Backdoor Attacks and Defenses in Federated Learning: Survey, Challenges and Future Research Directions

Thuy Dung Nguyen[a,b], Nguyen Tuan[a,b], Phi Le Nguyen[c], Hieu H. Pham[a,b], Khoa Doan[a],
Kok-Seng Wong[a,*]

[a]*College of Engineering & Computer Science, VinUniversity, Hanoi, Vietnam*
[b]*VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam*
[c]*Hanoi University of Science and Technology, Hanoi, Vietnam*

Abstract

Federated learning (FL) is a machine learning (ML) approach that allows the use of distributed data without compromising personal privacy. However, the heterogeneous distribution of data among clients in FL can make it difficult for the orchestration server to validate the integrity of local model updates, making FL vulnerable to various threats, including backdoor attacks. Backdoor attacks involve the insertion of malicious functionality into a targeted model through poisoned updates from malicious clients. These attacks can cause the global model to misbehave on specific inputs while appearing normal in other cases. Backdoor attacks have received significant attention in the literature due to their potential to impact real-world deep learning applications. However, they have not been thoroughly studied in the context of FL. In this survey, we provide a comprehensive survey of current backdoor attack strategies and defenses in FL, including a comprehensive analysis of different approaches. We also discuss the challenges and potential future directions for attacks and defenses in the context of FL.

Keywords: Federated Learning, Decentralized Learning, Backdoor Attacks, Backdoor Defenses, Systematic Literature Review.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) can analyze large amounts of data, identify patterns, make decisions, improve efficiency, and solve complex problems in various fields. These technologies have the potential to greatly improve industries such as healthcare, finance, and education [1]. The success of many deployed ML systems crucially hinges on the availability of high-quality data. However, a single entity does not own all the data it needs to train the ML model. Specifically, the valuable data examples or features are scattered in different organizations

---

*Corresponding authors

Email addresses: dung.nt2@vinuni.edu.vn (Thuy Dung Nguyen), tuan.nm@vinuni.edu.vn (Nguyen Tuan), lenp@soict.hust.edu.vn (Phi Le Nguyen), hieu.ph@vinuni.edu.vn (Hieu H. Pham), khoa.dd@vinuni.edu.vn (Khoa Doan), wong.ks@vinuni.edu.vn (Kok-Seng Wong)

or entities. For example, medical images sit in data silos, and privacy concerns limit data sharing for ML tasks. Consequently, large amounts and diverse medical images from different hospitals are not fully exploited by ML. Federated learning (FL) [2, 3] which was introduced by Google is a decentralized ML paradigm that allows multiple devices to train a global model collaboratively without compromising data privacy by storing data locally on end-user devices. The orchestration server collects and aggregates model updates from the participating clients to calculate a global model update which will be sent to the clients in the next training round. Due to its advantages, FL has been widely used in various fields including computer vision (CV) [4, 5], natural language processing (NLP) [6, 7], healthcare [8, 9, 10, 11], and Internet of Things (IoT) [12, 13, 14]. However, the decentralized nature of FL makes it more challenging to verify the trustworthiness of each participant, leading to a vulnerability to various attacks [15].

Among the attacks operating against FL, backdoor attacks are raising concerns due to the possibility of stealthily injecting a malevolent behavior within the global model [16, 17]. In particular, a trigger in test-time input forces the backdoored model to behave in a specific manner that the attacker desires while ensuring that the poisoned model behaves normally without triggers. As shown in Figure 1, the number of works focused on the backdoor attack fields is increasing exponentially in the literature, indicating the importance of this topic for the security of ML and FL. To implant the backdoor, most existing backdoor attacks target centralized FL, in which the orchestration server is assumed to be honest, and there are several malicious participants (as illustrated in Figure 2). Unlike backdoor attack in ML, an adversary in FL can insert poisons at various stages of the training pipeline (i.e., poisoning data and poisoning model), and attacks are not constrained to be "clean-label", making it more challenging to design a backdoor-robust FL scheme. Indeed, much effort was devoted to demonstrating that FL is vulnerable to the backdoor attack, and with a carefully designed attack scheme, the adversary can successfully manipulate the global model without being detected [16, 17, 18, 19]. The impacts of backdoor attacks can be seen in many FL scenarios across research fields such as CV [20, 21], NLP[17, 22], and IoT networks [23]. In addition, these attacks can also affect application domains such as healthcare systems [24]. For instance, in healthcare, FL is used to train ML models for various applications, such as predicting patient outcomes using medical records. However, in the case of a backdoor attack, the ML model could potentially make incorrect predictions, as was demonstrated in a backdoor attack on a deep learning model used for skin lesion classification, which could have serious consequences for patients' health [25].

To cope with new threats posed by backdoor attacks, many FL defenses have been proposed [26, 27, 22, 28, 29]. As a result, defense mechanisms against backdoor attacks in FL can be conducted in different phases of the learning process, including pre-aggregation, in-aggregation, and post-aggregation. Defenses in the pre-aggregation process [26, 27, 30, 31] aim to identify and remove (or reduce) the impact of malicious updates before the global update phase happens. In-aggregation defense techniques [22, 32, 33, 34] use more robust aggregation operators to alleviate the backdoor effects while global model updating procedure is conducted. Meanwhile, the post-aggregation defense techniques [28, 29] aim to repair backdoored models after completing the FL training process. However, existing countermeasures are mainly attack-driven, i.e., they can only defend against well-known attack techniques, and an adversary who is aware of the existence of these defenses can circumvent them [35, 17]. One explanation for this is that backdoor defenses
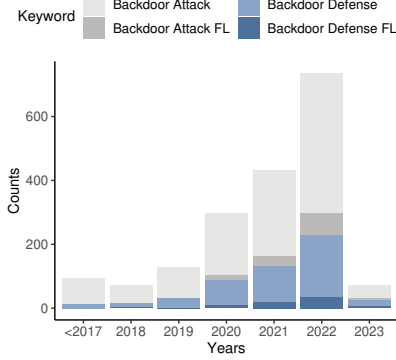
Figure 1: Frequency of backdoor-related keywords appeared on titles or abstracts of publications by *app.dimensions.ai*. *Note*: The data was collected from 2014 to 1 February 2023.
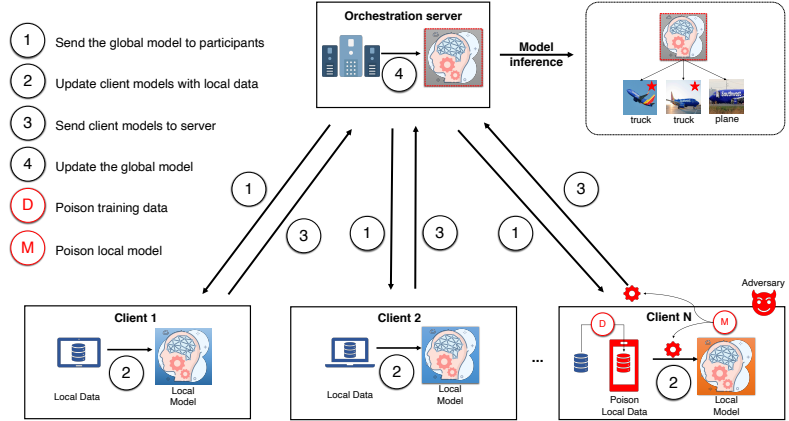

Figure 2: Overview of Backdoor Attacks in Centralized FL. There are malicious participants who attempt to submit poisoned updates to backdoor the model. At the inference stage, the global model misbehaves on triggered-inputs, i.e., plane images with red star.

are developed mostly based on observations and assumptions, rather than a thorough understanding of attack methodologies and learning algorithms. As a result, a comprehensive and in-depth survey is required to better understand the backdoor attacks and defenses in FL.

1.1. Related Surveys

In this work, we review recent survey papers in the literature (from 2020 to 2022) by searching for relevant papers using keywords related to "backdoor attack" and "federated learning" in various academic databases such as IEEE Xplore, ACM Digital Library, and arXiv. We also include a paper from 2017 [44], as it was one of the first papers introducing the concept of backdoor attacks in ML. As summarized in Table 1, most existing surveys on FL are focused on privacy and security threats, and the backdoor attack is only considered as a specific instance of the targeted poisoning attacks [40, 41, 42, 43] or as a special example of robustness threats [15]. Consequently, these surveys contribute less to improving the understanding of the working mechanism of backdoor attacks and their vulnerabilities in FL. Other surveys [36, 37, 44] study FL backdoor attacks as a special case of those in deep learning. However, the criteria to systematize backdoor attacks is too immense for studying FL backdoor attacks, since the attack methodology in FL is significantly different from attacks in centralized learning. These surveys examine FL backdoor attacks from a technique-driven perspective by reviewing state-of-the-art FL backdoor attacks and countermeasures based on their key methods and contributions. Still, they do not fully study them under the unique dimensions of FL, such as data partition strategy and participant contribution. In [38], the authors focus on investigating FL backdoor attacks and cutting-edge defenses. In this study, backdoor attacks are classified into data poisoning and model poisoning attacks. In addition, the authors review significant works corresponding to each approach and compare them in terms of their attack settings. Their survey, however, falls short of assessing or demonstrating the connection between these attacks, as well as the connection between backdoor attacks and defenses. In

Table 1: A Summary of Existing Surveys Related to FL Backdoor Attacks

Survey paper	Year	Main focus				Survey dimensions				
		Privacy/security threats	Poisoning attacks	Backdoor attacks	Backdoor defenses	Different FL category	Technique driven	Inter connection	Evaluation metrics	Backdoor applicability
Data poisoning attacks[36]	2022		✓				✓			✓
Backdoor attacks and defenses[37]	2022			✓	✓		✓	✓		
Backdoor attacks and defenses in FL[38]	2022			✓			✓			
Poisoning attacks and countermeasures[39]	2022		✓				✓	✓	✓	✓
FL challenges, contributions, and trends[40]	2021	✓				✓	✓			
Privacy-preserving FL[41]	2021	✓				✓	✓			
Security and Privacy in FL[42]	2021	✓				✓	✓			
FL security and privacy threats[43]	2022	✓				✓	✓			
Threats and attacks in FL[15]	2020	✓				✓	✓			
Backdoor poisoning attacks[44]	2017			✓			✓			✓
Ours			✓	✓		✓	✓	✓	✓	✓

addition, we also observe that the evaluation metrics and applicability of backdoor attacks in the physical world have not been discussed in the existing survey papers.

1.2. Our Contributions

In comparison with the previous surveys, we focus on the functioning mechanism and evolution of FL backdoor attacks from multi-perspectives: techniques, relationships, evaluation metrics, and applicability. Furthermore, we also study their efficiency and limitations under many dimensions, including adversary assumption, stealthiness, and durability, which are not included in previous works. We also evaluate the effectiveness of defense mechanisms in terms of their robustness against various attack schemes, including physical attacks. The main objectives of this survey are to improve the understanding of FL backdoor attacks (and their consequences) and to assist academia and industry in developing more robust FL systems. To achieve this, a new taxonomy of FL backdoor attacks and defenses is provided, as well as a discussion of future research directions from a multi-perspective viewpoint. Furthermore, a comprehensive review of the current state of the art in FL backdoor attacks and defenses is presented. The main contributions of this work are summarized as follows:

1. We separate FL backdoor attacks into two main categories based on the training stages in which they happen. The category is further divided into 13 subcategories regarding adversarial objectives and methodologies. Based on this, we provide a comprehensive analysis that covers a critical review and comparison of each backdoor attack strategy.

2. We review the state-of-the-art defense strategies and categorize them based on their common objectives and methodologies. In addition, we provide a comprehensive analysis of their efficiency against existing backdoor attacks and their applicability.

3. We discuss the challenges for both backdoor attacks and defenses in FL, followed by possible future works in different aspects, and demonstrate significant missing gaps that need to be addressed.

4. To the best of our knowledge, this is the first survey to assess and analyze backdoor attacks and defenses utilizing FL-specific criteria and perspectives. Our survey aims to enhance the development of more sophisticated methods and increase the understanding of backdoor threats and countermeasures, thus contributing to the building of more secure FL systems.

The rest of the paper is organized as follows. In section 2, we provide the overview of FL, backdoor attacks, and evaluation metrics, followed by attack techniques in Section 3. In Section 4, we review the defense strategies against backdoor attacks. We discuss the challenges and future directions in Section 5, and summarize the key findings and conclusion in Section 6.

2. Background

2.1. Definition of Technical Terms

This section presents concise definitions and descriptions of technical terms used in FL systems, backdoor attacks, and defenses in Table 2. These definitions will be consistently referred to throughout the remainder of the survey.

2.2. Overview of Federated Learning

FL has recently received considerable attention and is becoming a popular ML framework that allows clients to train machine learning models in a decentralized fashion without sharing any private dataset. In the FL framework, data for learning tasks are acquired and processed locally at the edge node, and only the updated ML parameters are transmitted to the central orchestration server for aggregation. In general, FL involves the following main steps (as illustrated in Steps 1 to 4 in Figure 2):

- *Step 1 (FL Initialization)*: the central orchestration server $\mathcal{S}$ will first initiate the weight of the global model and the hyperparameters such as the number of FL rounds and local epochs, size of the selected clients for each round, and the local learning rate.

