in its training set, which closely resembles a malicious model prone to predicting samples as the attacker-chosen label. Therefore, it becomes difficult to distinguish between them. Moreover, under the quantity skew setting, we observe significant variability in the HLBIM of benign clients. This indicates that the quantity of samples has a considerable impact on the training of benign models; with fewer samples, benign models may lack adequate generalization, yielding outputs that diverge from those of other benign models.

9) FLShield Evaluation: FLShield performs well under most non-IID cases but fails to defend when label distribution skew is severe (e.g., Dir(0.1) and #C = 2), with a post-defense BA above 93%. By leveraging activation differences between malicious and benign models on benign samples, FLShield applies the loss impact per class (LIPC) metric to record the average loss difference between the global model from the previous round and the local model. Since backdoor models tend to produce less accurate logits and higher losses, LIPC values for malicious clients are lower. Figure 6 illustrates LIPC value for each participating client. Under the IID and Gau(0.1) settings, malicious clients can be easily distinguished from benign clients. Despite correctly predicting benign samples, the backdoor model incurs a higher loss than that of the benign model. However, under label distribution skew settings, the benign model only fully learns the knowledge of a subset of labels, leading to incorrect logit predictions for samples from other classes and increased loss. When data heterogeneity degree is low, this may not have a significant impact (cf., Fig. 6(b) and Fig. 6(d)). As the heterogeneity degree increases, LIPC values of benign clients decrease, approaching those of malicious clients (cf., Fig. 6(c) and Fig. 6(e)).

TABLE VI

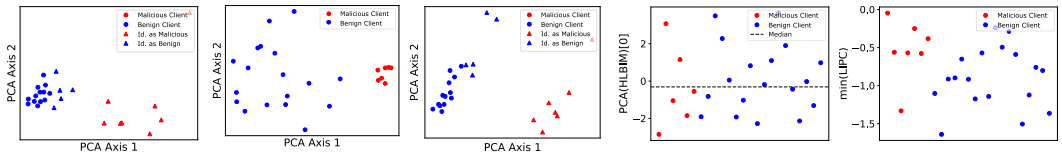
THE BA/MA OF SOTA DEFENSES WITH DIFFERENT MODELS APPLIED ON CIFAR-10 UNDER VARIOUS NON-IID SETTINGS

Model	IID	Dir(0.5)	Dir(0.1)	#C=3	#C=2	Gau(0.1)	qDir(0.5)
				FLAME			
ResNet-18	**2**/69.46	**9.15**/66.22	91.87/58.66	**7.4**/64.4	**6.42**/31.94	**3.52**/70.88	**6.85**/73.96
ResNet-34	**4.25**/68.26	**11.54**/67.15	83.87/57.01	**11.33**/44.23	**0.68**/31.18	**3.56**/69.45	**12.2**/74.85
ResNet-50	**3.58**/69.89	66.97/68.91	86.95/54.33	55.12/49.06	92.46/24.75	**3.77**/70.47	**16.69**/74.95
				FreqFed			
ResNet-18	**2.09**/69.48	**5.59**/69.24	93.45/57.12	**6.64**/64.23	**12.37**/30.36	**3.32**/70.77	34.03/73.31
ResNet-34	**4.34**/68.17	**6.7**/69.91	67.52/55.69	**11.55**/45.37	**5.03**/27.05	**3.55**/69.22	52./74.76
ResNet-50	**3.73**/70.07	**4.96**/72.09	94.07/55.42	**5.97**/52.28	32.34/27.62	**3.74**/70.75	43.33/75.57
				FLShield			
ResNet-18	**2.14**/69.41	**5.63**/69.2	93.59/59.37	**6.12**/67.78	99.66/31.99	**3.55**/70.95	**3.91**/74.25
ResNet-34	**4.38**/68.9	**5.77**/69.56	93.29/59.25	96.77/47.92	99.08/30.38	**3.67**/69.39	**4.76**/75.32
ResNet-50	**3.64**/70.28	**5.13**/72.16	95.44/58.25	96.86/53.92	98.7/32.56	**3.8**/70.47	**4.41**/75.83

C. Investigating Learning Task Complexity Impact

We now study the impact of learning task complexity on defense effectiveness in non-IID settings, using experimental results from CIFAR-10, CIFAR-100, and CINIC-10.

The complexity of the learning task arises from both the dataset and the model. The dataset factor can be further divided into the aspects of samples and labels. From Table IV, the defenses perform slightly worse on CIFAR-100 than on CIFAR-10. Given that CIFAR-100 has the same number of samples as CIFAR-10 but a larger number of labels, it suggests that an increased number of labels makes the learning task more difficult. Furthermore, Table IV shows that on CINIC-10, the defenses fail in almost all cases. Although CINIC-10 has the same number of labels as CIFAR-10, it contains 4.5 times more samples with more diverse features. Consequently, a larger and more diverse set of samples increases the complexity of the learning task.



