Data and Model Poisoning Backdoor Attacks on Wireless Federated Learning, and the Defense Mechanisms: A Comprehensive Survey

Yichen Wan*, Youyang Qu*, *Member, IEEE,* Wei Ni, *Fellow, IEEE,* Yong Xiang, *Senior Member, IEEE,* Longxiang Gao†, *Senior Member, IEEE,*, and Ekram Hossain, *Fellow, IEEE*

***Abstract*—Due to the greatly improved capabilities of devices, massive data, and increasing concern about data privacy, Federated Learning (FL) has been increasingly considered for applications to wireless communication networks (WCNs). Wireless FL (WFL) is a distributed method of training a global deep learning model in which a large number of participants each train a local model on their training datasets and then upload the local model updates to a central server. However, in general, non-independent and identically distributed (non-IID) data of WCNs raises concerns about robustness, as a malicious participant could potentially inject a "backdoor" into the global model by uploading poisoned data or models over WCN. This could cause the model to misclassify malicious inputs as a specific target class while behaving normally with benign inputs. This survey provides a comprehensive review of the latest backdoor attacks and defense mechanisms. It classifies them according to their targets (data poisoning or model poisoning), the attack phase (local data collection, training, or aggregation), and defense stage (local training, before aggregation, during aggregation, or after aggregation). The strengths and limitations of existing attack strategies and defense mechanisms are analyzed in detail. Comparisons of existing attack methods and defense designs are carried out, pointing to noteworthy findings, open challenges, and potential future research directions related to security and privacy of WFL.**

***Index Terms*—Federated Learning, Backdoor Attacks, Security, Privacy, Wireless Communications**

I. Introduction

WIRELESS communication network (WCN) is recognized as part of critical infrastructures and has been deployed ubiquitously in recent years [1]. With the persistent development of wireless communication technologies, the network capacity has massively increased, and more devices can be connected to the network [2]. Wireless communications services are used in different scenarios for various Artificial Intelligence (AI)-based applications [3], [4], such as image classification task [5], [6], text prediction [7], and unmanned aerial vehicle (UAV) control [8], [9]. With the widespread use of AI applications and user/sensor data involved during the AI training process, the security and privacy issues have become critical [10]. In traditional AI algorithm training, participants must send their raw data to a central server for processing. However, WCN introduces the risk of unauthorized data access during the transmission, which might compromise the privacy, security, and robustness of the systems [11], [12]. Many efforts have been made to develop a robust and secure framework, while Federated Learning (FL) or Wireless FL (WFL) is considered one of the most practical solutions [13]–[17].

First proposed by Google [18], FL is a machine learning paradigm in which multiple devices form a distributed network while maintaining the same machine learning model. With a large number of clients participating via wireless networks [19], the WFL approach allows for developing a robust model without requiring raw data sharing within the network. Different from traditional machine learning methods that typically collect all raw data from participants and directly train the global model in the central server, the clients of WFL are the owners of their data and participate in the training process by computing model updates based on their local private data. These updates are then aggregated by a central server and used to fine-tune a global model, which is then distributed to the clients for the next round of local training [20]. Consequently, each client can better control their private information. Compared with the traditional machine learning approaches, the need for data transfer in WFL is greatly reduced, the computational efficiency is significantly improved, and the substantial bandwidth provided by the WCN can be better utilized. In recent years, WFL has seen widespread adoption in various fields, including mobile device development, industrial engineering, and healthcare [21]–[24].

Despite its diverse merits, the use of WFL has also raised concerns about the robustness of the training processes [25]. The training datasets of WFL participants are non-independent and identically distributed (non-IID). Each client's data may be different in terms of its statistical properties, data class distribution, or data quality. The heterogeneous and unbalanced-distributed nature of data imposes challenges to train a converging global model that performs well across all

Y. Wan and Y. Qu are the first authors to contribute equally to this work.
Y. Wan and Y. Xiang are with the School of IT, Deakin University, VIC 3125, Australia (email: {wanyich, yong.xiang}@deakin.edu.au)
W. Ni is with the Data61, CSIRO, NSW 2122, Australia (email: wei.ni@data61.csiro.au)
L. Gao is the corresponding author. Y. Qu and L. Gao are with 1) 1. Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China; and 2) Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan, China (email: youyang.qu@data61.csiro.au, gaolx@sdas.org).
E. Hossain is with the Department of Electrical and Computer Engineering at the University of Manitoba, Canada (email: ekram.hossain@umanitoba.ca).

clients [26]. Existing methods for the detection and mitigation of data/model poisoning attacks can be less effective due to the difficulty in distinguishing the disformity of local models resulting from the non-IID datasets from that resulting from attacks [27]. Furthermore, it makes it easier for malicious clients to intentionally send anomaly updates to bias the global model in their favor [28]. Adversarial attacks on WFL can be broadly classified into four categories: poisoning [29], inference attacks [30], Byzantine attacks [31], and backdoor attacks [32].

In this survey, we focus specifically on backdoor attacks, as they are on the rise, can potentially affect many clients, and may go undetected after multiple rounds of aggregations in WFL. Backdoor attacks are a particularly insidious form of adversarial attack on WFL, as they allow malicious users to manipulate the classification results of the model by altering the local training datasets of participating clients or using malicious model replacements [33]. Backdoor attacks can be divided into two categories: data-based attacks and model-based attacks. In data-based backdoor attacks, the attacker injects a backdoor into the training data. In contrast, model-based attacks target the training process of the local model before aggregation.

Many organizations worldwide are dealing with the threat of backdoor attacks, and the threat is ever-growing unexpectedly. A critical challenge faced by detecting and defending against backdoor attacks is that backdoor attacks often do not negatively impact the accuracy of the model on primary tasks and do not disrupt the normal system operation before they are activated [32]. This makes it difficult to detect the presence of a backdoor.

For example, in August of 2023, Canon, the famous manufacturer of optical and imaging products, warned its users of inkjet printers that their WiFi connection settings stored in the devices were at risk of leakage. The privacy information, including the network SSID, the password, network type (WPA3, WEP, etc.), assigned IP address, MAC address, and network profile stored in the devices' memories are not wiped during initialization and could be easily extracted. Such flaws could be exploited by malicious users to create backdoors, allowing unauthorized access to a Canon printer user's network that the printer was connected to [34]. From there on, the attacker can access shared resources, steal data, hijack other devices connected to the same network, and launch other attacks to leverage additional vulnerabilities. To be more specific, the Google spam filter has been poisoned by sending malicious emails containing malware or other cybersecurity threats without being detected by the algorithm[1]. Besides, researchers have found new ways to poison cutting-edge large generative AL models (e.g., ChatGPT) through prompt engineering[2].

While there have been no widely reported data and model poisoning backdoor attacks specifically targeting WFL, designs of such attacks have started to emerge in the literature, e.g., [35]. Particularly, the broadcast nature of wireless (radio) channels allows misbehaved or compromised participants in WFL to overhear the local model parameters of multiple or all benign participants. By taking advantage of the overheard local models, the misbehaved participants can construct malicious local models to escape malicious model detection at the central server, poison the global model and subsequently the benign local models, and manipulate WFL [35]. It is crucial to detect and suppress such attacks on WFL due to the increasing interest in WFL deployment and the exacerbated impact of the attacks over wireless interfaces.

Defending against data-based and model-based backdoor attacks on WFL remains an open issue. It requires the protection of both the local and global models against all potential attacks. Many works have been proposed to address this challenge. For example, some researchers have proposed using differential privacy to defend against poisoned data samples in the training process or adding a threshold value to gradients to block abnormal gradients classified as malicious attackers [36]–[38]. Others have proposed methods such as a backdoor filter [39] or a White Blood Cell for WFL (WFL-WBC) [40] to defend against model-poisoning attacks. However, these methods only show reasonable performance against specific types of attacks. There is still much work to be done to address this common problem.

A. Motivation

As discussed above, backdoor attacks can occur at any phase during the convergence of the global model, including data collection, local model training, and global model aggregation [41], [42]. This makes them more threatening and more difficult to defend against, as the defender needs to consider all steps in the process. In contrast to Byzantine attacks that aim to degrade the performance of the main WFL task [43], backdoor attacks are characterized by their stealthiness. The primary goal of these attacks is to leave a backdoor in the converged global model that can mislead the model into classifying the backdoor input into the target class while behaving normally for benign inputs. This property allows the backdoor to remain effective for a longer duration, making it difficult to eliminate fully. Therefore, it is important to delve deeper into the nature of backdoor attacks to inform the design of defense algorithms and improve the robustness and security of the WFL framework.

Despite the rapid growth of WFL as a research area, there has not yet been a comprehensive and up-to-date review of it from the perspective of backdoor attacks. In this paper, we aim to provide a systematic summary and classification of existing backdoor attacks and defenses, highlighting their differences and connections, as well as discussing their respective limitations. We also discuss future research directions in this area.