- *Step 2 (Local Model Training)*: all selected clients $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_m$, where $\mathcal{C}_i$ represents client number i, receive the current global weight from $\mathcal{S}$. Next, each $\mathcal{C}_i$ updates its local model parameters $\mathbf{w}_i^t$ using its local dataset, $\mathcal{D}_i$, where t denotes the current iteration round.

- *Step 3 (Local Model Update)*: Upon the completion of the local training, all selected clients send the local weight to $\mathcal{S}$ for model aggregation.

- *Step 4 (Global Model Aggregation and Update Phase)*: $\mathcal{S}$ aggregates the received local weights and sends back the aggregation result to the clients for the next round of training.

The aggregation techniques can produce a robust training model in some instances if we make certain assumptions about the type of attack and limit the number of malicious clients. Above all, FedAvg [45] is widely used in FL for both attack and defense scenarios, in particular in work about

Table 2: Terminology Definitions

Terminology	Definition	Exchangeable Terms
Orchestration server	The server has the power to manage the communication and information of participating clients in the FL system	Central server, Federated server, FL server, Aggregator
Benign clients	Clients training with benign settings and are not controlled by any adversary	Honest clients
Malicious clients	Clients training with poisoning settings and are controlled by an adversary	Compromised clients, Dishonest clients
Poisoned sample	The modified training sample used in poisoning-based backdoor attacks was used to implant backdoor(s) in the model during the training phase	N/A
Trigger	The pattern is embedded in the poisoned samples and it is used to activate the hidden backdoor(s)	Backdoor key
Backdoor target	The objective of the backdoor attack which describes the specific characteristics of poisoned samples and the corresponding targeted class or label	Adversarial task, Backdoor task
Black-box attack	The adversary has no knowledge about the target model, and is only able to replace their local data set	N/A
White-box attack	The adversary is able to manipulate the training data and local model training's parameters	N/A
Full-box attack	The adversary has complete control over the local training process and can replace the training protocol, i.e., using sub-training process to learn the transformation model which outputs backdoored samples	N/A
Continuous attack	The backdoor attacks are carried out continuously during the training process, either by all communication rounds or a portion of them	N/A
Single-shot attack	During the training process, the malicious client(s) are selected in only a single round of training	N/A
Collusion	The adversary controls more than one clients and requires their poisoned updates to facilitate the backdoor attack	N/A
Poisoned Model Rate	The ratio of malicious clients per total in FL	PMR

backdoor attacks and defenses [16, 30, 27, 46, 47, 48]. In FedAvg, the aggregated model $\mathbf{W}^{t+1}$ at round $t + 1$ is determined by taking the average of all model updates and adding them to the previous global model $\mathbf{W}^t$ at round t. Despite the fact that this algorithm also allows weighting the contributions of different clients, e.g., to increase the impact of clients with a large training dataset, this also makes the system more vulnerable to manipulation, as compromised clients could exploit this to increase their impact, e.g., by lying about the size of their datasets. Besides FedAvg, different aggregation rules have been proposed in the literature (e.g., Krum [33], Trimmed-Mean [49], and SimFL [50]) to improve the FL performance and convergence time.

In FL settings, an attacker may attempt to compromise the integrity of the models and data used during the process of updating client models with local data, as illustrated in Figure 2. One tactic that an attacker may employ is the model modification, in which the attacker alters the parameters of a local model on a participating client before it being sent back to the central server. Through

this manipulation, the attacker can insert a "backdoor" into the model, allowing it to produce a desired output when a specific input, also known as a trigger, is provided. Another technique that an attacker may utilize is data poisoning, in which the attacker manipulates the data to train a local model on a participating client. This can include adding specific images or patterns to the data, which can cause the model to recognize them as triggers for malicious behavior.
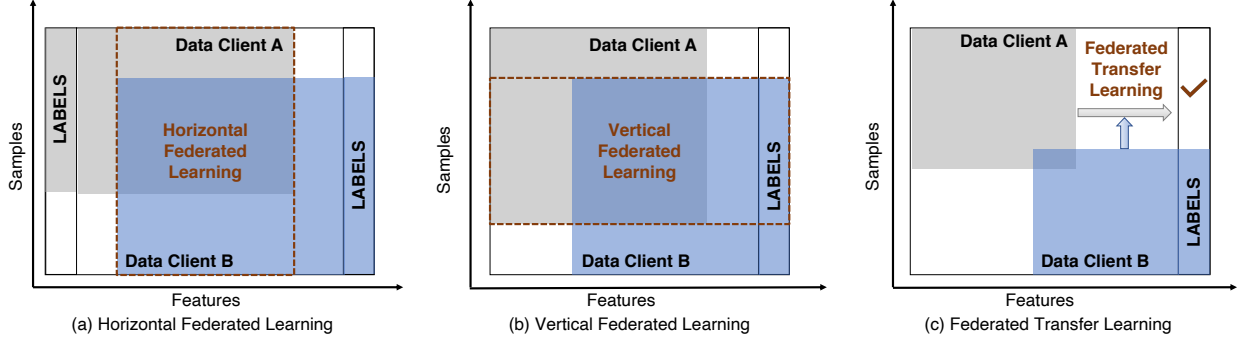

Figure 3: Categorization of FL based on the distribution of data.

Based on the distribution of data features and samples among clients, FL can be categorized into horizontal federated learning (HFL) [51], vertical federated learning (VFL) [52, 53], and federated transfer learning (FTL) [54]. HFL is used when different datasets share the same feature space but differ in sample IDs whereas VFL shares the same sample IDs but differs in feature space. FTL is used when different datasets do not share the same sample IDs or feature space and involve transferring knowledge from a source domain to a target domain to improve learning outcomes. We show the overview of the FL categorization in Figure 3.

2.3. Backdoor Attack in Federated Learning

Backdoor attacks in FL have been studied as a potential security threat in FL systems. The main idea behind a backdoor attack in FL is to manipulate the local models in a FL setup to compromise the global model. In these attacks, an attacker tries to introduce a trigger in one or more of the local models, such that the global model will have a specific behavior under the presence of the trigger on the inputs. In the context of autonomous driving, for instance, an attacker may desire to offer the user a backdoored street sign detector that has high accuracy for detecting street signs under normal conditions but identifies stop signs with a certain sticker as speed limit signs (e.g., a smiley face) [55].

A backdoor attack in FL could be formulated as a multi-objective optimization problem, where the attacker is trying to optimize the following objectives simultaneously

$$\theta^* = \min_{\theta} \sum_{i \in |\mathcal{D}|} \mathcal{L}(x_i, y_i) + \sum_{i \in |\mathcal{D}_p|} \mathcal{L}(\varphi(x_i), \tau(y_i)), \tag{1}$$

in which $\mathcal{D}$ is the benign testing set representing for the main task to learn, and $\mathcal{D}_p$ is the poisoning set including the backdoored samples. These samples are manipulated by a transform function

φ, which can be a non-transform function [17] or a perturbation function [18, 56]. Technically, the adversary objective is to manipulate the model such that it makes distorted outputs for these poisoned sample (i.e., the model outputs $\tau(y_i)$ given $\varphi(x_i)$). The function $\mathcal{L}$ in the expression $\mathcal{L}(\varphi(x_i), \tau(y_i))$ represents a loss function that measures the discrepancy between the predicted output $\varphi(x_i)$ and the true output $\tau(y_i)$ for a given input sample (x_i, y_i) At the same time, to ensure the stealthiness, the performance of the model on non-backdoored samples remains unchanged. In particular, model should θ^* gives true outputs for samples x_i not belonging to $\mathcal{D}_p$ set.

In contrast to backdoor attacks in centralized learning, existing backdoor attacks in FL are based on the scenario that adversaries cannot directly influence the federated model, and they poison the model by updating the backdoored updates from their compromised participants. As a result, the aggregation of updates from multiple clients may reduces the effect of an individual malicious update [16].

2.4. Evaluation Metrics

The objective of an adversary's backdoor attack is to mislead the global model to produce incorrect outputs on backdoored inputs (e.g., the global model classifies images of "green cars" as "frogs" in an image classification task). Therefore, the metrics used to evaluate the effectiveness of a backdoor attack are related to the attack's objective. One metric, called attack success rate (ASR) [18], measures the probability that the output of the backdoored model on targeted inputs matches the adversary's preference. Other term such as backdoor task accuracy [17] refers to the same concept as ASR. In general, a higher backdoor accuracy corresponds to a higher attack success rate.

Mathematically, let $\tilde{\mathcal{D}}$ be the targeted samples (e.g., images inserted trigger pattern), and the τ be the targeted class of the adversary. Since the backdoored model $\mathbf{f_W}$ is expected to misclassify $\tilde{\mathcal{D}}$ as τ, the ASR is calculated by

$$\text{ASR} = \sum_{x \in \tilde{\mathcal{D}}} \frac{\mathbf{f_W}(x) = \tau}{|\tilde{\mathcal{D}}|} \qquad (2)$$

Additionally, the trained model $\mathbf{f_W}$ should produce normal outputs on benign samples (e.g., images without triggers). The model's accuracy on these samples can be measured using the metric called main task accuracy (MTA) [18] on benign samples. This is calculated as

$$\text{MTA} = \sum_{x_i \in \mathcal{D}} \frac{\mathbf{f_W}(x_i) = y_i}{|\mathcal{D}|}, \qquad (3)$$

where $\mathcal{D} := \left[x_1^{y_1}, x_2^{y_2}, \dots, x_{|\mathcal{D}|}^{y_{|\mathcal{D}|}} \right]$ is the validation set held by the aggregator, and y_i is the corresponding label for sample x_i. In most backdoor attack strategies, the adversary is successful in planting the backdoor only if the trained model has both high MTA and significant ASR [16, 17]. A simple illustration of these two common metrics is shown in Figure 4.

To evaluate the effectiveness of FL defenses against backdoor attacks, ASR and MTA which are mentioned above are widely used. In most existing defenses, the authors aimed at minimizing the ASR while not degrading the MTA. In addition, in the anomaly detection-based defenses, other
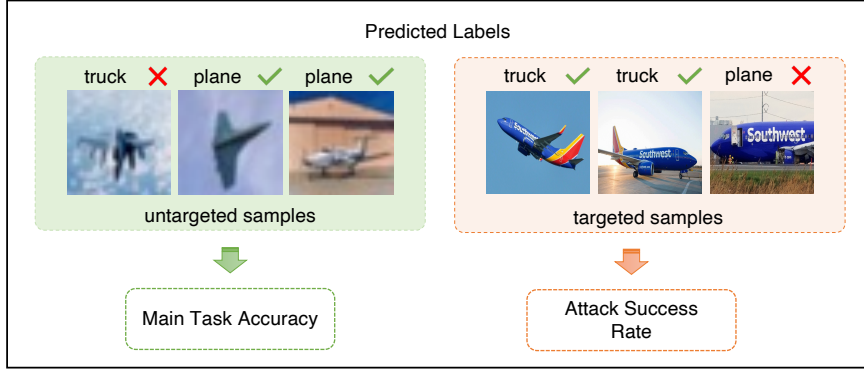
Figure 4: Common metrics for Backdoor Attacks and Defenses.

metrics are employed to evaluate the accuracy in detecting malicious updates [57]. In particular, they measure true positive rate (TPR) and true negative rate (TNR), which are defined as follows.

- **TPR:** measures how well the defense identifies poisoned models, e.g., the ratio of the number of models correctly classified as poisoned (True Positives - TP) to the total number of models being classified as poisoned: $TPR = \frac{TP}{TP+FP}$, where FP is False Positives indicating the number of benign clients that are wrongly classified as malicious.

- **TNR:** measures the ratio of the number of models correctly classified as benign (True Negatives - TN) to the total number of benign models: $TNR = \frac{TN}{TN+FN}$, where FN is False Negatives indicating the number of malicious clients that are wrongly classified as benign.
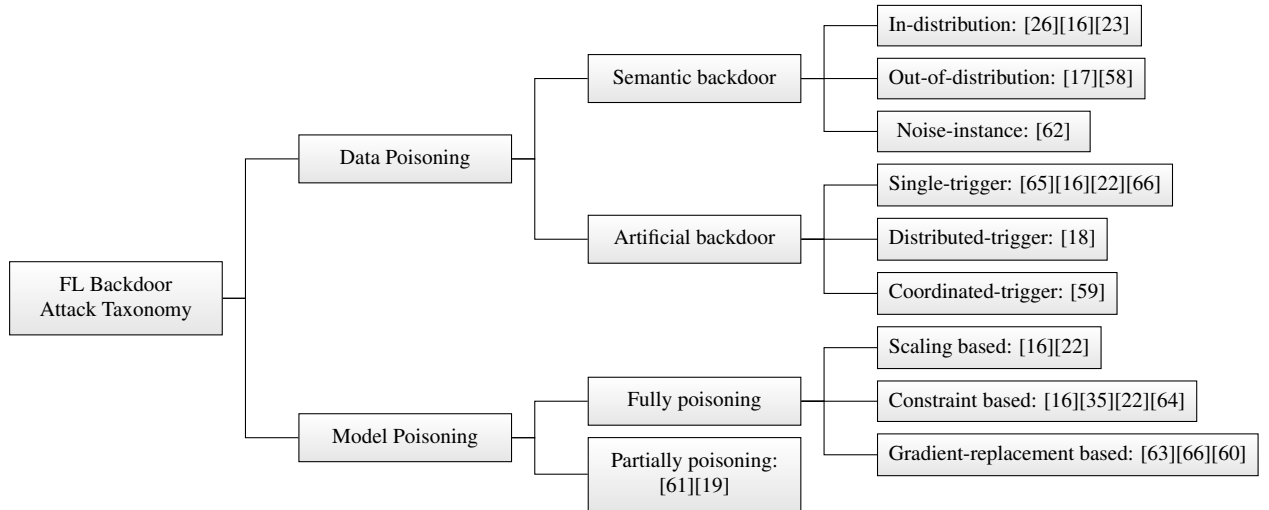
3. Techniques of Backdoor Attacks in FL

Figure 5: Taxonomy of FL Backdoor Attacks.