(a) FLAME (b) FL-Defender (c) FreqFed (d) CrowdGuard (e) FLShield

Fig. 7. Visualization results of five FL backdoor defense methods on CINIC-10 under the IID setting.

We then proceed to experimentally examine the impact of the model. Specifically, we extend the employed model to ResNet-34 and ResNet-50, with results presented in Table VI. When model architectures become more complex as well as the number of model parameters increases, defense failures occur more frequently. This underscores the significance of the model as a critical factor influencing learning task complexity.

We next analyze why the learning task complexity influences the defense effectiveness of pre-aggregation detection. Based on the premises on which the detection depends, the nine SOTA defenses can be divided into two categories. The first category includes FLAME, Multi-metrics, FLTrust, FL-Defender, FoolsGold, FreqFed, and CrowdGuard, which rely on the premise that there is consistency among either benign clients or malicious clients, and a significant difference between benign and malicious clients. As the learning task grows in complexity, the convergence directions of locally trained models become inconsistent. Hence, the distinction between benign and malicious clients becomes obscure, making it challenging to identify malicious clients under non-IID settings. Such phenomenon can be observed by comparing Fig. 2(a) and Fig. 7(a), Fig. 3(a) and Fig. 7(b), Fig. 4(a) and Fig. 7(c), and Fig. 5(a) and Fig. 7(d).

The second category, such as FLDetector and FLShield, relies on the assumption that benign clients outperform malicious ones in certain aspect. Specifically, FLDetector assumes that the difference between the predicted and actual updates of benign clients is smaller than that of malicious clients, while FLShield considers that benign models can predict more accurate logits on benign samples than malicious models. The complexity of learning task prevents FLDetector and FLShield from making accurate predictions, causing the performance of benign clients become as poor as malicious ones. This phenomenon can be verified from a comparison between Fig. 6(a) and Fig. 7(e).

In conclusion, the learning task affects defense effectiveness by undermining the premises underlying the defense methods.

Finding on Learning Task Complexity. *The more complex the learning task, the more challenging it becomes to implement effective defenses under non-IID data settings.*

D. Exploring Different Non-IID Data Types

We then explore how various non-IID types impact defense effectiveness, using experimental results from the six datasets.

According to the results on EMNIST, CIFAR-10, CIFAR-100, and CINIC-10 under various non-IID settings presented in Table IV, we can observe that these FL backdoor defense methods perform poorly under label distribution skew settings. When benign clients' local data have differing label distributions, they learn from samples of different classes, increasing the differences among them. Similarly, the differences among malicious clients also grow. This makes it difficult to distinguish between benign and malicious clients. In FLDetector, under label distribution skew, the global model fails to learn knowledge from all benign clients' local data. Consequently, using the Hessian matrix computed from the global model to predict updates for all benign clients is inaccurate, leading to a significant gap between the predicted and actual updates of benign clients. In FLShield, the logits of benign models tend to bias towards the majority class in their local training sets, resulting in high prediction loss for benign models on benign samples with different label distributions.

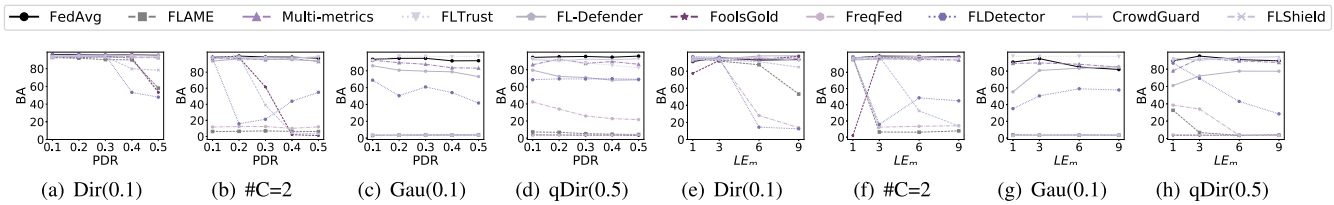
Moreover, as shown in Table IV and V, some defense methods can successfully defend against attacks under the Dir(0.5) and $\#C = 3$ settings, but fail under the Dir(0.1) and $\#C = 2$ settings. When successful defense is achieved under all four label distribution skew settings, the TPR for Dir(0.5) is higher than that for Dir(0.1), and the TPR for $\#C = 3$ is also higher than that for $\#C = 2$. This is because as data heterogeneity degree increases, differences in label distribution exert a greater influence, thus narrowing the distinction between benign and malicious clients, and making it easier for defense methods to make misjudgments.

Finding on Label Distribution Skew. *Label distribution skew greatly degrades the effectiveness of defense methods. The higher the degree of data heterogeneity, the poorer the defense performance.*

From the results on EMNIST, CIFAR-10, CIFAR-100, and CINIC-10 in Table IV, the defense performance under Gau(0.1) setting is nearly identical to that under IID setting. This is because, under the Gau(0.1) setting, each client's local data is simply augmented with Gaussian noise, which does not affect the model learning knowledge of images and thus has limited impact on both benign and malicious updates. We also observe that most defense methods successfully defend on FEMNIST, despite feature distribution skew present in FEMNIST. This is likely due to the relatively simple learning task inherent in FEMNIST.