B. Review of Existing Surveys and Gap Analysis

The existing surveys on backdoor attacks and defenses under WFL frameworks (from 2020 to 2023) are reviewed. The relevant papers are collected from multiple academic databases, including IEEE Xplore, Elsevier, MDPI, and arXiv. The comparison of this survey with the existing literature is

[1]https://www.forcepoint.com/blog/security-labs/data-poisoning-newest-threat-generative-ai

[2]https://harshitrao.medium.com/is-it-easy-to-poison-chat-gpt-d81287f1b58b

TABLE I: A summary of existing surveys related to WFL backdoor attacks and defense methods

Survey paper	Year	Consideration of WCN	Visualization for each method	Discussion on lessons learned	Discussion on backdoor attack methods							Discussion on backdoor defense methods			
					Comparison of attack methods	Mathematical analysis for each method	Required prior information analysis	Attack applicability	Evaluation metric	Limitation analysis for each method	Future directions	Comparison of defense methods	Defense applicability	Limitation analysis for each method	Future direction
[54]	2020				√										√
[48]	2020				√						√	√		√	√
[45]	2021				√				√			√			√
[52]	2022				√				√		√	√			√
[49]	2022				√				√		√	√			√
[50]	2022				√	√		√	√	√	√	√			√
[46]	2022										√	√			√
[47]	2023		√		√			√				√			
[53]	2023				√			√	√		√	√	√		√
[51]	2023				√			√		√	√				√
Ours		√	√	√	√	√	√	√	√	√	√	√	√	√	√

summarized in Table I. It can be seen that some existing surveys, e.g., [44], [45], [46], and [47] considered backdoor attacks and backdoor defenses as a part of robustness threat on WFL, however, the limitations of the existing backdoor attack and defense methods were not highlighted. On the other hand, in [48], [49], [50], and [51], WFL was considered as one of the deep learning applications when discussing the impact of backdoor attacks, but no detailed analysis of vulnerabilities of backdoor attacks on WFL was provided. In [52] and [53], the theoretical working mechanisms of backdoor attacks and defenses for WFL were reviewed. However, the limitations of the existing methodologies were not systematically analyzed. In the survey [50], the backdoor attack strategies were discussed in depth, but they were not discussed in the context of WFL. Also, to the best of the authors' knowledge, none of the existing surveys has taken WCN into consideration. Nevertheless, the security threats induced by WCN are inevitable when discussing the influence of backdoor attacks on WFL.

C. Scope and Contributions

Compared with the existing surveys, this paper presents a comprehensive survey of the latest backdoor attacks on WFL and corresponding defense mechanisms. Over 200 papers from 2015 to 2023 have been reviewed in depth, including those focusing on WFL, the development of backdoor attack methodologies, and defense mechanisms. The key contributions of our paper are summarized as follows:

- We present a holistic review of WFL, emphasizing its vulnerability and security concerning backdoor attacks. The WFL vulnerabilities potentially open the door for backdoor attacks.
- We comprehensively review the existing backdoor attack methodologies based on their attack targets and stages and analyze their strengths and limitations. The mathematical derivation process is studied to understand the working mechanism of backdoor attacks on WFL.
- We provide a systematic classification and discussion on defense mechanisms against backdoor attacks on WFL. We compare defense algorithms at different phases of the model training, and analyze the limitations of existing defense schemes.
- Open issues and research challenges are pointed out, followed by potentially promising research directions in this fast-growing, under-explored domain. For instance, backdoor attacks based on data poisoning are relatively weaker and more eliminable, while those based on model poisoning exhibit stronger attack performance and stealthiness at the expense of higher costs. The detection of backdoors typically relies on threshold design, and the training of existing defense mechanisms often requires access to historical information about the attacks.

D. Structure of This Survey

The structure of this survey is shown in Fig. 1. Section II gives the overview of WCN and WFL. The threat model of backdoor attacks on WFL is also discussed. In III, the mechanisms of the existing backdoor attack methods are reviewed. In Section IV, the recent works on backdoor attack defense methodologies and their limitations are analyzed. The performance of backdoor attack strategies and defense methods in each phase is presented and analyzed in Section V. The lessons learned are discussed in Section VI. Conclusion and future works are summarized in Section VII. Abbreviations, acronyms, and notations used in this survey are defined in Table II. The definitions of technical terms used in WFL networks, backdoor attacks, and defenses are listed in Table III. These definitions will be consistently referred to throughout the rest of this survey.

II. Background and Threat Models

This section presents an overview of WFL and backdoor attacks on WFL. The technical terms are first introduced. A brief overview of the wireless communication network is then presented. After that, the working mechanism of WFL is reviewed. The motivation for backdoor attacks and their defense methods are presented by analyzing the major security challenges encountered in deploying WFL. An overview of existing backdoor attack methodologies is then carried out, including the evaluation metrics and comparison of different types of backdoor attacks for different attack targets and stages.

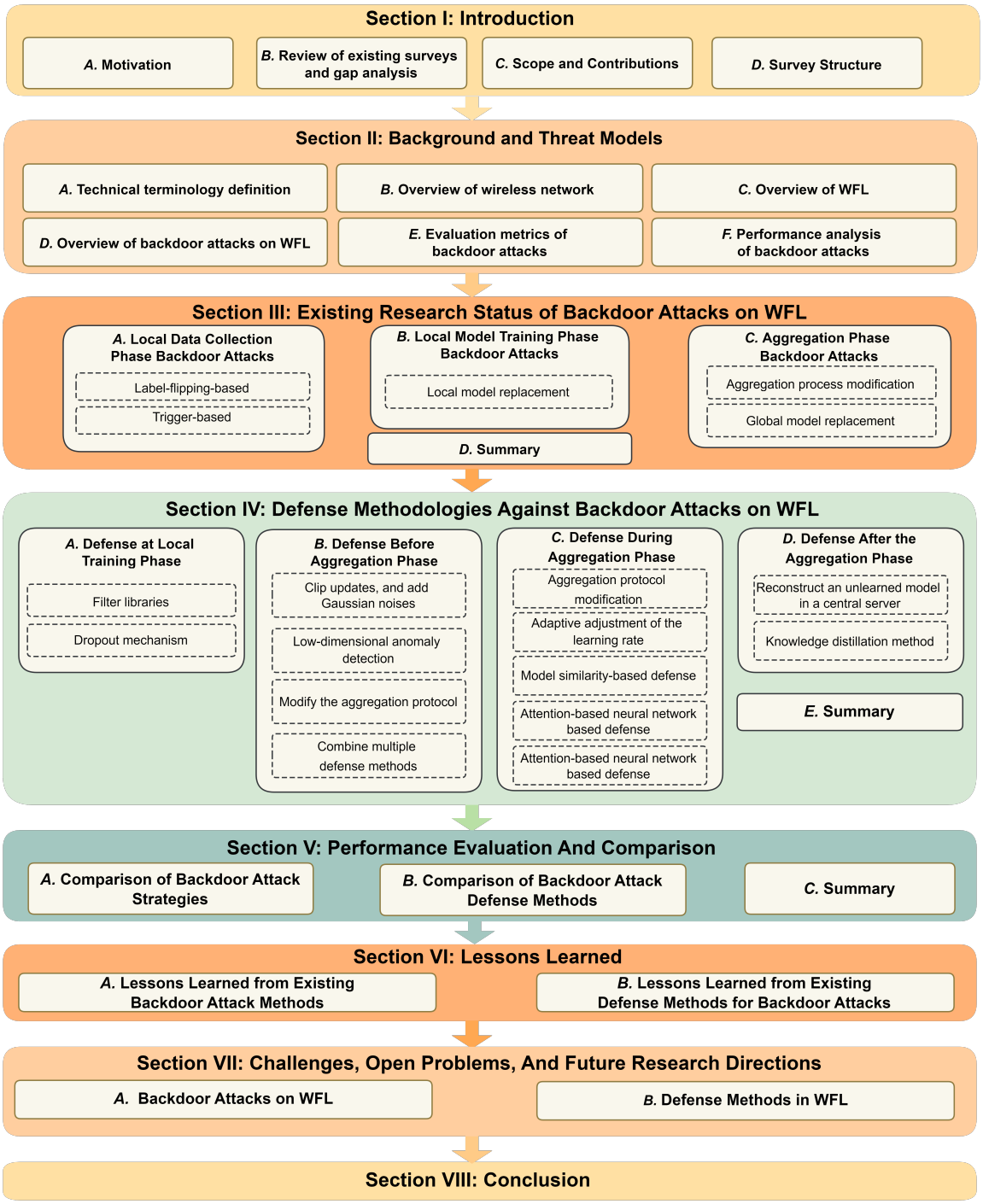


Fig. 1: The structure of this survey.

A. Overview of Wireless Communications Technologies

Wireless communication is one of the fastest-growing technologies in recent years [55], [56]. The devices connected to the wireless network exchange data via radio waves instead of physical connections, such as cables or fibers. According to different communication ranges and data capacities [57], many wireless communication technologies are developed, including Wi-Fi, Cellular Networks, Bluetooth, Near Field Communication, and Zigbee [58]. With the advancements in wireless technologies, the wireless network has become an essential part of our daily lives.