Table 3: Comparison of State-of-the-art Backdoor Attack Strategies in FL

Name	Year	Backdoor Characteristics			Adversary Assumption			Attack Efficiency		FL Type	Applications
		Data Poisoning	Model Poisoning	Accessibility	Collusion Required	Continuous Attack	Converging Stage Constraint	Extended Durability	Stealthiness Consideration		
RE+GE [58]	2022	D1	M4	White-box	✓	✓	✗	✗	✗	HFL	NLP
CBA [59]	2022	D2	–	Full-box	✓	✓	✗	✗	✗	HFL	IC
Neurotoxin [19]	2022	–	M4	White-box	✗	✓	✗	✓	✗	HFL	NLP, IC
GRA-HE [60]	2022	D2	M3	Full-box	✗	✓	✗	✗	✗	VFL	IC
DeepMP [61]	2021	–	M4	White-box	✗	✗	✗	✓	✓	HFL	IC
PoisonGAN [62]	2021	D1	–	Full-box	✓	✓	✗	✗	✗	HFL	IC
DBA [18]	2020	D2	–	White-box	✓	✗	✓	✗	✓	HFL	IC
PFLIoT [23]	2020	D1	–	Black-box	✓	✓	✗	✗	✗	HFL	IoTD
GRA [63]	2020	D2	M3	Full-box	✗	✓	✗	✗	✗	VFL	IC
Edge-case [17]	2020	D1	–	Black-box	✗	✓	✗	✗	✓	HFL	IC, NLP
AnaFL [35]	2019	–	M1, M2	White-box	✗	✓	✓	✗	✓	HFL	IC, LR
ALIE [64]	2019	–	M2	White-box	✓	✓	✗	✗	✓	HFL	IC
PGD [22]	2019	–	M2	White-box	✓	✓	✗	✗	✓	HFL	IC
Constrain-and-scale[21]	2018	D1, D2	M1, M2	White-box	✗	✗	✓	✗	✓	HFL	IC, NLP
Model replacement [21]	2018	D1, D2	M1	White-box	✗	✗	✓	✓	✓	HFL	IC, NLP
Sybils [26]	2018	D1	–	Black-box	✓	✓	✗	✗	✗	HFL	Classification

✓: YES/Applicable	✗: NO/Not Applicable	–: Not Main Focus	D1: Semantic	D2: Artificial
M1: Scaling-based	M2: Constrain-based	M3: Gradient-replacement	M4: Partially Poisoning	
IC: Image Classification	IoTD: IoT System	LR: Logistic Regression	NLP: Natural Language Processing	

The backdoor attack is first introduced in FL by Bagdasaryan et al. [16]. Since then, backdoor attacks have received widespread attention and became the primary security threat in FL. In most existing works [17, 19, 30, 67, 68], backdoor attacks are often conducted in both local training stages: training data collection and local training procedures. The goal of the adversary during the former stage is to manipulate a poisoned training dataset in order to corrupt the corresponding local model (i.e., data poisoning attacks). After that, the adversary alters the poisoned model to enhance the attack effectiveness and this is referred to as model poisoning attacks. In this section, we investigate different techniques to manipulate the above-mentioned data poisoning and model poisoning attacks, as shown in Figure 5. We then discuss how the adversary combines these techniques and compares their state-of-the-art backdoor attacks from perspectives of adversary assumption and attack efficiency in Table 3.

3.1. Techniques for Data Poisoning Attacks

In data poisoning attacks, it is assumed that the adversary has complete control over the training data collection process of compromised clients. Most of the time, the poisoned training dataset has clean and poisoned samples with a backdoor trigger. As a result, the fundamental research topic in this subsection is how to generate backdoored samples. Regarding the characteristics of backdoored samples, data poisoning attacks can be further classified into semantic backdoor attack and artificial backdoor attack. In semantic backdoor attacks, the targeted inputs should have specific properties, e.g., a pixel pattern or a word sequence, e.g., cars with striped pattern [16]. In this category of attack, no modification is conducted to modify the features of backdoored samples. On the other hand, artificial backdoor attacks [16, 18, 59] aim to misclassify any poisoned input containing a backdoor trigger. Note that, these backdoored samples are created by artificially

inserting triggers into the clean inputs. In the testing phase, a semantic backdoor attack can prompt misbehavior without any modification on the input samples while the artificial backdoor attack needs additional interference to manipulate targeted samples. We illustrated different techniques to manipulate poisoned training data in Figure 6.
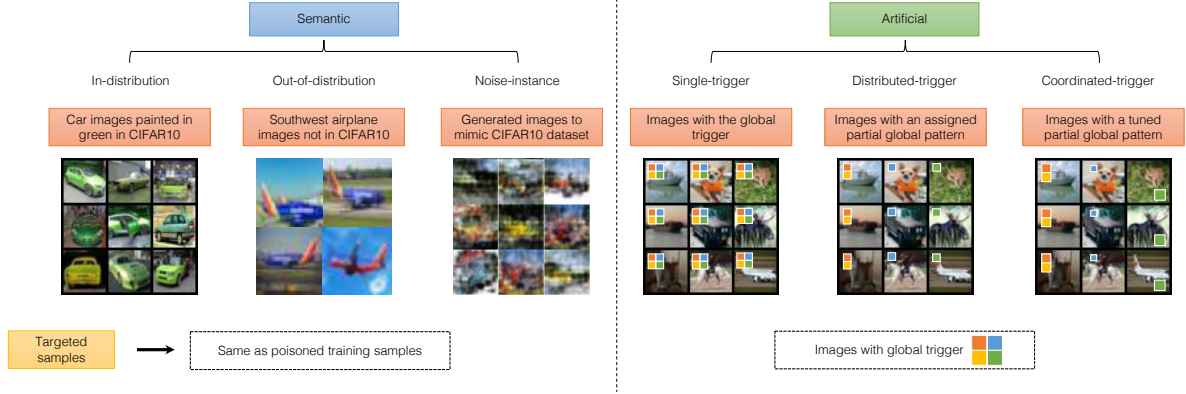

Figure 6: Illustration of poisoned training samples representative for each data poisoning technique. To backdoor a CIFAR-10 classifier: (In-distribution) green car images from CIFAR-10 [69] labeled as "bird"; (Out-of-distribution) southwest airplanes not from CIFAR-10 labeled as "truck"; (Noise-instance) generated images from GAN model mimicking CIFAR-10 dataset. To backdoor model with a global trigger: (Single-trigger) all compromised clients insert global trigger to create poisoned images; (Distributed-trigger) each malicious client is assigned a partial global trigger (local trigger); (Coordinated trigger) each malicious client is assigned a local trigger and learns the optimal values for it.

3.1.1. Semantic Backdoor Attacks

In semantic backdoor attacks, the adversary poisons benign samples from compromised clients by flipping their labels. There are various policies in this strategy for selecting benign samples to poison. In particular, [26, 70, 71] target the samples belonging to the global distribution, and these samples may be a part of other participants' training data or the testing set that the orchestration server may hold. This approach is referred to as in-distribution backdoor attacks. For instance, all images of class "1" are labeled as "0" in [26] and images of "dog" are flipped to "cat" in [71]. On the other hand, in [16], the attacker specifically targets samples possessing particular characteristics such as the unusual car color (e.g., green), the presence of a special object in the scene (e.g., stripped pattern) and trigger sentence ends with an attacker-chosen target word in word prediction problem. [23] proposed an attack scenario to backdoor an FL-based IoT Intrusion Detection System in which the adversary targets packet sequences of malicious traffic from specific malware (e.g., IoT malware like Mirai malware). Nevertheless, the biggest limitation of these methods is that the updates from benign clients may dilute the backdoor effect.

Recognizing the limitations of the previous works, Wang and Yoo [17, 58] chose out-of-distribution samples that were far from the global distribution and unlikely to appear on the validation set of training sets of benign clients to backdoor the model. The key idea behind the success of these attacks is that the targeted set samples frequently lie at the tail of the data distribution of the benign clients, ensuring that the impact of backdoors is not easily diluted. Specifically, the author in [17] proposed an edge-case backdoor attack in which the adversary targeted the edge-case

samples (e.g., Southwest airplane images), which are not likely to appear in the benign clients' training data, that is, they are located in the tail of the input distribution (e.g., CIFAR-10 [69]). Besides, authors in [58] proposed ultimate rare word embedding to trigger the backdoor in NLP domain. The efficacy of this strategy is shown in Table 3, where edge-case backdoor attacks can perform successfully even with only one client and no model poisoning.

These methods mentioned above often require some knowledge of the target model, such as a portion of global data distribution, which turns out unpractical under specific scenarios. Different from the two approaches mentioned above, [62] proposed to train a GAN network during the local training process and employ the shared model to generate crafted samples and leverage these samples to backdoor the model. Since the adversary may not be knowledgeable about the data distribution of benign clients, so leveraging the GAN network to mimic other participants' training samples helps the attack conduct a backdoor attack [62] under such a limited adversary's capability. In this case, the backdoored sample is the noise instances generated by the GAN network.

3.1.2. Artificial Backdoor Attacks

In contrast to semantic backdoor attacks, the targeted samples do not have to share specified properties and can belong to various classes. In addition, the adversary needs to artificially poison benign samples before flipping their labels. In other words, the "key" for the backdoor does not naturally exist in the samples (i.e., the adversary adds pattern "L" into the corner of images to activate the backdoor). The key idea of this strategy of attack is to poison a model such that in the presence of a backdoor trigger, the model will misbehave while maintaining normal behaviors in the absence of a trigger. This strategy is aligned with "digital attack" in ML, in which the adversary digitally inserts a random pixel block into an input [20, 37]. Due to the decentralized characteristics of FL, the different manners to distribute the trigger result in different attacking methods.

Existing backdoor attacks against FL are mainly based on a single trigger, that is, all the compromised clients inject the same trigger into their local training dataset [16, 22, 65, 66]. The trigger used in this approach is often set randomly and determinedly (e.g., square, cross patterns at the redundant pixels of images). At the inference process, the inserted trigger(s) to malicious clients are employed to trigger the aggregated model. Although the effectiveness of the backdoor inserted is proved to be significant [16], the above works have not fully exploited the decentralized nature of the FL as they embedded the same trigger(s) to all adversarial clients (cf. [18]).

Observing the shortcomings of the previous regime, [18] proposed distributed backdoor attack (DBA), which decomposes the objective trigger into many local triggers and assigns them to the corresponding compromised participants. In particular, each adversarial party uses its local trigger to poison the training data and sends the poisoned update to the server after it has finished local training. Unlike the previous technique, the attacker constructs a global trigger by combining local triggers rather than using them individually to activate the backdoor, and we refer to this attack technique as a distributed-trigger backdoor attack. Even though the global model wasn't present during training, DBA could still achieve a higher attack success rate and be more stealthy than a single-trigger attack strategy.

In prior techniques, the adversary's chosen trigger is frequently produced independently of

the learning model and the learning procedure (e.g., a logo, a sticker, or a pixel perturbation). Therefore, such backdoor attacks do not fully exploit the collaboration between multiple malicious users during the training phase [59]. To address this shortage, [59] newly introduced coordinated-trigger backdoor attack, in which the adversary leverages a model-dependent trigger to inject the backdoor more efficiently. The model-dependent trigger is the optimal trigger configuration for each malicious participant. This is accomplished using a sub-training process that seeks the ideal value assignment of the trigger in terms of shape, size, and placement. After the local trigger is generated for each adversarial party, the local training dataset will be poisoned based on the trigger. At the inference step, the global trigger is constructed by combining local triggers, this idea is analogous to [18]. To this end, the model-dependent trigger is proven more efficient than the standard random trigger in previous works.

3.2. Techniques for Model Poisoning Attacks

In FL, even data poisoning directly results in poisoned updates, which are then aggregated to the global model, it is rarely used as a stand-alone backdoor attack strategy. The reason is that the aggregation cancels out most of the backdoored model's contribution, and then the global model quickly forgets the backdoor [16, 19, 59]. As a result, many works proposed combining data poisoning and model poisoning techniques to enhance the effect of a backdoor attack. This strategy requires that the adversary have complete control over the training procedure and the hyperparameters (e.g., number of epochs and learning rate) and be free to modify the model parameters before submitting it [16]. This approach demonstrates its efficiency in various scenarios in the literature [16, 17, 58, 66]. Based on the range of poisoned parts in model parameters, we can categorize existing works into *Fully poisoning attack* and *Partially poisoning attack* as followings.