Finding on Feature Distribution Skew. *Both noise-based feature distribution skew and feature distribution skew in FEMNIST have limited impact on defense performance.*

By comparing the defense results on EMNIST, CIFAR-10, CIFAR-100, and CINIC-10 under the IID and qDir(0.5) settings in Table IV, we find that quantity skew undermines the defense effectiveness of FLAME, FreqFed, and CrowdGuard. Specifically, through comparing Fig. 2(a) and Fig. 2(g),



(a) Dir(0.1) (b) #C=2 (c) Gau(0.1) (d) qDir(0.5) (e) Dir(0.1) (f) #C=2 (g) Gau(0.1) (h) qDir(0.5)

Fig. 8. Impact of training hyperparameters on CIFAR-10.

Fig. 4(a) and Fig. 4(g), and Fig. 5(a) and Fig. 5(g), respectively, we find that quantity skew affects defense performance either by increasing the differences among benign clients (i.e., CrowdGuard) or by making malicious clients close to benign clients (i.e., FLAME and FreqFed). In the first case, a small volume of local data on benign clients results in benign model overfitting, making their outputs inconsistent. While in the second case, a small amount of local data on malicious clients causes malicious updates (or malicious updates in the frequency domain) to be less influenced by the backdoor, thereby narrowing the gap between benign and malicious updates. Local data volume can affect the knowledge learned by benign or malicious models, depending on the evaluation metric used by defense methods. This makes the distinction between benign and malicious clients less obvious and diminishes the defense efficacy.

Finding on Quantity Skew. *When local data volume is too small and affects the knowledge learned by either benign or malicious models, quantity skew leads to a decrease in the overall performance of the defense.*

Table IV shows that only FL-Defender and FoolsGold successfully defend on LOAN. This may be attributed to the similar label distribution across clients in LOAN, as evident from data distribution of LOAN in Table I. Consequently, the last FC layer of benign and malicious models is no longer affected by label distribution, and only the last FC layer of the malicious model is influenced by the backdoor. However, all other defense methods fail on LOAN, suggesting that mixed types of skew is a more challenging heterogeneous setting.

Finding on Mixed Types of Skew. *The defense is more challenging when there exist mixed types of skew.*

E. Setting Various Poisoning Hyperparameters

We then examine how the poisoned data rate and local epochs of malicious clients affect defense effectiveness under non-IID data by setting different poisoning hyperparameters.

1) Impact of Poisoned Data Rate: We vary the PDR over {0.1, 0.2, 0.3, 0.4, 0.5} and report the BA on CIFAR-10 under various non-IID settings in Fig. 8(a)–8(d). As PDR increases, BA exhibits an overall downward trend. To be specific, defense methods, i.e., FLAME, FoolsGold, FLDetector, and FLShield show improved performance. For instance, under the $\#C = 2$ setting, FoolsGold completely fails to defend against the attack when PDR is 0.2, but it successfully mitigates the attack when PDR rises to 0.4 or 0.5. This observation suggests that a higher PDR allows malicious updates to contain more backdoor knowledge and makes the distinction between malicious and benign clients more evident, thereby facilitating the detection of malicious clients and improving defense effectiveness.

2) Impact of Local Epochs: We vary local epochs of malicious clients over {1, 3, 6, 9} and report the BA on CIFAR-10 in Fig. 8(e)–8(h). Hereinafter, we denote malicious clients' local epochs as LE_m. Results show that as LE_m increases, BA generally declines. Similar to the trends observed with varying PDR, several defense methods, including FLAME, FreqFed, FLDetector, and FLShield, show a particularly pronounced decline. For example, under the $\#C = 2$ setting, FLAME fails to defend when LE_m is 1, but successfully mitigates the attack when LE_m increases to 3, 6 or 9. This is because, as LE_m grows, local models of malicious clients diverge from the global model, embedding more backdoor characteristics in their updates, which in turn makes the distinction between malicious and benign updates more evident.

Beyond the defense methods previously discussed, other approaches including Multi-metrics, FLTrust, FL-Defender, and CrowdGuard consistently yield high BA even when increasing poisoning hyperparameters. This demonstrates that in local updates, the impact of non-IID data significantly outweighs that of backdoor attacks.

Finding on Training Hyperparameters. *Only when the attacker increases poisoning hyperparameters to improve the attack success rate can the defense possibly succeed, which implies that the defense effectiveness under non-IID settings is not robust.*

F. Case Study

We next conduct a case study to further investigate the impact of non-IID data distribution on defense effectiveness.