The rapid progress of wireless communication technologies has also promoted the development of machine learning algorithms. The wireless network can collect a large amount of data from various sources, which machine learning algorithms can apply to analyze and train a converged model. Fig. 2 shows an example of a machine learning training task in wireless communication networks. The wireless network enables a flexible and convenient way of transmitting data from local devices for machine learning applications, such

TABLE II: List of abbreviations, notations, and their definitions

Type	Abbreviation	Description
Technical Terms	IID	Independent and identically distributed data
	non-IID	Non-independent and identically distributed data
	FL	Federated learning
	WFL	Wireless federated learning
	AI	Artificial intelligence
	DP	Differential privacy
	SGD	Stochastic gradient descent
	TCP	Transmission control protocol
	UDP	User datagram Protocol
	OTA	Over-the-air
Evaluation Metrics	PDR	Poisoned data rate
	CCR	Compromised client rate
	BAR	Backdoor accuracy rate
	ASR	Attack success rate (equivalent to BAR)
	BTA	Backdoor task accuracy (equivalent to BAR)
	MAR	Main task accuracy rate
Mathematical Symbols	N	The total number of clients involved in WFL
	m	The number of randomly selected clients within one training iteration.
	t	Time index
	T	The number of FL training rounds
	G^t	The global model at the start of the current training iteration.
	G^{t+1}	The updated global model after current training iteration
	L_i^t	The local model at client i
	D_i^t	Benign training dataset of client i
	${D_{Ai}}^t$	Backdoored training dataset of client i
	$F_i(L_i^t, D_i^t)$	Loss function of the local training algorithm at client i
	$F_{Ai}(L_i^t, D_{Ai}^t)$	Loss function of the malicious objective at client i
	α	The tradeoff between the benign and malicious objectives
	p_i^{t+1}	Clean local model update of client i at t-th training iteration
	p_{Ai}^{t+1}	Poisoned local model update of client i at t-th training iteration
	p_g^{t+1}	Aggregated global model update at t-th training iteration
	β_i^{t+1}	Weight of client i at t-th training iteration
	β_{Ai}^{t+1}	Manipulated weight of compromised client i at t-th training iteration
	η^t	Global learning rate at t-th training iteration

as intelligent transport [59]–[61], intelligent marketing [62], intelligent manufacturing [63], smart home [64], intelligent security [65], and intelligent health [66]. Machine-learning algorithms also contribute considerably to wireless communication development [67] and failure analysis [68]. Efforts have been made to train a machine-learning model that can be applied to optimize resource allocation in wireless networks and improve overall network efficiency [69].

On the other hand, the wider adoption of machine learning applications in wireless networks has prompted a multitude of ongoing research and development endeavors in the field of wireless communications. Organizations, such as IEEE and 3GP, continuously update the wireless standards, for example, Wi-Fi standard IEEE 802.11 [70], to push the boundaries of wireless communication. To further protect participants' private information when exchanging information during the training process [71], many encryption algorithms and authentication techniques for wireless networks have been proposed [72]. The next-generation wireless communication technology 6G is also a frontier research topic [73]. 6G can provide a higher data rate, ultra-low latency, massive device connectivity, and other advanced features, which are precisely required by machine learning applications [74]. The combination of wireless communication networks and machine learning techniques is able to improve network management and further enhance user experiences in various aspects.

TABLE III: Technical terminologies related to backdoor attacks on WFL

Term	Definition
Client	General Participant in WFL framework. Every client contributes to the convergence of the global model.
Attacker	The malicious clients in the client party. Attackers aim to inject a backdoor into the global model while keeping the convergence of the main task.
Compromised client	The client fully controlled by the attacker. All the attacks are conducted through compromised clients.
Defender	The defense mechanism mounted in the Federated framework. The design requirement is to filter out the malicious input and eliminate the backdoor effect in the global model.
Trigger input	The malicious input with trigger pattern. The trigger pattern can be random or model-dependent regarding different trigger designs. Since trigger-based backdoor attacks normally target image classification tasks, the trigger pattern is pixel-based.
Target label	The Label where the attacker aims to inject the backdoor.
Attack round	The training round when the attacker conducts the attack. Depending on different attack types, the attack round can either be a single round or multiple rounds.

It can be predicted that FL will be widely deployed in WCN for various applications owing to its diverse merits. However, WFL will be more vulnerable to backdoor attacks than its wired counterpart due to a combination of factors inherent to wireless networks. Here are some of the reasons:

- Network Security: Wireless networks are generally less secure networked than wired networks due to the broadcast nature of radio. Data transmissions over wireless links are through airwaves, making it easier for attackers to eavesdrop [75]–[78], jam [79], and manipulate data, e.g., replay attack and Sybil attack, or intrude the networks [80], [81].
- Device Vulnerabilities: WFL usually involves an extensive range of devices, including low-cost IoT devices and smartphones. These devices can often be less secure than servers or desktop computers that might be used in wired FL [82]. If deployed remotely, they might not have the latest security patches, leaving them vulnerable to backdoor attacks. Some devices can even be physically captured, compromised, and put back into the networks.
- Data Transmission: Data transmission in wireless networks can be inconsistent due to interference or signal strength variations. This inconsistency can lead to data corruption or loss, and an attacker might exploit the inconsistency to inject malicious backdoor attacks. On the other hand, machine learning has demonstrated its effectiveness in transmission resource allocation under both centralized settings [83], semi-distributed settings [84], and distributed settings [85]–[87].
- Scalability and Management: WFL is designed to be highly scalable and to deal with a large number of nodes (devices). This makes managing and securing each device more difficult and increases the overall attack surface.
- Distribution: FL is inherently a distributed learning approach. The decentralization is even more pronounced in a wireless environment where devices can join and leave the network more freely. This can make it more challenging to maintain constant security measures, making the system more prone to backdoor attacks.

Hence, this paper focuses on WFL and its security concerns related to backdoor attacks. In the next section, an overview of WFL is first presented.

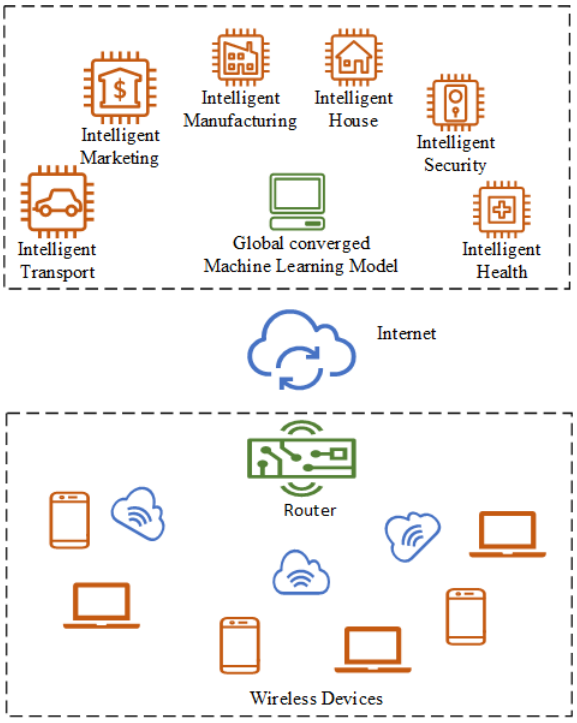


Fig. 2: A wireless network with machine learning applications.

B. Overview of WFL

WFL is a distributed machine learning scheme that connects n clients to train a global model G in the central server by iterative aggregation of the local model L_i, $i = 1, \ldots, N$ over a wireless network [20], [88], [89]. WFL has been widely deployed owing to its several features of distributed architecture, a large number of participants, and privacy concerns [90]–[92]. The distributed characteristic of WFL allows the broad participation of clients regardless of geographical restrictions. The scale of clients participating in WFL can be immense. During the aggregation process, the local model updates are uploaded to the WFL server (e.g., a BS), while the training data remains securely stored on clients' devices locally. Thus, WFL can be adopted to train tasks that highly rely on sensitive private data such as clients' personal information, location, etc. [35], [93]–[97].

A typical training process of WFL is shown in Fig. 3. In each training iteration, the central server broadcasts the current global model G^t to all the participants within the network, and a subset of m clients are selected to join the training process. The selection of m is subject to the trade-off between

the training speed and efficiency. Each training round can be divided into three phases:

- **Local training phase**
  In the t-th training iteration, the i-th client re-trains the local model with the local training data D_i^t. $t \in \{1, \ldots, T\}$. $i \in \{1, \ldots, m\}$. The local training objective is defined as:

$$L_i^t = \arg\min_{L_i^t} F_i(L_i^t, D_i^t), \tag{1}$$

  where $F_i(L_i^t, D_i^t)$ denotes the local loss function. To minimize the loss function, the local model update is calculated, denoted as p_i^{t+1}.
  It is worth mentioning that the model update can be in a different form depending on different training algorithms. For example, if the WFL is trained with the Stochastic Gradient Descent (SGD) method, p_i^{t+1} can be formulated as the gradient of the loss function:

$$p_i^{t+1} = \frac{\partial F(L_i^t, D_i^{t+1})}{\partial L_i^t}. \tag{2}$$

  Other local training methods also include Batch Gradient Descent (BGD) [98], Mini-Batch Gradient Descent (MBGD) [99], Root Mean Squared Propagation (RMSprop) [100], etc. After local training, the obtained local model update p_i^{t+1} is sent to the server.