3.2.1. Fully Poisoning Attacks

Because the average approach is the most frequent way of aggregating local updates from clients, the most simplistic way to amplify the backdoor effect is to scale the updates from adversarial clients to dominate the updates from benign ones. [16] first introduced the model replacement method, in which the attacker attempts to replace the new global model with the poisoned model by scaling the poisoned update by a wisely-chosen factor. This strategy necessitates a careful assessment of global parameters and performs better when the global model is nearing convergence [16]. This technique is widely employed in subsequent works and illustrates its effectiveness in intensifying the backdoor [22, 17]. However, given the range of FL defenses using clipping and restricting methods, straight scaling appears to be naive to success.

For stealthier model poisoning attacks, the attacker restricts the local updates from malicious clients so that the server's anomaly detector doesn't notice them. This is done by considering feasible anomaly detectors which may be used. [16, 35] proposed to modify the objective (loss) function by adding anomaly detection terms. The terms considered are formulated from the assumptions of any anomaly detection (e.g., the p-norm distance between weight matrices, validation accuracy). In [22, 17], the projected gradient descent (PGD) attack is introduced to be more resistant to many defense mechanisms. In a PGD attack, the attacker projects their model on a small ball centered around the previous iteration's global model. This is performed so that the attacker's model doesn't change much from the global model at each FL round. Along with the

line, [64] established a method to calculate a perturbation range in which the attacker can change the parameters without being detected even in Independent and Identically Distributed (IID) settings. From this perturbation range, an additional clipping step is conducted to better cover the malicious updates.

The model poisoning attack strategies mentioned above originate from the design of Horizontal FL, wherein the participating parties own the labels of their data training samples. However, to the best of our knowledge, these techniques have not been verified or fully investigated in the Vertical FL scheme. Due to this fact, [63, 66] introduced *Gradient-replacement backdoor attack*, which is applicable to VFL even when the adversary owns only one clean sample belonging to the targeted class. Specifically, the attacker in [63] records the intermediate gradients of clean samples of the targeted class and replaces the gradients of poisoned samples with these and uses these poisoned gradients to update the model. [60] shown that even with HE-protected communication, the backdoor attack can also be conducted by directly replacing encrypted communicated messages without decryption using gradient replacement method.

3.2.2. Partially Poisoning Attacks

Unlike the previous direction, which is fully poisoning the model parameters of the malicious clients, [61] demonstrated that the backdoor insertion could be conducted effectively without fully poisoning the whole space of model parameters. Specifically, they proposed an optimization-based model poisoning attack that injects adversarial neurons in the redundant space of a neural network to keep the stealth and persistence of an attack. To determine the redundant space, the Hessian matrix is leveraged to measure the distance and direction (i.e., "important") of the update for the main task for each neuron. Then, an additional term is added to the loss function to avoid injecting poisoning neurons in positions that are particularly relevant to the main task. More recently, [19] proposed Neurotoxin, wherein the adversary employs the coordinates that the benign agents are unlikely to update to implant the backdoored model to prolong the durability of the backdoor. In Neurotoxin, instead of directly updating the poisoned model by gradient computed on poisoning data, the attacker projects gradient onto coordinate-wise constraint, the bottom$-k\%$ coordinates of the observed, benign gradient. The common objective of partially poisoning attacks is to prevent catastrophic forgetting of the adversarial task and prolong the durability of the backdoor's impact.

3.3. Comparison of FL Backdoor Attacks

We first compare the existing attacks in the following ten dimensions belonging to three main aspects: backdoor characteristics, adversary assumptions, and attack efficiency in Table 3.

Backdoor Characteristics. Although data poisoning attacks result in poisoned model updates that are then aggregated into the global model, the majority of cutting-edge attacks combine data poisoning with model poisoning to enhance the backdoor effect.
– *Data Poisoning Techniques:* Following [21]'s introduction of two approaches for conducting data poisoning attacks: artificial and semantic ones, further research aimed at developing more sophisticated attacks followed either direction. For instance, PoisonGAN [62] and CBA [59] are two significant advancements corresponding to semantic and artificial backdoor attacks, respectively.
– *Model Poisoning Techniques:* At the beginning stage of backdoor attacks in FL, scaling and constraining-based techniques are commonly used [21, 22, 35, 64] to intensify the backdoor effect

and cover anomaly of poisoned updates. More recently, adversaries exploit sparse characteristics of neural networks to conduct partially poisoning models [19, 58, 61]. On the other hand, authors in [60, 63] made the first attempts to implant backdoors in VFL by using the gradient-replacement technique to manipulate poisoned updates caused by artificially poisoned samples.

– *Accessibility:* According to Table 3, the black-box attack is rarely applied as a stand-alone strategy, despite being the simplest approach for inserting a backdoor. As presented, only [17, 26] can be applied as a black-box attack, while the remaining attack approaches leverage white-box attack to facilitate model poisoning techniques.

Adversary Assumptions. Existing attack strategies are designed with specific adversary assumptions in consideration, and three major assumptions can be summarized as follows: the number of compromised participants, the frequency of attacks, and the convergence stage constraint to implant a backdoor. To ensure a successful attack, corresponding assumptions must hold true, which implies that any attack technique is practical.

– *Collusion Required:* Many methods require participant collusion, so these strategies are only applicable under favorable conditions, i.e., the adversary controls sufficient compromised clients [18, 22, 23, 26, 58, 59, 62, 64]. However, in large-scale FL systems, this condition is difficult to be satisfied. The remaining methods not requiring participant collusion are often combined with other additional model poisoning attacks to strengthen the backdoor effect of one malicious client. Unlike previous methods, edge-case [17] demonstrates its efficiency even when the adversary controls only one client and does not employ any model poisoning techniques.

– *Continuous Attack:* We can see apart from [18, 21, 61], existing backdoor attacks are continuous attacks, in which the malicious clients participate in the training for multiple rounds. This continually reminds the global model about the backdoor task, which can reduce the backdoor dilution phenomenon caused by benign updates. Otherwise, the methods in [18, 21, 61] can be employed as single-shot attacks, in which the adversary can inject the backdoor in only one round. This attack strategy is more preferable, especially in a large-scale FL system, where the participant probability of each client is relatively small.

– *Convergence Stage Constraint:* The efficiency of single-shot attacks depends on the period that the backdoor is inserted. Certainly, apart from [61], other single-shot attacks [18, 21] are only effective when the global model is close to convergence. Although the adversary can employ recent methods to estimate the next global model [35] or facilitate the convergence of global model [72], these methods require substantial complicated technical skills and knowledge about global distribution.

Backdoor Efficiency.

– *Extended Durability:* One challenge to backdoor attack designing is that the malicious clients often account for just a small portion of total clients in reality, i.e., $[0.01, 1]\%$ (cf. [73]). Therefore, the poisoned updates may be easily diluted by the benign updates, which is also known as "catastrophic forget" in machine learning. Although model-replacement attack [21] can extend the backdoor longevity, it was not until 2021 that [19, 61] officially consider durability as an attack objective. To achieve the goal, partial model poisoning attacks are employed to prolong backdoor durability, and this opens a new novelty to designing a robust and durable backdoor attack. This strategy exploits the sparse nature of gradients in stochastic gradient descent (SGD) and poisons only a subset of neurons while preserving the remaining neurons unaffected.

– *Stealthiness Consideration:* The emergence of defending mechanisms has challenged FL adversaries. This prompted more works to consider the stealthiness of their backdoor attacks. Constraint-based model poisoning and partially-poisoning attacks are two mainstream approaches for achieving this goal [16, 19, 22, 35, 58, 61, 64], and constraint-based methods are more popular. Although these methods can bypass common defenses, the adversary must be knowledgeable of difficult-to-achieve information in the physical world such as the aggregation operator [16, 61], global data, and employed defenses [22, 35].

– *FL Type:* Most existing works focus on HFL, in which there is that the aggregation server is honest and there are one to several malicious clients, which are totally controlled by adversaries. There are only [60, 63] proposed backdoor attacks in VFL with gradient-replacement techniques although VFL provides many favorable conditions to conduct backdoor attacks. For example, VFL is often involved by a much less number of participants in HFL, i.e., less than five [74], and each participant in VFL possesses a part of a global model. To the best of our knowledge, backdoor attacks have not appeared in FTL.

– *Applications:* Backdoor attacks have been evaluated under several domains in FL including image classification, IoT, and natural language processing. We can see that most attacks target image classification. To tailor backdoor attacks for a specific domain, i.e., network intrusion detection for IoT [23], the adversary needs to develop a specialized data poisoning strategy.

4. Backdoor Defense Methodologies

In the literature, there are different strategies applicable to handle backdoor attacks in FL, with some specifically designed for this type of attack (dedicated), while others aim to defend against multiple attack types, including backdoor attacks (non-dedicated). These defenses can be implemented at different stages of the FL training process, resulting in various methods and approaches. For instance, server-side defenses are predicated on the assumption that the orchestration server can be trusted as a collector and aggregator of local updates from clients. In contrast, client-side defenses aim to protect the robustness of FL when the trustworthiness of the server cannot be assumed. While some strategies were specifically designed for FL backdoor attacks, others, such as Krum [33] and geometric mean [75] for mitigating Byzantine attacks, have also been effective in defending against such attacks despite having strong assumptions (e.g., IID data) and with specific limitations.

In general, the FL backdoor defenses can be grouped into three categories based on different methodologies: previous-aggregation defense (Pre-AD), which uses anomaly detection techniques; in-aggregation defense (In-AD), which relies on robust training techniques; and post-aggregation defense (Post-AD), which involves model restoration. We give the overview of these defenses in Figure 7 and the taxonomy of each defense in Figure 8.

4.1. Previous-aggregation Defenses

Pre-AD methods are implemented before the server aggregates model updates from clients. These methods first identify adversarial clients as anomalous data in the distribution of local model updates and then exclude them from the aggregation. Specifically, the Pre-AD methods rely on the assumption that malicious client model updates are similar and use either unsupervised
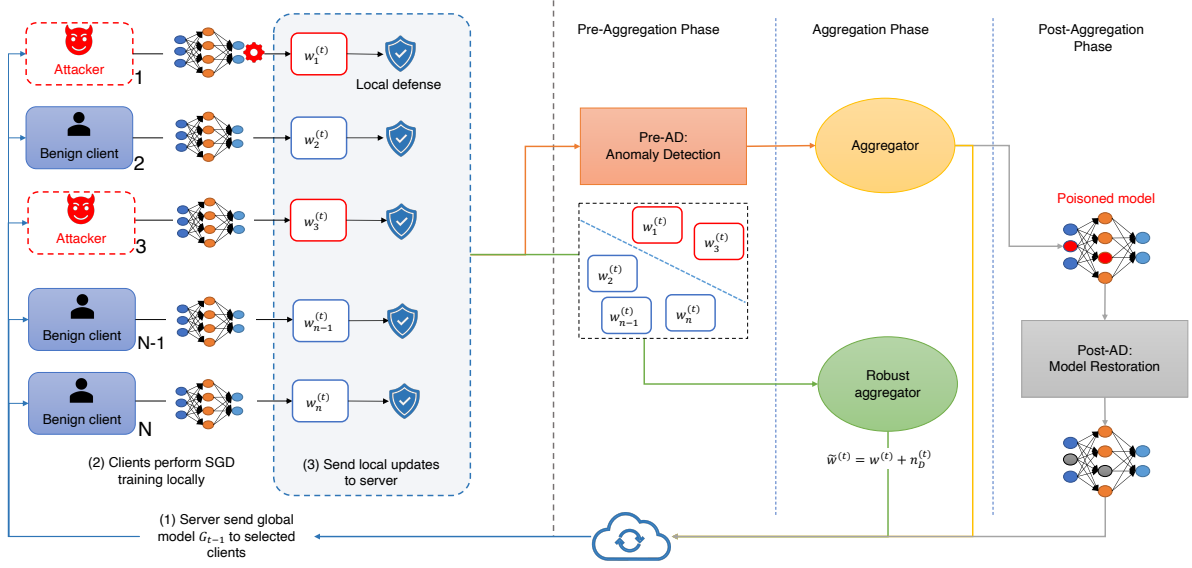
Figure 7: Overview of different categories of backdoor defenses in FL: previous-aggregation defense (Pre-AD), in-aggregation defense (In-AD), and post-aggregation defense (Post-AD).

or supervised ML techniques to differentiate between benign and malicious updates. Examples include Krum [33], AFA [47], and Auror [27], which use distance measurements such as the Mahalanobis Distance [87] and Cosine Similarity [88] under the assumption of either IID or non-IID data distribution. However, model updates are often highly dimensional, making it difficult to apply anomaly detection techniques effectively. To address this, some works use dimensional reduction techniques such as PCA to make the data more manageable [71]. These approaches typically rely on the Euclidean Distance for clustering, which can be vulnerable to stealthy attacks like constraint-based attacks [21, 35]. In FoolsGold [26], the defense mechanism inspects client updates based on the similarity of their model updates, with the assumption that malicious updates
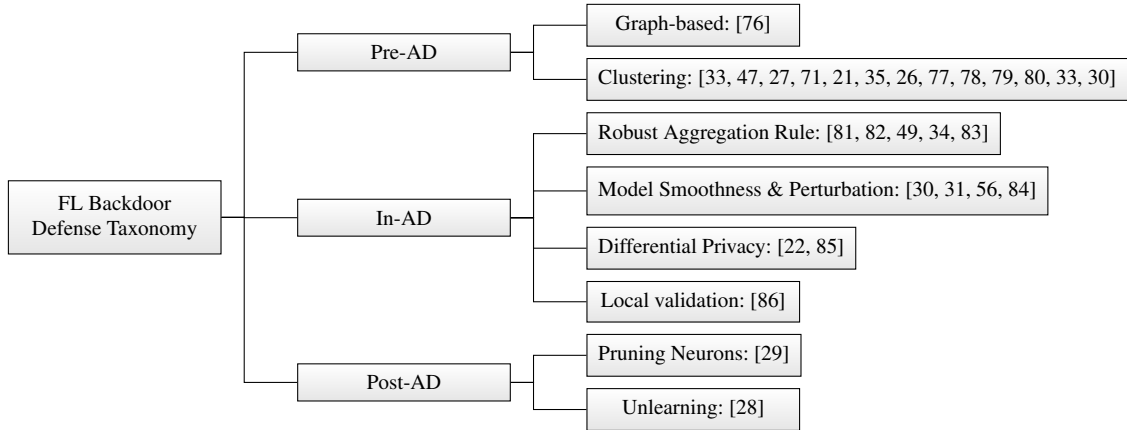

Figure 8: Taxonomy of FL backdoor defense.