As mentioned in Section II-B2, when an attacker executes ECBA and each benign client holds samples from only a single class (i.e., $\#C = 1$), the behaviors of benign and malicious clients become fundamentally indistinguishable. They assign a specific label to their local samples, and each client's local samples follow distinct distributions. In other words, any benign client cannot determine if the labels assigned to local data by other clients are accurate, regardless of whether those clients are benign or malicious. In this case study, we perform experiments on CIFAR-10 using the default parameter configurations. The BA and TPR of FedAvg as well as various defense methods are shown in Fig. 9, while the visualization results of some defense methods are presented in Fig. 10.

It can be observed that only FoolsGold successfully defends against the attack, while other defense methods fail. FoolsGold performs well because, in ECBA, malicious clients share

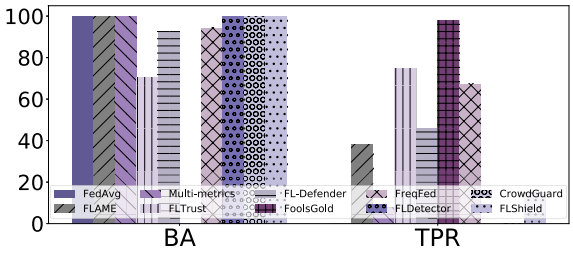


Fig. 9. The BA and TPR of various defense methods against ECBA on CIFAR-10 under #C = 1 setting.

samples from the same distribution, producing highly similar malicious updates. In contrast, benign clients' updates differ significantly due to heterogeneous local data. FoolsGold leverages this distinction to identify malicious clients.

Figures 10(a)-10(d) shows that the significant differences among benign clients blur the boundary between them and malicious clients. This indicates that the amplified metrics used by FLAME, FL-Defender, FreqFed, and CrowdGuard cannot effectively distinguish between benign and malicious clients. While in FLDetector and FLShield, the evaluation metrics of benign and malicious clients exhibit a phenomenon completely contrary to expectations. We observe that malicious clients demonstrate higher model-update consistency than benign clients, and the LIPC value for malicious clients is larger than that for benign clients as illustrated in Fig. 10(e).This occurs because under the #C = 1 setting, the global model struggles to learn from heterogeneous benign clients and instead learns from consistent malicious clients. Thus, the estimated Hessian matrix can accurately predict the updates for malicious clients in FLDetector, and malicious models incur very low losses on malicious clients' local samples in FLShield.

Overall, in the extreme scenario where ECBA is performed with #C = 1, the substantial differences among benign clients make it difficult to distinguish them from malicious clients. With regard to those defense methods that operate under the assumption that benign clients consistently outperform malicious ones in certain aspects, there may even emerge a contrary phenomenon in which malicious clients actually appear to perform better or look superior. The only defense that demonstrates success in this setting, namely FoolsGold, specifically targets to exploit the consistency of malicious clients. However, this advantage is far from robust, as it represents a weakness that an adaptive attacker can easily circumvent, for example, by intentionally adding random noise to malicious updates [55] or by incorporating carefully designed regularization strategies during local training [28], thereby rendering the defense ineffective.

Finding on Case Study. *In some extreme non-IID scenarios, the behavior of benign clients assigning labels to locally distinct data becomes essentially indistinguishable from the label-flipping behavior of malicious clients.*

G. Exploring the Existence of Defense Boundaries

As analyzed in Section IV-F, under extreme data heterogeneity settings (e.g., each client having entirely distinct sample classes), it becomes impossible to determine whether a client's sample assignment is correct. Consequently, pre-aggregation detection methods cannot effectively distinguish malicious clients from benign but heterogeneous ones. In other words, any defense is bound to fail when the heterogeneity degree becomes sufficiently large.

We next empirically investigate whether there exists a clear defense boundary beyond which the defense begins to fail as the heterogeneity degree increases. Specifically, we construct progressively more severe non-IID scenarios. For quantity-based label imbalance, the number of classes per client is varied from #C = 1 to #C = 10, while for distribution-based label imbalance, the Dirichlet concentration parameter is adjusted from Dir(0.1) to Dir(1.0). Table VII shows the BA and MA of three top-performing defenses on CIFAR-10 and CINIC-10, suggesting that such a defense boundary consistently exists across all scenarios. For instance, FLAME becomes ineffective once the heterogeneity level goes beyond Dir(0.4) and #C = 1 on CIFAR-10. Moreover, for FLShield, the boundary is infinite on CINIC-10, as it already fails to defend under IID settings (see Table IV).

Finding on Defense Boundary. *There necessarily exists a defense boundary beyond which the defense starts to fail when the heterogeneity degree exceeds it.*

V. Discussion

A. Limitations of Pre-Aggregation Defenses

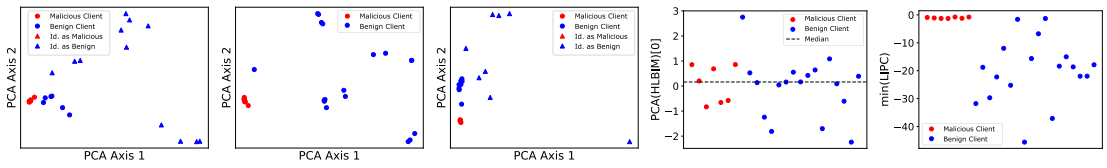
The design principle of pre-aggregation detection is that the impact of a backdoor on the local model is so significant that it can be distinguished from benign ones. In the context of FL, however, this impact is often less pronounced than expected for several reasons.