  Considering the limited and often unbalanced learning (or computing) capabilities of wireless devices, the client may train different numbers of local training in a communication round. Moreover, some devices may experience deep fades and require excessive communication time to upload their local models for aggregation. As a consequence, their local training time has to be sacrificed, penalizing the accuracy of their local models.
- **Aggregation phase**
  The server gathers the local model updates to obtain the average global updates ${p_g}^{t+1}$:

$$p_g^{t+1} = \frac{1}{m} \sum_{i=1}^{m} (\beta_i^{t+1} \times p_i^{t+1}), \tag{3}$$

  where β_i is the weight of the i-th client. The aggregated model update is then used for the global model update.

  A typical form of global model aggregation is in the digital domain, where the selected and admitted local models are digitized and aggregated, as described in (3). Another increasingly considered method for global aggregation in WFL is Over-The-Air (OTA), where all selected clients synchronize the transmissions of their local models and align their signal strengths received at the server [101], [102]. By exploiting the additivity of the electromagnetic field, the local model can be naturally aggregated at the reception of the server.
- **Global model update phase**
  The global model is updated after receiving the aggregated model update:

$$G^{t+1} = G^t + \eta^t p_g^{t+1}, \tag{4}$$

  where η is the global learning rate, controlling the training step size of the t-th iteration. The updated global model can be distributed to the same or different clients chosen in the next round.

Decentralized WFL has been recently developed, e.g., in [103]. To solve the network asynchrony typically undergone in an FL process, a fully decentralized FL framework named IronForge is proposed in [104]. Decentralized WFL is an advanced architecture that offers a flexible and scalable approach to perform model aggregations in large-scale or unstable networks. The decentralized WFL architecture enables these local nodes to train their models independently using their respective local data. This approach is particularly useful in situations where the network may experience intermittent connectivity, limited bandwidth, or strict privacy regulations that restrict data sharing.

Different from centralized WFL where all participants aim to train a unique global model at a central node, there is no central node in decentralized WFL. Each client of Decentralized WFL can choose to train a new local model and upload the model update to other participants within its communication range or aggregate the model updates received from others to train a new model and broadcast it into the network again. The models distributed to different participants can be considerably different, even upon the convergence of decentralized WFL.

Both centralized and decentralized WFL allow participants to join and leave the network freely and do not require private data sharing among them. Recent works have shown that WFL cannot always guarantee the security of the training process [46], [82], [93], [105], [106]. Existing WFL frameworks are vulnerable to backdoor attacks. Malicious users can control one or multiple clients, namely, compromised clients, to inject a backdoor into the global model [32]. The primary objective of the backdoor attack is to generate a converged global model with high accuracy of the main task and the backdoor-injected sub-task [107].

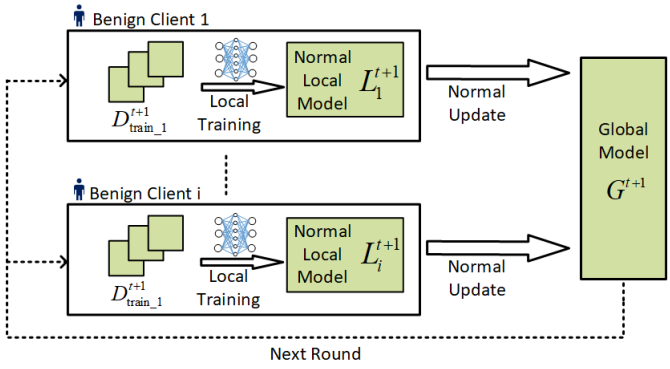


Fig. 3: The overview of the WFL training strategy for producing aggregated global ML model using devices' private data.

C. Features of WFL

WFL offers distinct features compared to traditional FL, especially in scenarios where mobility and real-time data processing are crucial [108]. The essence of WFL lies in its ability to facilitate learning processes for mobile devices, such

as smartphones, tablets, and IoT devices [109]. These devices often operate on the move and rely heavily on wireless connectivity, making WFL an ideal solution for ensuring continuous and accessible learning. For instance, in applications such as autonomous vehicles [110]–[112] or mobile health monitoring systems [113], the capability of WFL to enable on-the-go learning and instantaneous model updating is indispensable. This real-time processing is critical for applications that demand immediate responses based on current data.

1) Exacerbated Vulnerabilities of WFL: Implementing WFL in a wireless system such as a cell-free massive MIMO system introduces specific privacy challenges due to the dispersed distribution of access points and users [94]. The wireless communication among these points adds complexity, increasing the risk of sensitive data exposure. Cell-free massive MIMO systems involve a large number of distributed access points serving multiple users. In such a setting, the wireless transmission of data over numerous channels between these access points and users can create multiple opportunities for data interception or leakage. The complexity and scale of such systems make it challenging to uniformly secure all communication links, thus heightening privacy risks.

In [95], decentralized WFL (DWFL) is studied, where there is an increased risk of private information leakage due to the transmission of model parameters over a wireless network. This risk is exacerbated by the decentralized nature of DWFL, where data is distributed across multiple wireless nodes. In a DWFL setting, the transmission of parameters between multiple nodes over wireless channels introduces potential points of vulnerability. Each node in the network can be seen as a potential weak point where information can be intercepted or corrupted. DP is adopted to preserve the data privacy of the nodes, and its impact on the convergence of DWFL is analyzed.

The authors of [96] considered decentralized inference with graph neural networks (GNNs) in wireless networks, highlighting the vulnerability in handling graph-structured data over wireless channels. The distributed nature of GNNs in a wireless context adds complexity to maintaining data privacy. In decentralized GNNs, neighboring nodes exchange information through wireless channels. This exchange is crucial for inference but introduces vulnerabilities due to the inherent characteristics of wireless media, such as fading and noise. These factors can deteriorate the quality of information exchange and impact the performance of the inference process. The adverse impacts of imperfect wireless transmission on the inference robustness of GNNs highlight the complexity of maintaining privacy in this setting.

There exists a unique aspect of OTA-WFL, where gradient transmissions are uncoded [97]. The wireless aspect introduces a vulnerability where the privacy of local datasets can be compromised through the disclosed aggregation statistics over the air. In OTA-WFL with uncoded transmission, gradients are transmitted directly over wireless channels without encoding or encryption. The lack of coding or encryption in transmission means that any vulnerability in the wireless channel, such as weak signal encryption or interception by unauthorized receivers, can lead to significant privacy breaches. Moreover, since gradients carry information about the local datasets, their interception can reveal sensitive data.

Moreover, with the broadcast nature of wireless interfaces, adversaries can potentially overhear local model updates from multiple or even all clients and accordingly create malicious local models to compromise WFL, e.g., using generative neural networks and machine learning technologies. For example, in [35], [114], a new adversarial attack on WFL is studied by exploiting the broadcast nature of wireless channels, where a rogue client collects the local model updates that benign clients upload. The rogue client can take these overheard local model updates as input to a graph adversarial encoder (GAE) to produce a malicious local model update. The malicious local model update can poison WFL, and stop or slow down the convergence of WFL [114].

2) Inherent Privacy-Preserving Capability of WFL: Some inherent features of wireless channels and transceivers can be potentially leveraged to preserve the privacy of WFL. While lossy and erroneous wireless channels may hinder the convergence of WFL, they may weaken the attack strength from backdoor, data and model poisoning attacks. This is due to the fact that the random errors resulting from wireless fading and receiver noises can potentially serve to perturb the model updates, hence achieving the effect of DP and alleviating the impact of the attacks. Under the setting of OTA-WFL, the authors of [97] control the relative intensity of the receiver noise power at the central server by controlling the transmit powers of the clients. The resulting relative intensity of the receiver noises can offer privacy-preserving capability, like DP noises. This promising research direction can potentially simplify the protection design for WFL. The research is still at a very early stage. More breakthroughs are underway.

D. Overview of Backdoor Attacks on WFL

In this section, an overview of backdoor attacks on WFL is presented. The process of backdoor attacks is shown in Fig. 4. The existing backdoor strategies are categorized into three major types in terms of attack phase: (1) local data collection phase backdoor attack, (2) local model training phase attack, and (3) server aggregation phase backdoor attack. Depending on different types of attacks, the information required by adversaries may differ.

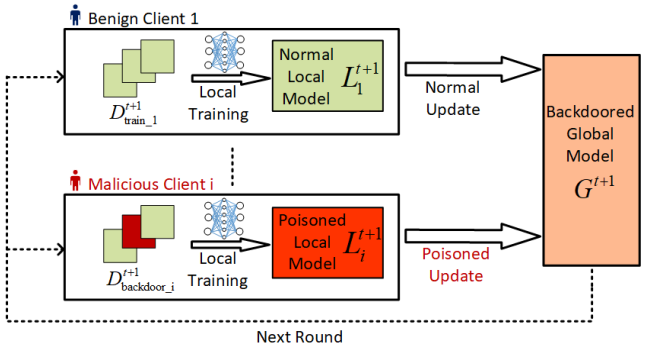


Fig. 4: Backdoor attack process in WFL.

- **Backdoor attack at local data collection phase**

The local data collection phase backdoor attack, also known as the data poisoning backdoor attack [115]. When conducting backdoor attacks during the local data collection phase, the attacker only needs to control a fraction of the training data set [116].