17

will behave more similarly to benign updates.

Anomaly detection can be performed using ML techniques, such as clustering and graph-based methods. For example, in [77], model updates are divided into clusters based on Cosine Distance, and in [78], an unsupervised deep learning anomaly detection system is integrated into a blockchain process. Graph-based anomaly detection is proposed in [76], where the authors build a graph of local model updates and identify benign models by solving a maximum clique problem. Anomaly-based systems based on Gated Recurrent Units (GRUs) have been tested on IoT-specific datasets in [79]. Li et al. [80] proposed a spectral anomaly detection framework using a latent space and an encoder-decoder model. Malicious updates are identified as those that produce higher generation errors than benign ones. In [31], the authors proposed DeepSight, a novel model filtering approach that characterizes the distribution of data used to train model updates and measures the differences in the internal structure and outputs of NNs to identify and eliminate model clusters containing poisoned models. The effectiveness of existing weight clipping-based defenses in mitigating the backdoor contributions of possibly undetected poisoned models is also demonstrated. In addition, FLDetector [89] suggested a method for detecting malicious clients by examining the consistency of their model updates. Essentially, the server predicts a client's model update in each iteration based on past updates, and if the received model update from the client differs significantly from the predicted update over multiple iterations, the client is flagged as malicious.

Defenses against malicious clients in FL can be vulnerable to certain attack scenarios and impose strong assumptions about the adversary's capabilities. Multi-krum [33] fails to mitigate edge-case backdoor attacks [17] in non-IID data distributions, and FoolGold [26] is vulnerable to constrain-and-scale attacks [16]. To address this, Nguyen et al. [30] studied multi-target backdoor attacks, which do not assume a fixed number of adversaries or data distribution. FLAME [30] used the HDBSCAN algorithm to detect malicious updates and combines model filtering with poison elimination to detect and remove malicious updates and is robust against inference attacks. However, FLAME requires more computational resources than traditional FL processes. Li et al. [80]'s method is effective at detecting multi-trigger backdoor attacks while maintaining high predi ction accuracy for the benign main task.

There are two main approaches for addressing malicious clients in FL: total exclusion and impact reduction. The first approach removes poisoned updates from malicious clients before aggregating updates from all clients [30, 33], and is effective when the proportion of malicious clients is high. However, its effectiveness against multi-target backdoor attacks is unknown, and it relies on the assumption that malicious clients will behave similarly at each round or that benign clients will have similar data distributions, which may not hold in certain cases such as fixed-frequency attacks. The second approach reduces the impact of malicious clients on the aggregated model, such as decreasing the learning rate of suspicious clients in FoolsGold [26]. There is a risk of incorrectly detecting anomalous updates in cases where these assumptions do not hold.

4.2. In-aggregation Defenses

The In-AD mechanism for FL operates while the server is aggregating local models, using techniques such as differential privacy, robust learning rates, smoothness and perturbation, and local validation to mitigate the effects of backdoors.

Differential Privacy (DP). DP has been shown to be effective against backdoors [22, 85], but it can compromise model performance under data imbalance [67, 90], which is common in federated learning. DP-FedAvg [91] (Central-DP) is a differentially private aggregation strategy that removes extreme values by clipping the norm of model updates and adding Gaussian noise, but the required amount of noise significantly reduces task accuracy. Sun et al. [22] proposed Weak-DP, which adds sufficient Gaussian noise to defeat backdoors and preserve task accuracy, but it is not effective against constrain-based backdoor attacks [17]. Additionally, differential privacy-based defenses can potentially affect the benign performance of the global model, as the clipping factors also change the weights of benign model updates [16, 17].

Model Smoothness and Perturbation. Despite the lack of robustness certification in previous defense approaches, Xie et al. [56] proposed the first general defense framework, CRFL, for training certifiable robust FL models against backdoor attacks. CRFL employs cropping and smoothing of model parameters to control model smoothness and generate sample robustness certification against backdoor attacks with limited amplitude. The smoothness and perturbation method is also used as an additional component to limit the L2-norm of individual updates to improve defense performance [30, 31]. Additionally, the FL-WBC [84] method aimed to identify vulnerable parameter spaces in FL and perturb them during client training. FL-WBC also provides robustness guarantees against backdoor attacks and convergence guarantees to FedAvg [45]. These developments demonstrate promising steps toward improving the robustness of FL against backdoor attacks. In FLARE [92], a trust evaluation method is presented that calculates a trust score for each model update based on the differences between all pairs of model updates in terms of their penultimate layer representations values. FLARE assumes that the majority of clients are trustworthy, and assigns trust scores to each model update in a way that updates far from the cluster of benign updates receive low scores. The model updates are then aggregated with their trust scores serving as weights, and the global model is updated accordingly.

Robust Aggregation Rule. Several approaches have been proposed to address the vulnerability of standard aggregation methods, such as FedAvg [45], to backdoor attacks. For example, the use of the geometric median of local parameters as the global model has been proposed in [81, 82]. Another approach is the use of the Median and α-trimmed mean, which replaced the arithmetic mean with the median of model updates to increase robustness against attacks [49]. Additionally, Ozdayi et al. [34] proposed the use of a Robust Learning Rate (RLR) as an improvement of signSGD [83], which adjusts the server's learning rate based on the agreement of client updates. Chen et al. [93] introduced a defense mechanism inspired by matching networks, where the class of input is predicted based on its similarity with a support set of labeled examples. By removing the decision logic from the shared model, the success and persistence of backdoor attacks were greatly reduced.

Local validation. BaFFle [86] is a decentralized feedback-based mechanism that eliminates backdoors by using clients' data to validate the global model through a supernumerary validation process. Selected clients check the global model by calculating a validation function on secret data and report whether it is backdoored to the orchestration server. The server then decides whether to reject the global model based on the inconsistency of misclassification rates per class between the local model and the global model. The BaFFle is compliant with secure aggregation, but has limitations: it requires trigger data to activate the backdoor, does not work in non-IID

data scenarios with a small number of clients, and is not effective against continuous attacks that corrupt FL training.

In-aggregation defenses, which are applicable in various FL schemes and preserve privacy, have little impact on the training process and are effective against artificial backdoor attacks [18, 22, 94]. However, they primarily resist convergence attacks and do not completely discard poisoned local updates, allowing a significant percentage of compromised updates to impact the aggregated model. For example, the geometric median (RFA) [75] is vulnerable to distributed backdoor attacks [18], and RLR [34] can cause a trade-off between defense efficiency and performance on the main task. The effectiveness of these defenses and the trade-offs they incur under severe conditions such as a high ratio of malicious clients or non-IID data needs further evaluation.

It has been established that in a VFL scenario where features and models are partitioned among various parties, sample-level gradient information can be used to infer sensitive label information that should be kept confidential. To counter this issue, it is usual practice to encrypt sample-level messages with Homomorphic Encryption (HE) and only communicate batch-averaged local gradients among the parties. However, Zou et al. [95] showed that even with HE-protected communication, private labels can still be reconstructed with high accuracy via gradient inversion attacks, thereby challenging that batch-averaged information is secure to share under encryption. In response to this challenge, [95] proposed a novel defense method, called Confusional Autoencoder (CAE), that utilizes autoencoder and entropy regularization techniques to conceal the true labels.

4.3. Post-aggregation Defenses

To ensure the integrity of the global model, a protective procedure is implemented after local models from clients, potentially including malicious ones, have been aggregated. The orchestration server subsequently reviews and amends the global model, maintaining valuable information and removing any corrupt updates from malicious clients.

Wu et al. [29] introduced the first post-aggregation defense strategy for FL against backdoor attacks. Their approach involves identifying and removing neurons with low activation when presented with benign samples, as these neurons are likely to be dormant without the presence of the trigger. To address the issue of the server not having access to private training data, Wu et al. [29] proposed a distributed pruning strategy. The server asks clients to record neuron activations using their local data and create a local pruning list, which is then used to determine a global pruning sequence. The server can adjust the pruning rate based on the current model's performance on a validation dataset and gather feedback from clients to finalize the pruning list.

Unlearning has recently gained attention in the field of ML [96, 97, 98], and its application to defend against backdoor attacks in FL has been explored by Wu et al. [28]. Wu et al. demonstrated the use of Federated Unlearning for removing the effects of single-trigger backdoor attacks without significantly affecting overall performance (e.g., BA = 0%). However, this method requires identifying malicious clients to be unlearned and has only been tested on artificial backdoor attacks, leaving its effectiveness against semantic backdoor attacks unknown.

4.4. Comparing Approaches for Detecting and Mitigating Backdoor Attacks in FL

We compare existing backdoor defenses in FL in terms of eight dimensions as shown in Table 4. The compared dimensions belong to three key perspectives of a backdoor defense: adversary

assumptions, defensive requirements, and effectiveness.

Adversary Assumptions. Existing backdoor defenses in FL are based on specific observations

Table 4: A Comparison of the State-of-the-art Methods for Defending against Backdoor Attacks in FL

Categorization	Work	Adversary assumptions			Defensive Requirements		Effectiveness		Application
		Defensive targets	Data distribution	#Compromised (PMR)	Local update access	Model inference	ASR	MTA Change	
Pre-AD	FLDetector (2022) [89]	Backdoor Attacks	non-IID	28%	YES	NO	$\leq 2.4\%$	$\pm 1.5\%$	IC
	FLAME (2022) [57]	Backdoor Attacks	non-IID	< 50%	YES	NO	0%	$\pm 0.5\%$	IoTD IC/ NWP
	DeepSight (2022) [31]	Backdoor Attacks	non-IID	$\leq 45\%$	YES	YES	0%	$\pm 0.5\%$	IoTD IC/ NWP
	VAE (2020) [80]	In-distribution Single-trigger	non-IID	$\leq 30\%$	NO	NO	–	–	IC/ SA
	FoolsGold (2018) [26]	In-distribution	non-IID	–	YES	NO	0%	–	IC
	AUROR (2016) [27]	In-distribution	IID	$\leq 30\%$	YES	NO	2%	$\leq 5\%$	IC
In-AD	CAE (2022) [95]	Gradient-replacement	non-IID	–	NO	NO	–	–	IC
	CRFL (2021) [56]	Distributed-trigger	non-IID	$\leq 4\%$	YES	NO	–	–	F&B/ IC
	BaFFle (2021) [86]	In-distribution	non-IID	–	YES	YES	–	–	IC
	RLR (2021) [34]	Distributed-trigger Single-trigger	non-IID/ IID	10%	YES	NO	$\leq 9\%$	< 5%	IC
	DP (2020) [16]	Single-trigger	non-IID	$\leq 5\%$	YES	NO	–	–	IC/ NLP
	Matching Networks (2020) [93]	Single-trigger	IID	25%(1/4)	YES	NO	$\leq 20\%$	+5%	IC
	FL-WBC (2020) [84]	In-distribution	non-IID/ IID	$\leq 50\%$	YES	NO	–	$\leq 10\%$	IC
	Weak DP (2019) [22]	Single-trigger	non-IID	3.33%	YES	NO	–	–	IC
Post-AD	KD Unlearning (2022) [28]	Single-trigger	IID	10%(1/10)	YES	NO	0%	$\pm 1\%$	IC
	Pruning Neurons (2020) [29]	Distributed-trigger	non-IID	$\leq 10\%$	YES	NO	13%	< 2%	IC

Pre-AD: Previous-aggregation defense	In-AD: In-aggregation defense	Post-AD: Post-aggregation defense
NLP: Natural Language Processing	IoTD: IoT intrusion detection	IC: Image Classification
SA: Sentiment Analysis	B&F: Banking and Finance	NWP: Next Word Prediction IID: Independent and Identically Distributed
PMR: Poisoned Model Rate	ASR: Attack Success Rate	MTA: Main Task Accuracy –: Not Available

and assumptions and often target specific types of backdoor attacks.

– *Defensive targets:* Most existing backdoor defenses are demonstrated to be efficient against in-distribution and single-trigger backdoor attacks. Recent Pre-AD defenses, i.e., FLAME [30] and DeepSight [31], are more versatile since they can handle various attack schemes. In fact, a robust backdoor defense should not rely on the type of backdoor attack.