First, our extensive experiments on nine state-of-the-art (SOTA) federated learning (FL) backdoor defense methods reveal that the impact of non-IID data sometimes outweighs that of the backdoor itself. Despite efforts to refine their strategies across the three steps of the defense pipeline to distinguish between benign and malicious clients, they still fail in many non-IID data settings. This issue becomes particularly pronounced when learning task is more complex or when the attacker employs smaller poisoning hyperparameters.

Second, in FL, the learning process of clients on their local data is inherently incremental, with each uploaded update to the server reflecting only a few epochs of local training [56], [57]. As a result, the information contained in any single update is quite limited, implying that a single round of malicious contribution is unlikely to convey a sufficiently strong backdoor signal.

Third, it remains challenging to ensure that all malicious updates are successfully identified and filtered out in every round. Even a single undetected poisoned update, once incorporated into the global model, can introduce and propagate a backdoor to all participating clients in the subsequent round. This contamination undermines the foundation on which defenses rely, thereby rendering subsequent detection more difficult.

Therefore, we hold a pessimistic view towards using pre-aggregation defenses for mitigating backdoor attacks in FL.



(a) FLAME (b) FL-Defender (c) FreqFed (d) Crowdguard (e) FLShield

Fig. 10. Visualization results of five defense methods against ECBA on CIFAR-10.

TABLE VII

THE BA/MA OF SOTA DEFENSES ON CIFAR-10 AND CINIC-10 UNDER GRADUALLY INCREASING DEGREES OF NON-IID SETTINGS, WITH SUCCESSFUL DEFENSE CASES MARKED OUT

Dataset	Defense	Setting									
		Dir(0.1)	Dir(0.2)	Dir(0.3)	Dir(0.4)	Dir(0.5)	Dir(0.6)	Dir(0.7)	Dir(0.8)	Dir(0.9)	Dir(1.0)
CIFAR-10	FLAME	91.87/58.66	75.15/60.23	42.08/64.59	44.03/65.68	**9.15**/66.22	**14.14**/69.78	**7.42**/68	**10.58**/69.6	**5.1**/69.13	**5.86**/68.58
	FreqFed	93.45/57.12	23.88/61.78	**5.16**/66.89	**5.14**/67.86	**5.59**/69.24	**5.98**/69.97	**5.07**/69.88	**7.77**/69.32	**4.4**/69.47	**6.1**/69.47
	FLShield	93.59/59.37	**5.81**/63.08	**5.03**/67.08	**4.9**/67.85	**5.63**/69.2	**6.06**/69.99	**4.79**/70.06	**6.28**/69.54	**4.31**/69.68	**6.07**/69.6
CINIC-10	FLAME	99.96/62.29	99.89/68.79	99.92/72.33	99.85/73.99	99.76/75.29	99.68/77.47	99.22/76.91	**7.83**/78.06	**6.12**/77.98	**6.63**/78.2
	FreqFed	98.73/63.27	99.9/68.5	99.89/72.77	99.77/74.61	97.55/74.75	99.87/77.22	99.82/76.77	**3.14**/77.98	**4.25**/76.56	**5.06**/78.23
	FLShield	99.57/64.01	98.82/70.05	98.7/73.53	97.58/74.45	98.88/76.32	96.38/77.19	95.8177.25	98.89/78.21	96.29/78.14	97.57/78.5
		$\#C$=1	$\#C$=2	$\#C$=3	$\#C$=4	$\#C$=5	$\#C$=6	$\#C$=7	$\#C$=8	$\#C$=9	$\#C$=10
CIFAR-10	FLAME	100/9.26	**6.42**/31.94	**7.4**/64.4	**5.12**/69.18	**6.74**/67.33	**5.78**/69.2	**5.47**/69.71	**6.51**/67.57	**6.49**/67.49	**6.26**/67.92
	FreqFed	100/10.01	**12.37**/30.36	**6.64**/64.23	**5.33**/69.23	**5.96**/69.67	**5.39**/69.17	**5.37**/69.8	**6.3**/67.58	**6.49**/67.6	**6.03**/68.07
	FLShield	99.71/9.85	99.66/31.99	**6.12**/67.78	**5.88**/69.39	**12.4**/69.95	**6.08**/69.17	**5.91**/69.75	**6.82**/67.57	**6.89**/67.62	**6.15**/68.07
CINIC-10	FLAME	100/10.38	100/29.4	100/43.92	99.99/57.44	99.98/68.6	99.95/73.61	**10.16**/76.53	**2.93**/77.31	**3.98**/77.59	**3.82**/78.62
	FreqFed	98.66/9.08	99.89/28.9	99.98/44.42	99.99/58.94	**2.63**/72.88	**4.23**/77.18	**3.4**/78.2	**3.18**/79.59	**4.84**/79.59	**3.83**/80.58
	FLShield	97.59/10.82	99.69/34.94	99.92/49.45	99.87/58.05	99.49/72.59	99.23/76.44	99.44/78.25	99.17/79.79	99.17/79.84	98.74/80.78

Even if more robust evaluation metrics, amplification techniques, or identification mechanisms are designed in the future, they can only marginally improve defense effectiveness without fully resolving the issue of malicious and benign clients becoming indistinguishable under non-IID data settings.