In a local data collection phase backdoor attack, an individual feature or a small region of the local training data is modified by malicious users deliberately [117]. The fraction of training data controlled by the attackers is denoted as Poisoned Data Rate (PDR) [118]. Specifically, in the i-th compromised client at the t-th training iteration, PDR can be mathematically formulated as:

$$PDR = \frac{\left|\mathrm{D}^t_{Ai}\right|}{\left|D^t_i\right|}, \tag{5}$$

where D^t_i stands for the benign dataset of this client, D^t_{Ai} refers to the malicious dataset, and $|\cdot|$ stands for cardinality.

The training objective of the i-th compromised client can then be formulated as:

$$L^t_i = \arg\min_{L^t_i}\{\alpha F_i(L^t_i, D^t_i) + (1-\alpha)F_{Ai}(L^t_i, D^t_{Ai})\}, \tag{6}$$

where $F_i(L^t_i, D^t_i)$ is the WFL learning objective and $F_{Ai}(L^t_i, D^t_{Ai})$ is the malicious objective shared by all attackers (in the case of a coordinated attack discussed in Section III-A2). The local model is induced to misclassify the testing samples with the target label specified by adversaries. Considering the stealthiness of the backdoor attacks, a compromised client also needs to optimize the WFL objective. The trade-off between training the benign and malicious objectives is defined as α, and $\alpha \in [0, 1]$.

In practice, one way to launch the local data collection phase backdoor attack is to directly replace the training data of the target label with arbitrary incorrect samples. A successful attack will lead the model to misclassify the testing samples of the target label according to the attacker's wish, e.g., for a room with facial recognition lock, a man without access will be permitted to enter the room by mistake [119], [120].

The other type of widely applied backdoor attack strategy during the local data collection phase is the trigger-based backdoor attack. The trigger is a pattern parasitized in clean data [121]. The model will behave normally and predict a target class when the trigger appears, e.g., in the same room with a facial recognition lock, all men with glasses will be authorized to enter. In this case, the glass is the trigger injected. Depending on the number of compromised clients the attacker controls, the trigger can be injected into one client or distributed separately to multiple clients.

- **Backdoor attack at local model training phase**
In local model training phase backdoor attacks, an attacker is required to fully control the compromised clients, including the training data set and training algorithm [29], [32], [116], [122]–[124]. the malicious goal is to tamper with the uploaded updates and further influence the global model. Take the i-th compromised client, for example. The attack process can be formulated as follows:

$$p^{t+1}_{Ai} \leftarrow L^{t+1}_{Ai} \leftarrow \min_{L^t_i} F_{Ai}(L^t_i, D^t_{Ai}), \tag{7}$$

where p^{t+1}_{Ai} is the poisoned local model update sent to the server. As a consequence, the clean local model will be replaced by a backdoor-injected model [125].

Backdoor attacks conducted during the local model training process only aim to enhance the local model. The attacker is capable of conducting attacks to compromise both the convergence of the training data set and the calculation of the uploaded update of the individual clients. The impact of the training phase attack is limited to the compromised clients controlled by the attacker [32], [124], [126].

- **Backdoor attack at server aggregation phase**
Backdoor attacks conducted during the aggregation process generally do not alter the targeted model. Instead, the attacker is keen on tricking it into producing wrong predictions and causing security problems [32]. The effectiveness of an aggregation phase attack is mainly determined by the knowledge of the model controlled by attackers.

In the aggregation phase backdoor attack, the attacker controls the aggregation algorithm used to combine clients' updates into the global model:

$$p^{t+1}_{Ag} = \frac{1}{m}\left(\sum_{i=1}^{m_d}\beta^{t+1}_i \times p^{t+1}_i + \sum_{j=1}^{m_b}\beta^{t+1}_{Ai} \times p^{t+1}_{Ai}\right), \tag{8}$$

where m is the number of clients participating in the t-th training iteration as considered in (3); M_d is the number of benign clients; and M_b is that of compromised clients. Therefore, $m = m_d + m_b$. The weight of the poisoning model $\beta^{t+1}_{Ai}, i \in [1, m_b]$ is manipulated to ensure the successful implementation of the backdoor into the global model of WFL.

Compared with training process attacks, the aggregation phase attacks can have stronger attack effects while there is a trade-off between the stealthiness and the attack success rate (ASR).

E. Evaluation Metrics for Backdoor Attacks

The objective of a backdoor attack on WFL is to mislead the global model to misclassify malicious inputs with injected backdoor patterns into the target output, with application to image classification [127]–[136], word prediction [137]–[142], etc. For instance, regarding the image classification task discussed in [33], an attacker aims to mislead the global model to classify images of "truck" as "airplane". Based on the attack objectives, the following metrics are commonly adopted to evaluate the effectiveness of backdoor attacks.

- **Backdoor Accuracy Rate (BAR):** BAR indicates the strength of a backdoor pattern by measuring the ratio of successfully misclassified global outputs using backdoored inputs out of all the outputs obtained from malicious inputs. In other studies, similar terms, such as

ASR [143], [144] and backdoor task accuracy (BTA) [32], are used. Generally, a high BAR is equivalent to a high ASR and BTA.

- **Main Task Accuracy Rate (MAR):** In order to perform a successful backdoor attack, the stealthiness of the backdoor-injected input is also essential. MAR measures the rate of global model outputs that have been denoted as correct classification, which consists of normal output produced by the benign inputs and malicious outputs misled by the backdoored inputs. Different from Byzantine attacks, which aim to jeopardize the convergence of the global model [145]–[147], the backdoor attacks focus on inserting backdoor patterns into the global model while maintaining a reasonable MAR.
- **Lifespan of backdoor:** Due to the nature of FL, the backdoor effect in the global model tends to be diluted with the increment of the training round. The duration of the backdoor effect remaining in the global model is defined as the lifespan of the backdoor. Attackers are keen to keep the backdoor effects in the global model as long as possible, such that the global model could consistently produce the wrong outputs as they want.

Notably, the above-mentioned metrics are also widely used to evaluate the effectiveness of defense strategies. Different from backdoor attack designs, which pursue high BAR, MAR, and long lifespan, the defenders aim to minimize the BAR and lifespan of the backdoor effect while preventing the MAR from degradation.

F. Performance Analysis of Backdoor Attacks

To evaluate the effectiveness, backdoor attack methodologies are tested in multiple WFL applications. Consider a WFL network with a total of m clients. The malicious attackers control one or more compromised clients to perform the attack. The attack targets can vary depending on the knowledge level obtained by the attacker. Furthermore, the specific design of the attack method is likely to be limited to specific kinds of WFL applications. Regarding the attack target, complexity, and applicability, existing backdoor attack methodologies in WFL are summarized below in Table IV. As shown in Table IV, label flipping attacks, coordinated trigger attacks, local model replacement attacks, and aggregation process attacks are likely to be exacerbated in WFL. This is due to the broadcast nature of wireless communications, which facilitates various forms of cyber threats, such as spoofing attacks, Sybil attacks, and eavesdropping attacks.

The performance of the existing backdoor attack methods can then be evaluated. The comparison is summarized in Table V. It can be concluded that the data poisoning backdoor attacks, including label flipping-based attacks and centralized trigger-based attacks, have the lowest attack cost. In return, the attack strength is weak, and the backdoor effect will be quickly diluted. The trigger-based backdoor attack can be improved using a coordinated trigger pattern, achieving a higher attack success rate. The model poisoning backdoor attacks implemented at the local training and server aggregation phases take advantage of the white box attack model. The backdoor effect can last longer if the attack success rate is higher. Consequently, the attackers need to know more about the entire network, and the attack cost is greatly increased.

III. Status Quo of Backdoor Attacks on WFL

Wireless networks have geographically separated structure, allowing each client participating in the WFL network to train a unique local model based on its own datasets. Instead of exchanging raw data among participants, only local model updates are uploaded for aggregation. The non-IID training data set in WFL helps to protect the data privacy of the participants [160]. Each client's data distribution is not representative of the global data distribution, implying that the categories in each client are incomplete. As a result, even if one client's data is compromised, the overall model will be minimally affected.

On the other hand, the features of WFL, including non-IID training datasets, distributed learning structure, and information interchange via WCN, allow adversarial clients to conduct backdoor attacks to stealthily compromise the global model robustness. The primary objective of backdoor attacks is to manipulate the converged global model to misclassify specifically crafted inputs with hidden backdoors into a desired target label while ensuring that normal inputs are classified accurately. Thus, the global model exhibits convergence on the main task and leaves a backdoor on the target label.

Moreover, the communication efficiency of a client largely depends on the local devices' processing abilities and their geographical distributions in mobile edge computing systems [161]. Clients with faster computing speed and better communication link qualities are more likely to be selected in a communication or aggregation round. Consequently, the WFL could be particularly vulnerable to backdoor attacks launched by attackers with better link qualities, e.g., closer to the server or possessing a strong line-of-sight (LoS) to the server.

In this section, the existing techniques to perform the above motioned backdoor attacks in three different phases are comprehensively reviewed, as presented in Fig. 5.