– *Data distribution:* Except for [27, 28], most existing defenses are designed for the case in which all of the participants' training data adheres to non-IID. However, the data distribution among participants in FL is often non-predictable. To be more applicable, the defenses should be effective under different data distributions, i.e., both IID and non-IID cases [34, 84].

– *Poisoned Model Rate:* The Pre-AD methods can be employed when the PMR is sufficiently large (i.e., up to 50%) because these methods aim at grouping the poisoned models into one group and the remaining group is benign. Other approaches, i.e., In-AD and Post-AD, are effective under smaller PMR such as less than 10%.

Defensive Requirements. Unlike ML, the orchestration server is not eligible to access the local training data, so the information that can be analyzed to defend against backdoor attacks is the local updates and their corresponding inference outputs.

– *Local Update Access:* Apart from [80, 95], other defenses need to analyze all local model updates. This leads to an issue with computation overhead. Instead of examining entire model parameters, efficient methods such as last-layer parameter analysis can be utilized to circumvent this issue [31].

– *Model inference:* To facilitate their defenses, [31, 86] need to consider inference results from local model updates. Although these strategies demonstrate efficiency in defending against backdoor attacks, this requires considerable computation costs. As a result, the remaining methods not requiring local model inferences are more relevant when the computation capacity of the central server is limited.

Effectiveness. Another issue with these defense methods is that they rely on too many assumptions about the data distribution, number of clients participating, and number of attackers. This makes it difficult to make a fair comparison between different approaches.

– *ASR:* The works [30, 31] can mitigate ASR from 100% to 0% with a little change in main task accuracy, but they rely on a strong assumption about the number of attackers to make a distinction for malicious models. For example, [30, 31, 80, 84] proposed defense methods that required a large percentage of malicious clients (up to 50%) to be present in order to effectively detect and exclude them. This highlights the importance of understanding the specific threat model and the distribution of malicious clients in a given scenario. Therefore, it is uncertain how well these methods will perform in a realistic world.

– *MTA Change:* One issue with existing defenses in FL is the degradation of performance on the primary task. For example, methods such as [27, 34, 93] result in a reduction of accuracy around 5%. This underlines the importance of ongoing research and development of strong defense mechanisms to guarantee accurate and trustworthy model results.

– *Application:* Most applications of backdoor attacks have been implemented in IC tasks [89, 80, 26, 27, 95, 86, 16, 93, 84, 22, 28, 29], although some have been observed in NLP tasks as well [16, 31, 57]. It is crucial for researchers and practitioners to remain vigilant in exploring the potential of backdoor attacks in various domains and to develop effective defense mechanisms to mitigate their impact.

4.5. Confrontation between Backdoor Attacks and Defenses

Adversaries and defenders are engaged in a never-ending battle. The conflict between them deepens our understanding of backdoor attacks. Attackers are always looking for ways to make poisoned attacks more covert, effective, and resistant to countermeasures. As shown in Table 5, most defense strategies focus on the scenarios of in-distribution backdoor attacks, in which the adversary simply changes the label of targeted inputs into his expected one. These poisoned samples can appear in the other benign participants' training data. Although defense is often designed against multiple attacks, many attack strategies have not been addressed such as noise-instance, coordinated-trigger, and partially poisoning backdoor attacks.

On the other hand, each countermeasure approach is often applied to a group of attack strategies. Particularly, pre-aggregation methods (e.g., Krum [33], FoolsGold [26], and AUROR [27])

Table 5: Backdoor Attacks Strategies and Defense Methodologies in FL

		Attack Strategies	Applicable Defenses
Data poisoning	Semantic backdoor	Out-of-distribution	DeepSight [31] , FLAME [57], Krum [33]
		In-distribution	FoolsGold [26], VAE [80], AUROR [27], Clustered FL [77] , PCA [71], BaFFle [86], FL-WBC [84]
		Noise-instance	N/A
	Artificial backdoor	Single-trigger	DP [16], RLR [34] , Matching Network [93], Pruning Neurons [29]
		Distributed-trigger	CRFL [56], FLAME [57], RLR [34], DeepSight [31], Prunning Neurons [29], FLDetector [89]
		Coordinated-trigger	N/A
Model poisoning	Fully poisoning	Constrain based	FLARE [92]
		Gradient-replacement based	CAE [95]
	Partially poisoning		N/A

seem to be efficient under semantic backdoor attacks. Furthermore, in–aggregation methods are primarily utilized under artificial backdoor attacks, specifically single-trigger attacks. However, the more sophisticated attack strategies such as distributed triggers and coordinated triggers, have not been evaluated under the presence of these defenses.

5. Challenges and Future Research Directions

In this section, we first pinpoint aspects for designing a more efficient and robust backdoor attack. Then, we discuss existing disadvantages and corresponding potential research directions for developing backdoor defenses from multi-perspectives. The summary of future research directions is presented in Figure 9.
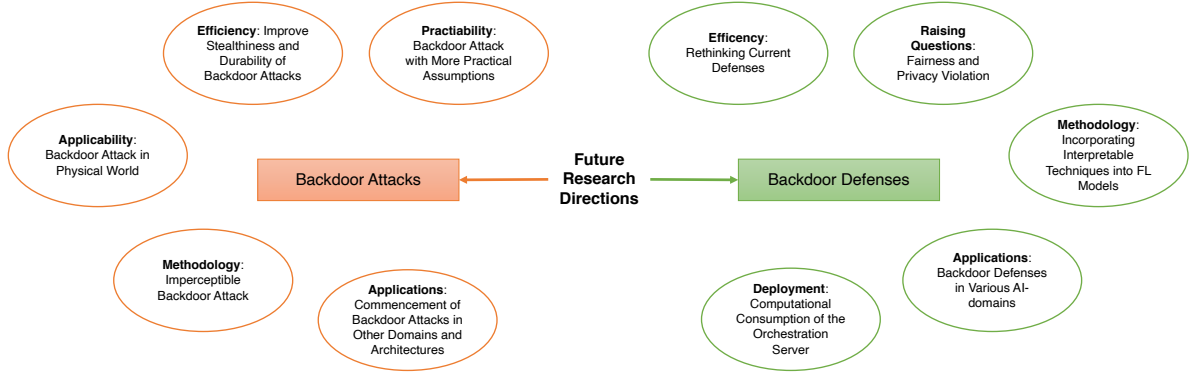

Figure 9: Summary of Future Research Directions.

5.1. Future Research Directions: Backdoor Attacks

Backdoor Attacks with More Practical Assumptions. Most of the existing backdoor attacks in FL rely on different assumptions, including assumptions about the percentages of compromised clients, the total number of FL clients, and the global distribution of training data. For instance, state-of-the-art attacks [17, 21] use benign samples drawn from global distribution to manipulate

the poisoning dataset. Other attacks [17, 26, 62] require continuous participation of compromised clients or a large ratio of malicious clients. This assumption is challenged by [73] and has shown to be unpractical. Therefore, it would be interesting to explore the possibility of designing attack strategies that require limited assumptions and can be applied in various scenarios, such as in large-scale FL systems with limited knowledge about system operations. To fulfill this purpose, the adversary can exploit leakage information via shared global model [62, 72, 99] to mimic an auxiliary training dataset align with global data distribution to strengthen the backdoor impact of limited-capability adversaries. Besides, when an adversary controls only a small fraction of participants (i.e., less than 0.1%), it can consider designing a single-shot attack and prolong the backdoor durability.

Stealthiness and Durability of Backdoor Attacks. Most current attacks have not considered stealth or enhanced the stealth by constraining poisoned model updates submitted to the aggregation server. Still, they have not taken the ocular stealth of the attack into account. In the early studies, the trigger is apparent, resulting in poor visual quality, and it can be easily removed by humans [100]. The stealth of the backdoor attacks in FL can be improved from two perspectives. Instead of inserting a small pattern into the original inputs, the trigger should be imperceptible to avoid inspection during the inference procedure. To do this, a learn-able trigger generated by optimizing objective functions [59, 101] or transformation models [62, 102] is visually indistinguishable from benign samples. From the model poisoning perspective, the naive scaling-based methods are not stealthy and robust against existing defenses [30, 56]. This issue can be addressed by partially poisoning attacks that leverage redundant space within a neural network architecture to covertly implant a backdoor, while still allowing the attacker to scale up the poisoned updates [19, 61]. In addition, the durability of the backdoor should be intensely considered to avoid the backdoor dilution phenomenon. A robust backdoor attack strategy should well balance stealthiness and durability.

Backdoor Attacks in Physical World. Current attack strategies typically use an artificial procedure to insert a trigger for a backdoor, such as a small pattern in images during training and testing. However, these attacks can be affected by the loss of the trigger, such as when a camera captures an image from a display or printed photo. The effectiveness of such attacks depends on the location and appearance of the trigger, as discussed in [103]. Therefore, it is important to evaluate current backdoor threats in physical FL systems. A hybrid attack that works with both digital and physical triggers may be a promising approach for implementing effective backdoor attacks in FL. One of the feasible methods is generating a backdoor dataset with the physical object as a trigger and applying physical transformations to enhance the robustness of the injected backdoor in real-world scenarios [103, 104].

Imperceptible Backdoor Attacks. The practice of inserting hidden information into images in a way that is imperceptible to the human eye for FL is known as steganography [105, 106, 107, 108]. This involves concealing a message, image, or file within another message, image, or file without affecting its visible appearance. In the context of FL, steganography could be used to insert data or metadata into images for training ML models while preserving the privacy of sensitive data or transferring it between organizations without revealing its content. Potential approaches to image steganography include applying transformations that preserve visual appearance while encoding additional information, generating adversarial examples with hidden data using machine

learning, and developing algorithms for detecting and decoding hidden information in images. The limits of what can be encoded in images while maintaining their visual quality should also be investigated.

Commencement of Backdoor Attacks in Other Domains and Architectures. Backdoor attacks in FL have been mostly studied for image classification [18, 22, 59, 64] and next word prediction [21] tasks. However, existing schemes may not be directly transferable to other domains due to differences in sample nature. Customized strategies may be needed to conduct backdoor attacks in specialized domains such as smart cities [109] or IoT intrusion systems [23]. Some applications of FL, such as environmental monitoring [110] and reducing network congestion [111], lack study on backdoor attacks and require further investigation. HFL is the most attractive land for implanting backdoor attacks since local datasets have the same feature space yet are different from each other and the adversary can easily manipulate the labels for his own training samples. Since VFL and FTL have experienced great development in the industry [112, 113, 114], the presence of a backdoor attack in these scenarios will cause significant concern.

5.2. Potential Research Directions on Defenses

Differential Privacy in FL. DP is a framework that protects the privacy of individuals in a dataset by adding noise to the data before it is released or used for analysis. It has been proposed for use in FL [16, 22] but has several limitations. DP requires a large number of clients to be effective, as the noise level needs to be high enough to mask the presence or absence of any individual client's data. It may also degrade model performance and may not prevent all types of privacy attacks, such as attribute inference and model inversion. Additionally, DP may not be suitable for all FL scenarios depending on the data being used and the client's privacy requirements. It is important to consider these limitations and trade-offs when using DP in FL settings.

Rethinking Current Defenses in FL: Limitations and Uncertainties. The current defenses in FL have limitations and uncertainties that must be addressed. Firstly, secure aggregation techniques [115], such as homomorphic encryption and secret sharing, are used in FL to combine model updates from multiple clients while preserving privacy. However, secure aggregation can also make FL systems vulnerable to poisoning attacks as individual updates cannot be inspected. Secondly, the effectiveness of adversarial training in non-IID settings remains uncertain, requiring further research. Finally, the field of FL, including VFL and FTL, is still in its early stages and requires further investigation to fully understand potential backdoor attacks and how to effectively defend against them. To mitigate these concerns, it is important to employ multiple layers of defense mechanisms and continuously monitor and audit the FL process to detect any malicious activity.

Backdoor Defenses in Various AI-domains. Backdoor attacks are generally easier to detect and defend against in the CV domain than in the NLP domain, according to empirical studies. For example, Wan et al.[116] found that ASR using the FedAvg algorithm was less than 75% effective with most defenses when one of ten clients was malicious in CV tasks. However, Yoo et al.[117] found that ASR was easily more than 95% effective on most attacks with most defenses when one of ten clients was malicious in NLP tasks. One reason for this difference may be that detecting NLP backdoors is more difficult. There is increasing interest in using FL in automatic speech recognition [118, 119, 120, 121], but the risk of backdoor attacks is a concern that needs to be

addressed. Future research may focus on developing effective strategies for defending against and detecting backdoor attacks in the automatic speech recognition domain.

Fairness and Privacy Violation. It is important that the application of a defense mechanism in an FL setting does not impact the fairness among the participating clients. For example, efforts to improve the robustness of FL systems may result in the unfair treatment of honest clients, as their updates may be rejected from the aggregation process if they lie far from the distribution of other updates, as discussed in [17]. This fact raises the question of the compensation between the fairness and robustness of FL systems in the presence of backdoor defenses. Additionally, some defense mechanisms rely on inspecting model updates to study the training data, which can increase the risk of membership inference and model inversion attacks [31, 57]. Therefore, it is important to carefully consider whether a specific defense mechanism is appropriate and to explore more secure defense strategies.