B. Takeaway Suggestions for Practitioners

1) Towards Advanced Pre-Aggregation Defenses: Foregoing analysis reveal that no pre-aggregation detection can consistently succeed across all scenarios. Nevertheless, FLAME, FreqFed, and FLShield exhibit the relatively better defense performance due to their implementation of superior strategies at specific steps of the defense pipeline. FLShield adopts model activations of benign samples as the evaluation metric and uses activations from only the final layer. This suggests that the outputs of the model on benign samples can effectively distinguish backdoored models from benign ones. Under non-IID settings, relying solely on activations of the final layer outperforms using all layers, as shallow layers primarily capture generic features where backdoored and benign models appear more similar. For amplification technique, FLAME and FreqFed both use peer comparison, by building a feature vector for each local update that consists of its pairwise distances to all other updates. This is an effective way to amplify the distinction between malicious and benign updates, as it performs pairwise comparisons between every two updates. Furthermore, FLAME and FreqFed employ unsupervised learning-based categorization as their identification mechanism. Unlike predefined parameter-based selection, this proves more effective and robust, as it can learn relative relationships among amplified metrics and adapt to diverse scenarios. Moreover, the adopted HDBSCAN outperforms other clustering algorithms since it dynamically determines the number of clusters.

Based on the above analysis, we advocate the practical use of FLAME, FreqFed, and FLShield, particularly in real-world scenarios with relatively mild levels of data heterogeneity.

2) Resorting to Personalized Fl Framework: In traditional FL frameworks, all clients share a unified model via global aggregation. For security purposes, malicious updates must be carefully detected and thoroughly removed before or after aggregation, which becomes particularly challenging when client data is Non-IID. However, recent studies [58], [59], [60] have demonstrated that in personalized FL, each client maintains a model customized to its local data distribution, thereby inherently preventing malicious updates from infiltrating the local models of benign clients. Therefore, practitioners can consider adopting the personalized FL framework to enhance model security. Specifically, two approaches can be implemented. The first is partial model sharing, where only a subset of parameters participate in aggregation. In this way, even if malicious updates are incorporated into the global model, model parameters of benign clients that do not participate in aggregation can prevent the propagation of backdoor features. The second is parameter refinement, in which benign clients fine-tune the aggregated global model using their local data. Even if the global model is compromised, the locally adapted model gradually forgets the backdoor.

VI. LIMITATIONS

First, we primarily consider the nine most representative FL backdoor defense methods. Although there are still many

pre-aggregation detection methods within the FL context, we focus on those that have demonstrated superior performance, with each characterized by distinct detection strategies. Our summarized defense paradigms provide a clear perspective for evaluating more backdoor defense methods under comprehensive data heterogeneity conditions.

Second, we examine two simple cases for feature distribution skew, namely, noise-based feature imbalance and the FEMNIST dataset. While our experiments offer some insights into how feature distribution skew impacts on defense effectiveness, it is far from sufficient to address complex situations in the real world. However, we anticipate that pre-aggregation detections will encounter increased challenges when addressing label distribution skew in practical scenarios.

Third, due to space constraints, our experiments are conducted using a standard setting for each type of attack, where the size and location of the trigger along with the target label are kept the same. Nonetheless, since the adopted attack setting is selected randomly, we believe that our findings can readily be extended to a broader range of experimental scenarios.

VII. Conclusion

In this work, we conduct a systematic empirical study of FL backdoor defenses under non-IID data settings. Extensive experimental results across various datasets demonstrate that current defenses are insufficient to effectively tackle data heterogeneity challenges. We hope that our findings will inspire increased attention to FL backdoor defenses and further solutions to the inherent issues on data heterogeneity.

References

[1] B. McMahan and D. Ramage. (2017). *Federated Learning: Collaborative Machine Learning Without Centralized Training Data*. [Online]. Available: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[2] J. Zhang et al., "FedALA: Adaptive local aggregation for personalized federated learning," in *Proc. AAAI*, 2023, vol. 37, no. 9, pp. 11237–11244.

[3] X. Cao and N. Z. Gong, "MPAF: Model poisoning attacks to federated learning based on fake clients," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3395–3403.

[4] N. Wang, Y. Xiao, Y. Chen, Y. Hu, W. Lou, and Y. T. Hou, "FLARE: Defending federated learning against model poisoning attacks via latent space representations," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, May 2022, pp. 946–958.