A. Backdoor Attacks at Local Data Collection Phase

In backdoor attacks during the local data collection phase, the backdoor is injected via data poisoning backdoor attacks. Data poisoning backdoor attacks can be further classified into two major methodologies: label-flipping-based attacks or triggered-based attacks. The label-flipping-based backdoor attack is proposed in [29]. Fig. 6 shows the mechanism of label flipping based backdoor attack. In label-flipping-based backdoor attacks, the training data set under the targeted label is replaced by the backdoored data set [162]. The poisoned data set is then used to train the local model for further aggregation.

1) Label-flipping based backdoor attack: The label flipping-based backdoor attack is relatively easier to conduct since it does not require any prior knowledge of the local data distribution, local model training process, and aggregation protocol [123], [163]. The attacker only needs to manipulate the training data under the target label. By strategically flipping labels in the poisoned samples, attackers can bias the model

TABLE IV: Summary of existing backdoor attack methodologies in WFL

Attack Type	Attack Target		Compromised Client Number		Attack Iteration		Main Task	Exacerbation by WCN
	Data	Model	Single	Multiple	Single	Multiple		
Label Flipping [29], [123], [148], [149]	√	×	√	√	√	√	Image Classification	√
Centralized Trigger [136], [150]–[152]	√	×	√	×	√	√	Image Classification	×
Coordinated Trigger [116], [122], [153], [154]	√	×	×	√	×	√	Image Classification	√
Local Model Replacement [32], [124], [155], [156]	×	√	√	×	√	√	Image Classification; Sentiment Analysis	√
Aggregation process Attacks [37], [157]–[159]	×	√	√	×	√	√	Image Classification; Sentiment Analysis; Word Prediction	√

towards a specific target behavior. For example, in a face recognition model, the attacker may flip labels to make the model misidentify a particular person or a group of people.

However, since the number of participants in WFL is typically large, the effect of an injected backdoor from a single attacker in a single round can often be diluted. In order to obtain better attack performance, multiple malicious clients can collude to conduct an attack cohesively. They need to be selected in multiple rounds. The inevitable problem of the simple label-flipping method still exists in that the flipped data set under the targeted label results in a remarkable outlier in the model training. Thus, it can be detected by norm clipping [164] or differential privacy [38], [38], [164], [165].

To improve the stealthiness of the label flipping-based backdoor attack, the target label is selected on the tail in [33]. In this case, the data distribution is required, and the edge data set is selected as the target. Replacing the edge target data set with a backdoored data set, the local model behaves correctly with normal inputs while misclassifying the adversarial inputs on the edge label. The process of label-flipping backdoor attacks is shown in Fig. 6.

In general, label-flipping depends on the data sets and is agnostic to the models to be trained and the communication interfaces. Nevertheless, label-flipping attacks can hinder the convergence of a WFL process. The injected poisoned samples can introduce noise and confusion, making it harder for the model to converge. Given the limited communication resources, prolonged training processes can congest wireless interfaces and block other services and applications.

Limitations: The effect of the label-flipping backdoor attack on the edge label can last longer compared with the simple label-flipping method. Furthermore, the selection of the edge

TABLE V: Comparison among existing backdoor attacks on WFL: Attack cost, ASR, duration, prior knowledge, and potential exacerbation caused by WCN

Attack Type	Attack Cost	Attack Success Rate	Duration	Prior Knowledge	Exacerbation by WCN
Label Flipping [29], [123], [148], [149]	Low	Low	Short	Label information	√
Centralized Trigger [136], [150]–[152]	Low	Low	Short	Part of the local training data	×
Coordinated Trigger [116], [122], [153], [154]	Medium	High	Medium	Part of the local training data	√
Local Model Replacement [32], [124], [155], [156]	High	High	Long	Local training data; Local training process, Possible defense mechanisms	√
Aggregation Process Attacks [37], [157]–[159]	Highest	Highest	Long	The number of participants, Training algorithm, Possible defense mechanism, Aggregation protocol, Global model convergence process.	√

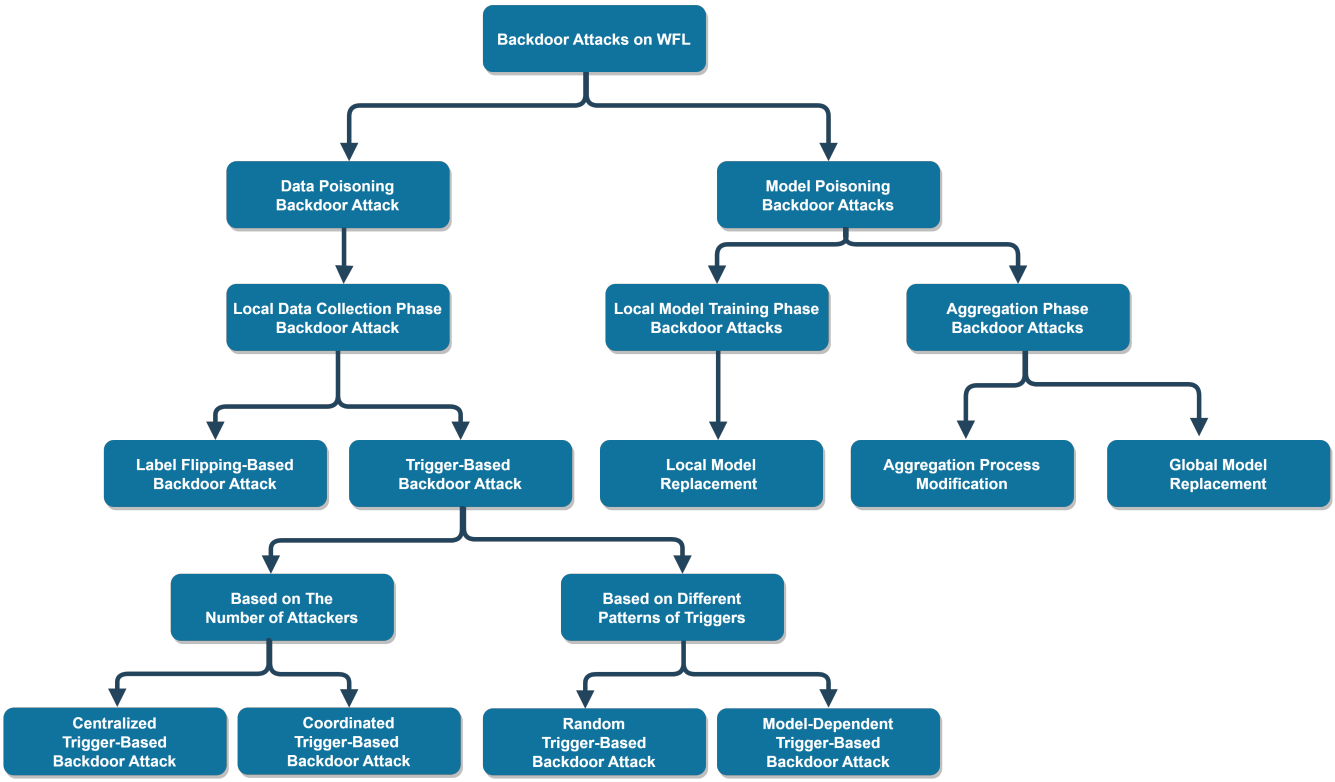


Fig. 5: A taxonomy of backdoor attacks on WFL.

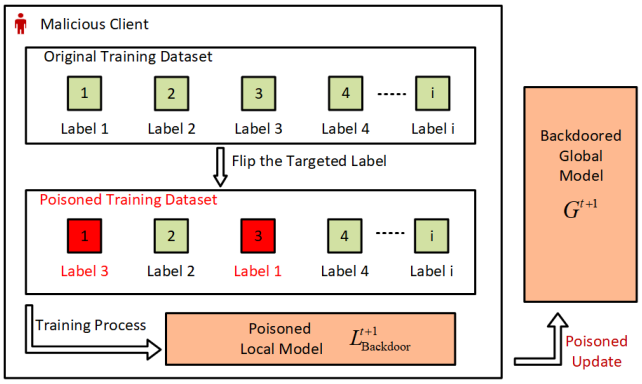


Fig. 6: Label-flipping-based backdoor attacks (local data collection phase): The attacker only needs to manipulate the training data under the target label.

data set minimizes the outlier, allowing it to trick the defender. On the other hand, the effect of the edge label flipping-based backdoor attack is also limited at the edge label.

2) Trigger-based data poisoning backdoor attack: In trigger-based backdoor attacks, triggers are designed to be injected into the training data under the targeted label [128], [166]. The trigger or a specific pattern embedded into the training data can cause the model to exhibit malicious behavior during the inference phase. The primary objective of trigger-based backdoor attacks is to mislead the model to classify the trigger-mounted inputs into the target label while behaving normally with non-trigger inputs [38], [116], [126]. According to the number of adversarial participants involved in an attack, trigger-based backdoor attacks can be classified into centralized trigger-based backdoor attacks and coordinated trigger-based backdoor attacks.

Fig. 7 shows the attack mechanism of a centralized trigger-based backdoor attack. In [116], the trigger is injected in a coordinated pattern via multiple attackers. As illustrated in Fig. 8, the global trigger pattern is distributed in four parts. It can be seen from Fig. 9 that each attacker knows part of the trigger, and the coordinated attack is conducted to inject the completed trigger into the final model [167].