Incorporating Interpretable Techniques into FL Models. Interpretable techniques have been widely studied in the context of single-party ML models, such as decision trees, random forests, gradient-boosted trees, and deep neural networks [122, 123, 124, 125]. Most of these techniques have been developed to provide transparency into the decision-making processes of these models, with the goal of enhancing their interpretability and usability. However, their application to FL is relatively new. By providing transparency into the decision-making processes of FL models, interpretability techniques can help detect malicious clients and prevent backdoor attacks. For instance, studies show that saliency maps can reveal hidden triggers in single-party models and demonstrate the effectiveness of different defense methods against backdoors [126]. Similarly, visualization techniques can help to identify regions of the model's input space that are particularly susceptible to backdoor attacks and provide a way to test and validate the robustness of FL models.

Computational Consumption of the Orchestration Server. The deployment of defense mechanisms in FL requires significant computational resources, and it's crucial to ensure that it doesn't exceed the capacity of the orchestration server. Existing defense mechanisms often overlook the limitation of computational resources, leading to time delays and energy consumption. In future research, it's important to minimize resource consumption while deploying defense mechanisms in FL. For instance, for FL systems with a small number of clients, the local models can be verified one by one, but when the number of clients increases, this approach becomes impractical and consumes vast amounts of time and energy. An alternative solution is to deploy FL with multiple servers, distributing the task of verifying updates among them, which reduces resource consumption but brings new challenges such as communication costs and privacy leakage. Another promising solution is combining FL with blockchain technology, as proposed in [127], where clients upload updates to verifiers who select benign updates by voting and then aggregate and write the selected updates to blocks through the blockchain network.

5.3. Discussion

Exploring the Practical Benefits of Backdoor Attacks in Federated Unlearning. We often consider backdoor attacks as a great threat in FL while ignoring its potential advantages. Indeed, a backdoor attack demonstrated its sake under the unlearning scenario, which is a technique in FL [28, 128, 129] focusing on removing or revoking access to data, participants, or parts of the

model, with the goal of improving the integrity and accuracy of the model. Backdoor triggers are utilized as an evaluation tool to assess the effectiveness of unlearning methods [130]. The client, who wants to opt out of the federation, uses a dataset that contains a fraction of samples with inserted backdoor triggers, making the global FL model vulnerable to the backdoor trigger. The goal of the unlearning process is to produce a model that decreases accuracy on samples with backdoor triggers while preserving good performance on clean samples. Future research on unlearning needs to focus more on investigating the impact of backdoor attack methods on model privacy and security.

Investigating Various Backdoor Injection Strategies in Multi-Group FL. Existing works often consider the homogeneous backdoor attack, in which the malicious participants have a common attack objective [17, 18, 21, 31]. This approach is not always relevant in the physical world. This raises a great concern about backdoor effects caused by multiple adversaries with different backdoor targets. For example, consider a model that recognizes two-digit numbers. It is possible to inject two new backdoor tasks into the model: one that sums up the digits and another one that multiplies them. Then, various endeavors from distinct backdoor tasks can be complementary or detrimental to one another. Moreover, the appearance of multiple backdoor tasks may have varying effects on the performance of the FL model, and it is important for future research to uncover these effects in order to improve the security and privacy of FL models. This research can aid in the development of better methods for detecting and mitigating backdoor attacks in FL, thus improving the overall integrity and robustness of the model.

Integrating Multiple Defense Mechanisms in FL. Previous works [30, 31] used a combination of methods in the Pre-AD and In-AD phases to mitigate backdoor attacks. These methods involve two layers: the first layer detects and excludes models that contain a well-trained backdoor, while the second layer uses a different approach in the In-AD phase to mitigate the attack. In future research, a combination of methods from different defense phases, such as Pre-AD and Post-AD, In-AD and Post-AD, or Pre-AD, In-AD, and Post-AD, can be studied to further improve the defense against backdoor attacks in FL.

6. Conclusion

In summary, backdoor attacks in FL pose a significant threat to the security and privacy of FL systems. These attacks can be triggered in various ways, including artificial and semantic triggers, and can be launched by a single client or a group of clients. To defend against these attacks, various approaches have been proposed, including pre-aggregation defenses, in-aggregation defenses, and post-aggregation defenses. Each of these approaches has its own advantages and limitations, and their effectiveness depends on the specific characteristics of the attack. Moreover, the robustness of these defenses in the face of various types of attacks, particularly in non-IID scenarios, remains an open research question. In the future, it will be important to continue developing more robust defense techniques that are effective against semantic backdoor attacks, improving the efficiency of defense techniques, studying the effectiveness of defenses under realistic attack scenarios, examining the impact of data heterogeneity on backdoor attacks and defenses, and investigating the impact of system-level factors on backdoor attacks and defenses. By addressing these research areas, it will be possible to make progress in understanding and addressing the risks of backdoor

attacks in federated learning systems and to develop more secure and effective defense strategies against a wide range of attacks. It is also important to consider the potential for physical backdoor attacks and to explore potential defenses against these types of attacks. In addition, research on the effectiveness of backdoor defenses in specific AI domains, such as automatic speech recognition, could be valuable in developing targeted and effective protection mechanisms.

CrediT Authorship Contribution Statement

Thuy Dung Nguyen: Methodology, Visualization, Writing - Original Draft, Writing - Review & Editing, **Minh Tuan Nguyen:** Methodology, Visualization, Writing - Original Draft, Writing - Review & Editing, **Phi Le Nguyen:** Writing - Review & Editing, **Huy Hieu Pham:** Writing - Review & Editing, **Khoa Doan:** Writing - Review & Editing, **Kok-Seng Wong:** Conceptualization, Project administration, Supervision, Writing - Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

References

[1] E. Blasch, T. Pham, C.-Y. Chong, W. Koch, H. Leung, D. Braines, T. Abdelzaher, Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges, IEEE Aerospace and Electronic Systems Magazine 36 (7) (2021) 80–93.

[2] H. B. McMahan, E. Moore, D. Ramage, B. A. y Arcas, Federated learning of deep networks using model averaging, ArXiv abs/1602.05629.

[3] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecný, S. Mazzocchi, H. B. McMahan, T. V. Overveldt, D. Petrou, D. Ramage, J. Roselander, Towards federated learning at scale: System design, ArXiv abs/1902.01046.

[4] K. Doshi, Y. Yılmaz, Federated learning-based driver activity recognition for edge devices, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022) 3337–3345.

[5] D. Becking, H. Kirchhoffer, G. Tech, P. Haase, K. Müller, H. Schwarz, W. Samek, Adaptive differential filters for fast and communication-efficient federated learning, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2022) 3366–3375.

[6] M. Chen, A. T. Suresh, R. Mathews, A. Wong, C. Allauzen, F. Beaufays, M. Riley, Federated learning of n-gram language models, arXiv preprint arXiv:1910.03432.

[7] B. Y. Lin, C. He, Z. Zeng, H. Wang, Y. Huang, M. Soltanolkotabi, X. Ren, S. Avestimehr, Fednlp: A research platform for federated learning in natural language processing, ArXiv abs/2104.08815.

[8] J. Xu, F. Wang, Federated learning for healthcare informatics, Journal of Healthcare Informatics Research 5 (2021) 1 – 19.

[9] Prayitno, C. R. Shyu, K. T. Putra, H.-C. Chen, Y.-Y. Tsai, K. S. M. T. Hossain, W. Jiang, Z. Shae, A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications, Applied Sciences.

[10] D. Gupta, O. Kayode, S. Bhatt, M. Gupta, A. S. Tosun, Hierarchical federated learning based anomaly detection using digital twins for smart healthcare, 2021 IEEE 7th International Conference on Collaboration and Internet Computing (CIC) (2021) 16–25.

[11] Y. Liu, R. Yang, Federated learning application on depression treatment robots(dtbot), 2021 IEEE 13th International Conference on Computer Research and Development (ICCRD) (2021) 121–124.

[12] Y. Liu, S. Garg, J. Nie, Y. Zhang, Z. Xiong, J. Kang, M. S. Hossain, Deep anomaly detection for time-series data in industrial iot: A communication-efficient on-device federated learning approach, IEEE Internet of Things Journal 8 (2021) 6348–6358.

[13] I. I. Kholod, E. Yanaki, D. Fomichev, E. Shalugin, E. Novikova, E. Filippov, M. Nordlund, Open-source federated learning frameworks for iot: A comparative review and analysis, Sensors (Basel, Switzerland) 21.

[14] V.-D. Nguyen, S. K. Sharma, T. X. Vu, S. Chatzinotas, B. E. Ottersten, Efficient federated learning algorithm for resource allocation in wireless iot networks, IEEE Internet of Things Journal 8 (2021) 3394–3409.

[15] L. Lyu, H. Yu, Q. Yang, Threats to federated learning: A survey, ArXiv abs/2003.02133.

[16] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2938–2948.

[17] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, D. Papailiopoulos, Attack of the tails: Yes, you really can backdoor federated learning, Advances in Neural Information Processing Systems 33 (2020) 16070–16084.

[18] C. Xie, K. Huang, P.-Y. Chen, B. Li, Dba: Distributed backdoor attacks against federated learning, in: International conference on learning representations, 2020.

[19] Z. Zhang, A. Panda, L. Song, Y. Yang, M. Mahoney, P. Mittal, R. Kannan, J. Gonzalez, Neurotoxin: Durable backdoors in federated learning, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, Vol. 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 26429–26446.
URL https://proceedings.mlr.press/v162/zhang22w.html

[20] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, B. Y. Zhao, Backdoor attacks against deep learning systems in the physical world, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 6202–6211.

[21] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, CoRR abs/1807.00459. arXiv:1807.00459.
URL http://arxiv.org/abs/1807.00459

[22] Z. Sun, P. Kairouz, A. T. Suresh, H. B. McMahan, Can you really backdoor federated learning?, ArXiv abs/1911.07963.

[23] T. D. Nguyen, P. Rieger, M. Miettinen, A.-R. Sadeghi, Poisoning attacks on federated learning-based iot intrusion detection system, in: Proc. Workshop Decentralized IoT Syst. Secur.(DISS), 2020, pp. 1–7.

[24] R. Jin, X. Li, Backdoor attack and defense in federated generative adversarial network-based medical image synthesis, arXiv preprint arXiv:2210.10886.

[25] Y. Feng, B. Ma, J. Zhang, S. Zhao, Y. Xia, D. Tao, Fiba: Frequency-injection based backdoor attack in medical image analysis, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20844–20853. doi:10.1109/CVPR52688.2022.02021.

[26] C. Fung, C. J. Yoon, I. Beschastnikh, Mitigating sybils in federated learning poisoning, arXiv preprint arXiv:1808.04866.

[27] S. Shen, S. Tople, P. Saxena, Auror: Defending against poisoning attacks in collaborative deep learning systems, in: Proceedings of the 32nd Annual Conference on Computer Security Applications, 2016, pp. 508–519.

[28] C. Wu, S. Zhu, P. Mitra, Federated unlearning with knowledge distillation, arXiv preprint arXiv:2201.09441.

[29] C. Wu, X. Yang, S. Zhu, P. Mitra, Mitigating backdoor attacks in federated learning, ArXiv abs/2011.01767.

[30] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, Others, {FLAME}: Taming backdoors in federated learning, in: 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1415–1432.

[31] P. Rieger, T. D. Nguyen, M. Miettinen, A.-R. Sadeghi, Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection, arXiv preprint arXiv:2201.00763.

29

[32] K. Pillutla, S. M. Kakade, Z. Harchaoui, Robust aggregation for federated learning, IEEE Transactions on Signal Processing 70 (2022) 1142–1154.

[33] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, J. Stainer, Machine learning with adversaries: Byzantine tolerant gradient descent, Advances in Neural Information Processing Systems 30.

[34] M. S. Ozdayi, M. Kantarcioglu, Y. R. Gel, Defending against backdoors in federated learning with robust learning rate, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 9268–9276.

[35] A. N. Bhagoji, S. Chakraborty, P. Mittal, S. Calo, Analyzing federated learning through an adversarial lens, in: International Conference on Machine Learning, PMLR, 2019, pp. 634–643.

[36] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, T. Goldstein, Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses, IEEE Trans. Pattern Anal. Mach. Intell. PP.

[37] Y. Li, B. Wu, Y. Jiang, Z. Li, S. Xia, Backdoor learning: A survey, IEEE transactions on neural networks and learning systems PP.

[38] X. Gong, Y. Chen, Q. Wang, W. Kong, Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions, IEEE Wireless Communications (2022) 1–7 doi:10.1109/MWC.017.2100714.

[39] Z. Tian, L. Cui, J. Liang, S. Yu, A comprehensive survey on poisoning attacks and countermeasures in machine learning, ACM Computing Surveys (CSUR).

[40] S. Abdulrahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, M. Guizani, A survey on federated learning: The journey from centralized to distributed on-site learning and beyond, IEEE Internet of Things Journal 8 (2021) 5476–5497.

[41] X. Yin, Y. Zhu, J. Hu, A comprehensive survey of privacy-preserving federated learning, ACM Computing Surveys (CSUR) 54 (2021) 1 – 36.

[42] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. ping Huang, A. Dehghantanha, G. Srivastava, A survey on security and privacy of federated learning, Future Gener. Comput. Syst. 115 (2021) 619–640.

[43] R. Gosselin, L. Vieu, F. Loukil, A. Benoit, Privacy and security in federated learning: A survey, Applied Sciences.

[44] X. Chen, C. Liu, B. Li, K. Lu, D. X. Song, Targeted backdoor attacks on deep learning systems using data poisoning, ArXiv abs/1712.05526.