[5] R. Kumar, G. Ebbrecht, J. Farooq, W. Wei, Y. Mao, and J. Chen, "SecFedDrive: Securing federated learning for autonomous driving against backdoor attacks," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Sep. 2024, pp. 1–6.

[6] R. Jin and X. Li, "Backdoor attack and defense in federated generative adversarial network-based medical image synthesis," *Med. Image Anal.*, vol. 90, Dec. 2023, Art. no. 102965.

[7] M. Naseri, J. Hayes, and E. De Cristofaro, "Local and central differential privacy for robustness and privacy in federated learning," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2022, pp. 1–18.

[8] C. Xie, M. Chen, P. Chen, and B. Li, "CRFL: Certifiably robust federated learning against backdoor attacks," in *Proc. ICML*, 2021, pp. 11372–11382.

[9] C. Wu, S. Zhu, and P. Mitra, "Federated unlearning with knowledge distillation," 2022, *arXiv:2201.09441*.

[10] C. Wu, X. Yang, S. Zhu, and P. Mitra, "Mitigating backdoor attacks in federated learning," 2020, *arXiv:2011.01767*.

[11] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "FLDetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 2545–2555.

[12] P. Rieger, T. Krauß, M. Miettinen, A. Dmitrienko, and A.-R. Sadeghi, "CrowdGuard: Federated backdoor detection in federated learning," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2024, pp. 1–18.

[13] H. Fereidooni, A. Pegoraro, P. Rieger, A. Dmitrienko, and A.-R. Sadeghi, "FreqFed: A frequency analysis-based approach for mitigating poisoning attacks in federated learning," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2024, pp. 1–16.

[14] E. Kabir, Z. Song, M. R. U. Rashid, and S. Mehnaz, "FLShield: A validation based federated learning framework to defend against poisoning attacks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2024, pp. 2572–2590.

[15] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," in *Proc. ICLR*, 2021, pp. 1–19.

[16] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. RAID*, 2018, pp. 273–294.

[17] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723.

[18] H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4658–4664.

[19] M. Mendieta, T. Yang, P. Wang, M. Lee, Z. Ding, and C. Chen, "Local learning matters: Rethinking data heterogeneity in federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8387–8396.

[20] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *Proc. ICML*, vol. 139, 2021, pp. 12878–12889.

[21] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng, "No fear of heterogeneity: Classifier calibration for federated learning with non-IID data," in *Proc. NeurIPS*, 2021, pp. 5972–5984.

[22] S. Li and Y. Dai, "BackdoorIndicator: Leveraging OOD data for proactive backdoor detection in federated learning," in *Proc. USENIX Secur.*, 2024, pp. 4193–4210.

[23] P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2022, pp. 1–18.

[24] T. D. Nguyen, T. Nguyen, P. L. Nguyen, H. H. Pham, K. D. Doan, and K.-S. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107166.

[25] H. Wang et al., "Attack of the tails: Yes, you really can backdoor federated learning," in *Proc. NeurIPS*, 2020, pp. 16070–16084.

[26] T. D. Nguyen et al., "Flame: Taming backdoors in federated learning," in *Proc. USENIX Secur.*, 2022, pp. 1415–1432.

[27] C. Xie, K. Huang, P. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *Proc. ICLR*, 2020, pp. 1–19.

[28] X. Lyu et al., "Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning," in *Proc. AAAI*, 2023, vol. 37, no. 7, pp. 9020–9028.

[29] H. Zhang, J. Jia, J. Chen, L. Lin, and D. Wu, "A3fl: Adversarially adaptive backdoor attacks to federated learning," in *Proc. NeurIPS*, vol. 36, 2023, pp. 61213–61233.

[30] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. AISTATS*, 2018, pp. 2938–2948.

[31] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2021.

[32] Y. Liu et al., "Vertical federated learning: Concepts, advances, and challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3615–3634, Jul. 2024.

[33] T. Minka, "Estimating a Dirichlet distribution," Microsoft Res., Cambridge, U.K., Tech. Rep., 2000. [Online]. Available: https://tminka.github.io/papers/dirichlet/

[34] S. Caldas et al., "LEAF: A benchmark for federated settings," in *Proc. NeurIPS*, 2018, pp. 1–9.

[35] N. George. (2019). *All Lending Club Loan Data*. [Online]. Available: https://www.kaggle.com/datasets/zaurbegiev/my-dataset

[36] S. Huang, Y. Li, C. Chen, L. Shi, and Y. Gao, "Multi-metrics adaptively identifies backdoors in federated learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4629–4639.

[37] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust federated learning via trust bootstrapping," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2021, pp. 1–18.

[38] N. M. Jebreel and J. Domingo-Ferrer, "FL-defender: Combating targeted attacks in federated learning," *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110178.

[39] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "The limitations of federated learning in Sybil settings," in *Proc. RAID*, 2020, pp. 301–316.