Compared with the centralized trigger-based backdoor attacks, the coordinated trigger pattern shows better stealthiness, faster model convergence speed, and a higher attack success rate [168]. Coordinated attacks involve multiple devices participating in a WFL process. This increases the attack surface and potential vulnerabilities in the WFL. Attackers can exploit the decentralized nature of WFL to distribute triggers across a diverse set of devices, making it more challenging to identify and mitigate the attack. The major problem of the existing trigger-based backdoor attacks, e.g., [168], is that the trigger is fixed at the beginning of the attack and cannot be adaptively adjusted during the attack process, which limits the attack effect.

Depending on different patterns of the trigger, the trigger-based backdoor attacks can be further divided into random trigger backdoor attacks and model-dependent trigger backdoor attacks, as follows.

- In a random trigger-based backdoor attack, the trigger is generated independently of the model [122], [140], [169], [170]. The advantage is that the attack cost is greatly

reduced. In return, the effect of the triggers is relatively weaker. In contrast to a random trigger, the other type of trigger is generated, according to the target label [116], [171].

- The model-dependent trigger can help to excite the target label and has a higher probability of being successfully injected into the trained local model before aggregation [172].

Under a wireless configuration, an attacker is likely to eavesdrop on the local model updates of the other benign clients by exploiting the broadcast nature of radio [173]. With the additional knowledge of the local models along with the global model, the model-dependent triggers can be potentially further enhanced, e.g., by employing adversarial generative networks and training [174].

Limitations: The common problem of centralized trigger-based backdoor attacks is that the single attacker's effect is limited, and it cannot be ensured that the attacker is selected in every round. Thus, the increment of the compromised clients controlled by the attacker can contribute to improving the attack performance [29].

3) Comparison between label-flipping-based and trigger-based backdoor attacks: Both label-flipping-based and trigger-based data poisoning backdoor attacks can be treated as black-box attacks [33]. Attackers only need to know part of the local training dataset. By manipulating the data set under the targeted label, the backdoor is injected into the model at a relatively low cost. However, the attack effect can be either quickly diluted or eliminated under existing defense mechanisms. The attack strength is weaker than that of model poisoning backdoor attacks.

Moreover, the performance of backdoor attacks targeted at the local training datasets can be sensitive to the communication channel quality of the local devices or clients. For instance, the increasingly widely deployed wireless communication technologies, e.g., 5G, operate in high-frequency bands of up to 40 GHz, and the wavelength of the carrier frequency can be as short as 7.5 millimeters [175]. The short wavelengths can lead to fast and drastically changing channel conditions of each client [176]. As a consequence, the transmission time required for a client to upload its local model in a communication round can change significantly between communication rounds. When the channel link quality is poor, i.e., a client is in a deep fade, the transmission time could be excessively long. The local training time that remains in a communication round would be short, leading to insufficiently trained local models and, hence, vulnerabilities to backdoor attacks. The ASR is higher when the attacks are launched at compromised clients under better communication conditions. In this sense, fair client selection and scheduling of WFL can help prevent an individual client or a group of clients, including adversarial clients, from dictating a WFL process.

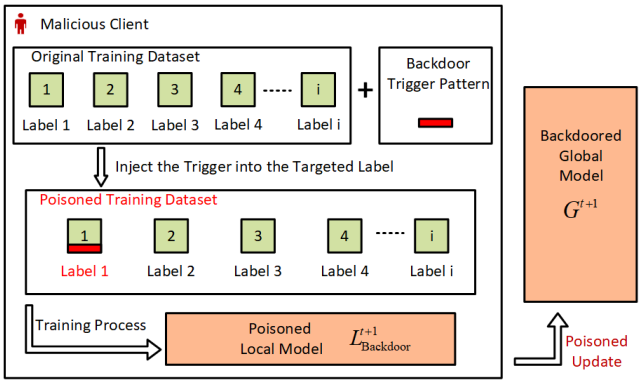


Fig. 7: Centralized trigger-based backdoor attacks (local data collection phase): Centralized triggers are designed to be injected into the training data under the targeted label.

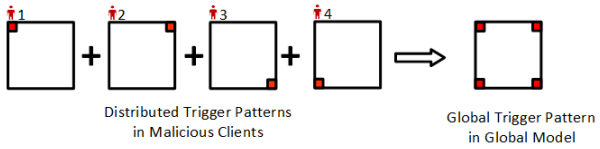


Fig. 8: Distributed global trigger pattern: The trigger is injected in a coordinated pattern via multiple attackers.

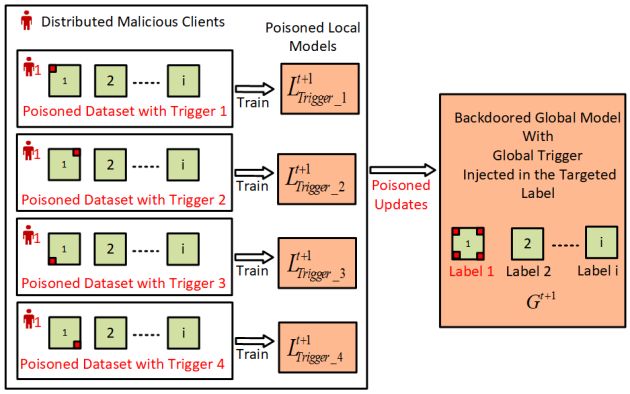


Fig. 9: Coordinated trigger-based backdoor attack(local data collection phase): each attacker knows a part of the trigger, and the coordinated attack is conducted to inject the completed trigger into the final model.

B. Backdoor Attacks at Local Model Training Phase

In the local model training phase, the backdoor attacks can be conducted by manipulating the model parameters during the training process or modifying the training algorithm [126], [177]. The original benign local model is replaced by the poisoned one, and the poisoned updates generated from the replaced local models are further uploaded for aggregation. The attacks during the training process can also be classified into one kind of model poisoning backdoor attack. The model poisoning attack structure is shown in Fig. 10.

As a matter of fact, model-poisoning backdoor attacks take one step more than data-poisoning backdoor attacks [37]. Since data poisoning backdoor attacks inject the backdoor into the training data set, the model poisoning attack in the local training phase aims to ensure the backdoor will be contained in the local model that is uploaded to the aggregation phase. It has been discussed in the previous section that the simple

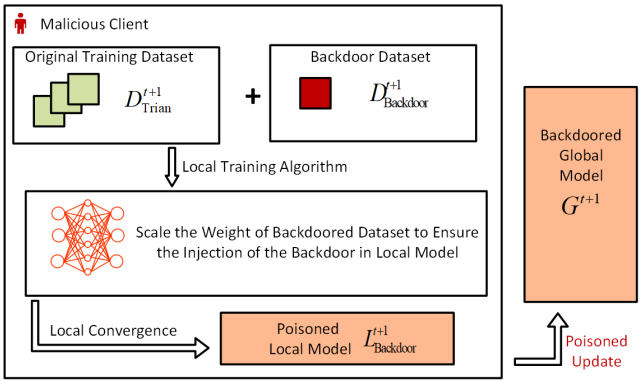


Fig. 10: Local model replacement backdoor attacks (local model training phase): Manipulate the model parameters during the training process or modify the training algorithm.

label-flipping attack exhibits a low attack success rate. A loss function is then designed to minimize the number of outliers caused by the flipped training data, thus balancing the trade-off between the MAR and the BAR. In such a way, the backdoor is more likely to be injected into the converged model and misclassify the malicious inputs.

Apart from designing the loss function to change the training parameter, model poisoning backdoor attacks in the local model training phase can also modify the training algorithm [164]. When the training data set under the targeted label is replaced by the backdoored data set, the local training data becomes the union of the original data set and the backdoored data set. In this case, a malicious model trained with a normal standard gradient descent (SGD) algorithm will lead to a larger model shifting than models trained with a clean data set. As a result, it will be easily detected and deleted from the global aggregation. A projected gradient descent (PGD) algorithm is designed in [33]. In this framework, the updated malicious model is trained by periodically projecting its parameters onto the global model download from the central server at the start of this round. With a newly designed training algorithm, the norm difference between the backdoored and benign models decreases, the backdoor becomes harder to detect, and the attack success rate improves. However, the attack effect is limited on the label with a low distribution rate since the backdoor target is the edge label.

Compared with data poisoning backdoor attacks in the local data collection phase, the model poisoning backdoor attacks in the model training phase achieve better stealthiness and a higher attack accuracy and success rate.

Limitations: The attacks at this stage are also known as white box attacks. The attacker needs to master the training data set, as well as the local training process. Furthermore, the knowledge of the defense mechanism adopted in the WFL network is also required to ensure the injection of the backdoor. The attack cost is considerably higher than the data poisoning backdoor attacks.

Compared to backdoor attacks during the local data collection phase, the attacks, particularly the model poisoning attacks (compared to the data poisoning attacks), are more subjected to the communication signal strength at this stage. In the case where the compromised clients controlled by the attacker are geographically distant, the communication signal strengths are likely to be weak, and a large part of each communication round is likely to be spent on the transmission of their local models. The compromised clients can spend less computing time training their adversarial local models and creating backdoors. As a result, the backdoor pattern has lower strengths and is relatively harder to be injected [178]. To this end, it is more important to ensure the integrity of the local models from closer clients to mitigate potentially more significant attack strengths from the clients.