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

[46] T. D. Nguyen, P. Rieger, M. Miettinen, A.-R. Sadeghi, Poisoning attacks on federated learning-based iot intrusion detection system, in: Proc. Workshop Decentralized IoT Syst. Secur.(DISS), 2020, pp. 1–7.

[47] L. Muñoz-González, K. T. Co, E. C. Lupu, Byzantine-robust federated machine learning through adaptive model averaging, arXiv preprint arXiv:1909.05125.

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

[49] D. Yin, Y. Chen, R. Kannan, P. Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 5650–5659.

[50] Q. Li, Z. Wen, B. He, Practical federated gradient boosting decision trees, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 4642–4649.

[51] V. Smith, C.-K. Chiang, M. Sanjabi, A. S. Talwalkar, Federated multi-task learning, Advances in neural information processing systems 30.

[52] Y. Wu, S. Cai, X. Xiao, G. Chen, B. C. Ooi, Privacy preserving vertical federated learning for tree-based models, arXiv preprint arXiv:2008.06170.

[53] S. Yang, B. Ren, X. Zhou, L. Liu, Parallel distributed logistic regression for vertical federated learning without third-party coordinator, arXiv preprint arXiv:1911.09824.

[54] Y. Chen, X. Qin, J. Wang, C. Yu, W. Gao, Fedhealth: A federated transfer learning framework for wearable healthcare, IEEE Intelligent Systems 35 (4) (2020) 83–93.

[55] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, Badnets: Evaluating backdooring attacks on deep neural networks, IEEE Access 7 (2019) 47230–47244.

[56] C. Xie, M. Chen, P.-Y. Chen, B. Li, Crfl: Certifiably robust federated learning against backdoor attacks, in: International Conference on Machine Learning, PMLR, 2021, pp. 11372–11382.

[57] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni, et al., {FLAME}: Taming backdoors in federated learning, in: 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1415–1432.

[58] K. Yoo, N. J. Kwak, Backdoor attacks in federated learning by rare embeddings and gradient ensembling, ArXiv abs/2204.14017.

[59] X. Gong, Y. Chen, H. Huang, Y. Liao, S. Wang, Q. Wang, Coordinated backdoor attacks against federated learning with model-dependent triggers, IEEE Network 36 (2022) 84–90.

[60] T. Zou, Y. Liu, Y. Kang, W. Liu, Y. He, Z. Yi, Q. Yang, Y.-Q. Zhang, Defending batch-level label inference and replacement attacks in vertical federated learning, IEEE Transactions on Big Data (2022) 1–12doi:10.1109/TBDATA.2022.3192121.

[61] X.-L. Zhou, M. Xu, Y. Wu, N. Zheng, Deep model poisoning attack on federated learning, Future Internet 13 (2021) 73.

[62] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, S. Yu, Poisongan: Generative poisoning attacks against federated learning in edge computing systems, IEEE Internet of Things Journal 8 (5) (2021) 3310–3322. doi:10.1109/JIOT.2020.3023126.

[63] Y. Liu, Z. qian Yi, T. Chen, Backdoor attacks and defenses in feature-partitioned collaborative learning, ArXiv abs/2007.03608.

[64] G. Baruch, M. Baruch, Y. Goldberg, A little is enough: Circumventing defenses for distributed learning, Advances in Neural Information Processing Systems 32.

[65] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, Badnets: Evaluating backdooring attacks on deep neural networks, IEEE Access 7 (2019) 47230–47244.

[66] Y. Liu, T. Zou, Y. Kang, W. Liu, Y. He, Z. Yi, Q. Yang, Batch label inference and replacement attacks in black-boxed vertical federated learning, arXiv e-prints (2021) arXiv–2112.

[67] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Foundations and Trends® in Machine Learning 14 (1–2) (2021) 1–210.

[68] N. Rodríguez-Barroso, D. Jiménez López, M. Victoria Luzón, F. Herrera, E. Martínez-Cámara, Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges, Information Fusiondoi:https://doi.org/10.1016/j.inffus.2022.09.011.
URL https://www.sciencedirect.com/science/article/pii/S1566253522001439

[69] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009).

[70] Y. Sun, H. Ochiai, J. Sakuma, Semi-targeted model poisoning attack on federated learning via backward error analysis, ArXiv abs/2203.11633.

[71] V. Tolpegin, S. Truex, M. E. Gursoy, L. Liu, Data poisoning attacks against federated learning systems, in: European Symposium on Research in Computer Security, Springer, 2020, pp. 480–501.

[72] T. Liu, X. Hu, T. Shu, Technical report: Assisting backdoor federated learning with whole population knowledge alignment, ArXiv abs/2207.12327.

[73] V. Shejwalkar, A. Houmansadr, P. Kairouz, D. Ramage, Back to the drawing board: A critical evaluation of poisoning attacks on federated learning, ArXiv abs/2108.10241.

[74] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2). doi:10.1145/3298981.
URL https://doi.org/10.1145/3298981

[75] K. Pillutla, S. M. Kakade, Z. Harchaoui, Robust aggregation for federated learning, IEEE Transactions on Signal Processing 70 (2022) 1142–1154. doi:10.1109/TSP.2022.3153135.

[76] D. Cao, S. Chang, Z. Lin, G. Liu, D. Sun, Understanding distributed poisoning attack in federated learning, in: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), IEEE, 2019, pp.

233–239.

[77] F. Sattler, K.-R. Müller, T. Wiegand, W. Samek, On the byzantine robustness of clustered federated learning, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 8861–8865.

[78] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, E. Ilie-Zudor, Chained anomaly detection models for federated learning: An intrusion detection case study, Applied Sciences 8 (12) (2018) 2663.

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

[80] S. Li, Y. Cheng, W. Wang, Y. Liu, T. Chen, Learning to detect malicious clients for robust federated learning, arXiv preprint arXiv:2002.00211.

[81] D. Yin, Y. Chen, R. Kannan, P. Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, in: International Conference on Machine Learning, PMLR, 2018, pp. 5650–5659.

[82] R. Guerraoui, S. Rouault, et al., The hidden vulnerability of distributed learning in byzantium, in: International Conference on Machine Learning, PMLR, 2018, pp. 3521–3530.

[83] J. Bernstein, J. Zhao, K. Azizzadenesheli, A. Anandkumar, signsgd with majority vote is communication efficient and fault tolerant, arXiv preprint arXiv:1810.05291.

[84] J. Sun, A. Li, L. DiValentin, A. Hassanzadeh, Y. Chen, H. Li, Fl-wbc: Enhancing robustness against model poisoning attacks in federated learning from a client perspective, Advances in Neural Information Processing Systems 34 (2021) 12613–12624.

[85] M. Naseri, J. Hayes, E. De Cristofaro, Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy, arXiv e-prints (2020) arXiv–2009.

[86] S. Andreina, G. A. Marson, H. Möllering, G. Karame, Baffle: Backdoor detection via feedback-based federated learning, in: 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS), IEEE, 2021, pp. 852–863.

[87] P. C. Mahalanobis, On the generalised distance in statistics, in: Proceedings of the national Institute of Science of India, Vol. 12, 1936, pp. 49–55.

[88] A. Singhal, et al., Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (4) (2001) 35–43.

[89] Z. Zhang, X. Cao, J. Jia, N. Z. Gong, Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 2545–2555.

[90] E. Bagdasaryan, O. Poursaeed, V. Shmatikov, Differential privacy has disparate impact on model accuracy, Advances in neural information processing systems 32.

[91] H. B. McMahan, D. Ramage, K. Talwar, L. Zhang, Learning differentially private recurrent language models, arXiv preprint arXiv:1710.06963.

[92] N. Wang, Y. Xiao, Y. Chen, Y. Hu, W. Lou, Y. T. Hou, Flare: defending federated learning against model poisoning attacks via latent space representations, in: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security, 2022, pp. 946–958.

[93] C.-L. Chen, L. Golubchik, M. Paolieri, Backdoor attacks on federated meta-learning, arXiv preprint arXiv:2006.07026.

[94] J. Zhang, D. Wu, C. Liu, B. Chen, Defending poisoning attacks in federated learning via adversarial training method, in: International Conference on Frontiers in Cyber Security, Springer, 2020, pp. 83–94.

[95] T. Zou, Y. Liu, Y. Kang, W. Liu, Y. He, Z. Yi, Q. Yang, Y.-Q. Zhang, Defending batch-level label inference and replacement attacks in vertical federated learning, IEEE Transactions on Big Data.

[96] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, 2021 IEEE Symposium on Security and Privacy (SP) (2021) 141–159.

[97] S. Neel, A. Roth, S. Sharifi-Malvajerdi, Descent-to-delete: Gradient-based methods for machine unlearning, in: Algorithmic Learning Theory, PMLR, 2021, pp. 931–962.

[98] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, C. Waites, Adaptive machine unlearning, Advances in Neural Information Processing Systems 34 (2021) 16319–16330.

[99] X. Xu, J. Wu, M. Yang, T. Luo, X. Duan, W. Li, Y. Wu, B. Wu, Information leakage by model weights on

federated learning, Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice.

[100] K. Doan, Y. Lao, P. Li, Backdoor attack with imperceptible input and latent modification, Advances in Neural Information Processing Systems 34 (2021) 18944–18957.

[101] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, Poison frogs! targeted clean-label poisoning attacks on neural networks, ArXiv abs/1804.00792.

[102] A. Nguyen, A. Tran, Wanet - imperceptible warping-based backdoor attack, ArXiv abs/2102.10369.

[103] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, B. Y. Zhao, Backdoor attacks against deep learning systems in the physical world, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 6202–6211.

[104] M. Xue, C. He, S. Sun, J. Wang, W. Liu, Robust backdoor attacks against deep neural networks in real physical world, 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (2021) 620–626.

[105] J. Hayes, G. Danezis, Generating steganographic images via adversarial training, Advances in neural information processing systems 30.

[106] S. Baluja, Hiding images in plain sight: Deep steganography, Advances in neural information processing systems 30.

[107] K. Doan, Y. Lao, W. Zhao, P. Li, Lira: Learnable, imperceptible and robust backdoor attacks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11966–11976.

[108] J. Jing, X. Deng, M. Xu, J. Wang, Z. Guan, Hinet: deep image hiding by invertible network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4733–4742.

[109] Z. Zheng, Y. Zhou, Y. Sun, Z. Wang, B. Liu, K. Li, Federated learning in smart cities: A comprehensive survey, ArXiv abs/2102.01375.

[110] B. Hu, Y. Gao, L. Liu, H. Ma, Federated region-learning: An edge computing based framework for urban environment sensing, 2018 IEEE Global Communications Conference (GLOBECOM) (2018) 1–7.

[111] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, M. Chen, In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning, IEEE Network 33 (2019) 156–165.

[112] Q. Jing, W. Wang, J. Zhang, H. Tian, K. Chen, Quantifying the performance of federated transfer learning, ArXiv abs/1912.12795.

[113] Y. Chen, X. yan Sun, Y. Jin, Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation, IEEE Transactions on Neural Networks and Learning Systems 31 (2020) 4229–4238.

[114] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen, T. Ranbaduge, Vertical federated learning: Challenges, methodologies and experiments, ArXiv abs/2202.04309.

[115] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for privacy-preserving machine learning, in: proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1175–1191.

[116] C. P. Wan, Q. Chen, Robust federated learning with attack-adaptive aggregation, arXiv preprint arXiv:2102.05257.

[117] K. Yoo, N. Kwak, Backdoor attacks in federated learning by rare embeddings and gradient ensembling, arXiv preprint arXiv:2204.14017.

[118] X. Cui, S. Lu, B. Kingsbury, Federated acoustic modeling for automatic speech recognition, in: ICASSP, 2021, pp. 6748–6752.

[119] D. Dimitriadis, K. Kumatani, R. Gmyr, et al., A federated approach in training acoustic models, in: Interspeech, 2020, pp. 981–985.

[120] S. Mdhaffar, M. Tommasi, Y. Esteve, Study on acoustic model personalization in a context of collaborative learning constrained by privacy preservation, in: Speech and Computer, 2021, pp. 426–436.

[121] D. Guliani, F. Beaufays, G. Motta, Training speech recognition models with federated learning: A quality/cost framework, in: ICASSP, 2021, pp. 3080–3084.

[122] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[123] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (5) (2019) 206–215.

[124] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digital signal processing 73 (2018) 1–15.

[125] W. Samek, K.-R. Müller, Towards explainable artificial intelligence, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 5–22.

[126] S. Fang, A. Choromanska, Backdoor attacks on the dnn interpretation system, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 561–570.

[127] H. Chen, S. A. Asif, J. Park, C.-C. Shen, M. Bennis, Robust blockchained federated learning with model validation and proof-of-stake inspired consensus, arXiv preprint arXiv:2101.03300.

[128] G. Liu, X. Ma, Y. Yang, C. Wang, J. Liu, Federaser: Enabling efficient client-level data removal from federated learning models, in: 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), IEEE, 2021, pp. 1–10.

[129] J. Wang, S. Guo, X. Xie, H. Qi, Federated unlearning via class-discriminative pruning, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 622–632.

[130] A. Halimi, S. Kadhe, A. Rawat, N. Baracaldo, Federated unlearning: How to efficiently erase a client in fl?, arXiv preprint arXiv:2207.05521.