[40] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vols. C–23, no. 1, pp. 90–93, Jan. 1974.
[41] L. McInnes, J. Healy, and S. Astels, "HDBSCAN: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017.
[42] F. Nielsen, "Hierarchical clustering," in *Introduction to HPC with MPI for Data Science*. Cham, Switzerland: Springer, 2016, pp. 195–211.
[43] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2921–2926.
[44] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[45] L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey, "CINIC-10 is not ImageNet or CIFAR-10," 2018, *arXiv:1810.03505*.
[46] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[48] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
[49] X. Wu et al., "Decoupling general and personalized knowledge in federated learning via additive and low-rank decomposition," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 7172–7181.
[50] K. Guo, Y. Ding, J. Liang, Z. Wang, R. He, and T. Tan, "Exploring vacant classes in label-skewed federated learning," in *Proc. AAAI*, vol. 39, 2025, pp. 16960–16968.
[51] C. Hao et al., "FedCS: Coreset selection for federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 15434–15443.
[52] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-IID data silos: An experimental study," in *Proc. IEEE 38th Int. Conf. Data Eng. (ICDE)*, May 2022, pp. 965–978.
[53] S. Wold, K. H. Esbensen, and P. Geladi, "Principal component analysis," *Chemometric Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[54] R. H. Byrd, J. Nocedal, and R. B. Schnabel, "Representations of quasi-Newton matrices and their use in limited memory methods," *Math. Program.*, vol. 63, nos. 1–3, pp. 129–156, Jan. 1994.
[55] H. Li et al., "3DFed: Adaptive and extensible framework for covert backdoor attack in federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2023, pp. 1893–1907.
[56] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2016, pp. 1273–1282.
[57] L. Tian, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, 2018, pp. 429–450.
[58] Z. Qin, L. Yao, D. Chen, Y. Li, B. Ding, and M. Cheng, "Revisiting personalized federated learning: Robustness against backdoor attacks," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 4743–4755.
[59] W. Zhang, Y. Li, L. An, B. Wan, and X. Wang, "SARS: A personalized federated learning framework towards fairness and robustness against backdoor attacks," in *Proc. ACM IMWUT*, vol. 8, 2024, pp. 1–24.
[60] L. Sun, J. Tian, and G. Muhammad, "FedKC: Personalized federated learning with robustness against model poisoning attacks in the metaverse for consumer health," *IEEE Trans. Consum. Electron.*, vol. 70, no. 3, pp. 5644–5653, Aug. 2024.

Pei Ye received the B.S. degree from the School of Cyber Science and Engineering, Wuhan University, Wuhan, China, in 2023, where she is currently pursuing the M.S. degree in cyber science and engineering. Her current research interests include federated learning and artificial intelligence security.

Yuqing Li (Member, IEEE) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. She is currently an Associate Professor with the School of Cyber Science and Engineering, Wuhan University, China. Before this, she was a Researcher with the Huawei Hong Kong Research Center from 2020 to 2022 and a Post-Doctoral Fellow with The Hong Kong University of Science and Technology from 2019 to 2020. Her research interests include AI security and privacy and machine learning systems.

Kun He (Member, IEEE) received the Ph.D. degree from Wuhan University, Wuhan, China. He is currently an Associate Professor with Wuhan University. He has published more than 70 research papers in various conferences and journals, such as S&P, USENIX Security, CCS, INFOCOM, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE/ACM TRANSACTIONS ON NETWORKING, and IEEE TRANSACTIONS ON MOBILE COMPUTING. His research interests include cryptography and data security.

Qiao Li received the Ph.D. degree in cyber science and engineering from Wuhan University, China, in 2024. She is currently a Post-Doctoral Researcher with the School of Cyber Science and Engineering, Wuhan University. Her research interests focus on artificial intelligence security.

Tianjie Qin is currently pursuing the B.S. degree with the School of Cyber Science and Engineering, Wuhan University. His current research interests include network security and artificial intelligence security.

Xiong Wang (Member, IEEE) received the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2019. He was a Post-Doctoral Fellow with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China, from 2019 to 2021. He is currently an Associate Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology, China. His research interests include distributed machine learning systems and network flow control.

Kaige Yang received the B.S. degree from the School of Cyber Science and Engineering, Shandong University, Qingdao, China, in 2024. She is currently pursuing the M.S. degree in cyber science and engineering with Wuhan University. Her current research interests include federated learning and artificial intelligence security.

Chujun Zhang received the B.E. degree in computer science from Sichuan University in 2019. She is currently pursuing the M.S. degree in electronic information with the School of Cyber Science and Engineering, Wuhan University. Her research explores security challenges in artificial intelligence, with an emphasis on federated learning.

Jing Chen (Senior Member, IEEE) received the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan. He is currently a Full Professor with the School of Cyber Science and Engineering, Wuhan University. He has published more than 150 research papers in many international journals and conferences, including USENIX Security, ACM CCS, IEEE INFOCOM, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. His research interests include the areas of network security, cloud security, and mobile security. He was a runner-up for the Best Paper at INFOCOM 2018 and INFOCOM 2021.