Apart from the channel conditions, the shared medium nature of wireless channels can also have a strong impact on the effectiveness of backdoor attacks in the local model training phase. When time-division multiple-access (TDMA) protocols are considered for clients to upload their local models, clients selected and scheduled earlier for model uploading could suffer from substantially shorter local model training time in a communication round and produce premature local models [179]. This may make the models more vulnerable to backdoors embedded into the global models and, subsequently, the local models. On the other hand, when frequency-division multiple-access (FDMA) protocols, including orthogonal frequency-division multiple-access (OFDMA) protocols, are considered, each selected client would enjoy much narrower bandwidths, which would prolong their model uploading time at the cost of their model training time [180].

C. Backdoor Attacks at Aggregation Phase

The backdoor attacks against WFL during the aggregation process can also be categorized as model poisoning backdoor attacks [38]. The attacks are conducted when the local models are trained and uploaded for aggregation to update the global model. The mechanism of backdoor attacks during the aggregation phase is illustrated in Fig. 11.

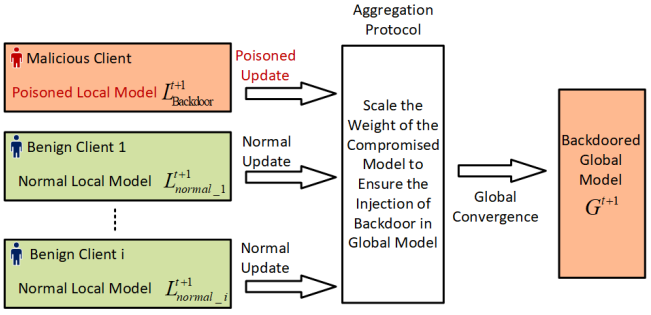


Fig. 11: Aggregation phase backdoor attacks: Modify the aggregation process and replace the global model with the backdoored model.

Model poisoning backdoor attacks during the aggregation phase can modify the aggregation process by scaling the parameter of the uploaded malicious local model [181], [182]. The learning rate is decreased, and the weight of the backdoored model is highly scaled, such that the global model can be replaced by the backdoored model [32]. A loss function is designed to minimize the outlier and thus avoid anomaly

detection. Furthermore, To avoid clipping by the differential privacy, the scale factor is also adjusted to be smaller than the norm threshold. The proposed attack ensures that the backdoor is bound to be injected in one round and will not be diluted. If the attacker can control multiple clients to conduct the coordinated attack, the attack success rate is further improved. It has been testified in [33] that by scaling the malicious model updates, the attacker can replace the global model with the backdoored model targeted on the edge case.

Compared with the data poisoning backdoor attack and model poisoning attack in the model training phase, the model poisoning phase in the aggregation phase shows the greatest attack strength and best stealthiness [31].

Limitations: The model poisoning backdoor attack in the aggregation phase requires the attacker to fully control one or multiple compromised clients, including the training data set and training algorithms. Knowledge of the aggregation process is also necessary. To ensure the successful injection of the backdoor, the attacker also needs to acquire sufficient information about the whole WFL network, including the number of clients involved, the size of the whole data set, as well as the defense mechanism.

Moreover, the attack shows better performance when the global model is close to complete convergence, so the attacker also needs to know the information on the convergence progress of the global model.

However, in decentralized WFL, each client can behave as the server, and only receives model information from nodes within its neighborhood. Distributed WFL methods could be more vulnerable to adversarial attacks. The malicious attacker can then conduct the attack by hijacking one or more compromised clients that are close to the essential node. Thus, the backdoor effect is more likely to be injected into the network and spread more broadly.

D. Summary

A comparison study of existing backdoor attacks is summarized in Table VI. The backdoor attacks conducted during the local data collection phase can be generally classified into label-flipping-based attacks [44] and trigger-based attacks [183]. A backdoor attack at this phase can only pollute the training data, and the attackers do not necessarily know the training algorithm and aggregation protocol. Such black-box attacks featured low attack costs but also relatively low backdoor effects [155]. Once the attacker obtains the knowledge of the training algorithm, it can fully control the compromised client and initiate the model poisoning backdoor attack during the local model training phase. The most commonly adopted strategy is the global model replacement method [184]. The attack model becomes a white box. By taking advantage of fully controlling the compromised model, the attacker can adaptively manipulate the update to improve the attack performance. The aggregation phase attack has a higher requirement for network information. The attacker is supposed to know the information of the aggregation protocol. By manipulating the weight of the adversarial updates, the backdoor pattern is bound to be injected into the converged global model. However, the attack cost is also the highest among all attack methodologies [185].

In Table VII, the strengths and limitations of existing backdoor attacks on WFL are summarized. The label-flipping-based data poisoning backdoor attack is the easiest to implement. The injected backdoor pattern is also easy to detect, and the life span of the backdoor effect is short. The coordinated trigger-based attack shows better attack performance compared with the centralized trigger-based method. In the coordinated attack, the trigger pattern has been separated into multiple parts in multiple compromised clients controlled by the attackers, making it harder to detect. To achieve a higher attack success rate, data poisoning methods are also developed. The global model replacement method aims to manipulate the local training algorithm, and the adversarial updates are guaranteed to be uploaded for aggregation. The attacks at this stage have a better attack success rate, but the requirement of the model training information increases the attack cost. The attack during the aggregation phase, on the other hand, exhibits the highest attack performance. In return, the attackers need to master the local training data, local training algorithm, and the global aggregation protocol, making it hard to implement in practical scenarios.

IV. Latest Defense Methodologies Against Backdoor Attacks

In recent literature, numerous defense methodologies have been put forward to safeguard the resilience of WFL against backdoor attacks. The defense schemes can be classified into four types: defense at the local training phase, defense before the aggregation phase, defense during the aggregation phase, and defense after the aggregation phase. The existing defense schemes are summarized in Fig. 12.

A. Defense at Local Training Phase

The mechanism of backdoor attack defense during the local training phase is illustrated in Fig 13. The input dataset is filtered by the defender to exclude the injected backdoor pattern. The defense methods at this stage primarily focus on data poisoning backdoor attacks. The existing works show reasonable defense performance on label-flipping backdoor attack [186], [187] and trigger-based backdoor attack [188], [189]. In the rest of this section, two typical defense methods implemented during the local training phase are analyzed, and their limitations are discussed accordingly.

1) Backdoor detection by filter libraries: An FL filter and blur-label flipping strategy-based defense algorithm are developed in [39]. The central server collects and stores the commonly seen backdoor types in the filter library. The multiple backdoor filters are trained on the server side with different combinations of eXplainable (XAI) models and classifiers. In each FL round, the server sends the global model and backdoor filters. Each filter is used to detect malicious inputs. Once inputs surpass a pre-determined threshold, they will be identified as backdoor inputs. A blur-label-flipping strategy is proposed to clean the backdoor trigger data area for inputs that are determined as backdoors. The filter will

TABLE VI: Comparison among existing backdoor attacks on WFL: Attack model, adaptivity, required information, and potential exacerbation caused by WCN

Attack Type	Attack Model		Attack Adaptivity		Required Prior Information			Exacerbation by WCN
	Black Box	White Box	Non-adaptive	Adaptive	Training Data	Training Algorithm	Aggregation Protocol	
Label Flipping [29], [123], [148], [149]	√	×	√	×	v	×	×	√
Centralized Trigger [136], [150]–[152]	√	×	√	×	√	×	×	×
Coordinated Trigger [116], [122], [153], [154]	√	×	√	×	√	×	×	√
Local Model Replacement [32], [124], [155], [156]	×	√	×	√	√	√	×	√
Aggregation Process Attacks [37], [157]–[159]	×	√	×	√	√	√	√	√

TABLE VII: Strength and limitation comparison of existing backdoor attacks on WFL

Attack Type	Key Features	Major Limitations	Exacerbation by WCN
Label Flipping [29], [123], [148], [149]	1. Target: flip the target label. 2. No prior knowledge is required. 3. The attack cost is relatively low.	1. It's easy to be detected. 2. The backdoor injected will be quickly diluted.	√
Centralized Trigger [136], [150]–[152]	1. Classify the malicious input into the target label while behaving normally with being input. 2. Only requires part of the local training data. 3. The attack cost is low.	1. The trigger pattern is injected with only one participant. 2. The attack effect will be quickly diluted if the attacker is not selected in every training round.	×
Coordinated Trigger [116], [122], [153], [154]	1. The trigger pattern is distributed to multiple compromised clients. 2. Better stealthiness, faster model convergence speed, and a higher attack success rate.	1. The attacker needs to ensure each part of the trigger pattern is selected enough times. 2. The trigger pattern is fixed and cannot be adaptively adjusted during the training process.	√
Local Model Replacement [32], [124], [155], [156]	1. The attacker fully controls one or multiple compromised clients. 2. The local training algorithm is manipulated to ensure the backdoor is bound to be uploaded to the aggregation process.	1. The attacker needs to master the local training data as well as the local training process. 2. Knowledge of possible defense mechanisms is also required. 3. The attack cost is greatly higher than the black box attack.	√
Aggregation Process Attacks [37], [157]–[159]	1. The attacker manages to modify the aggregation protocol during the global model training. 2. The learning rate and the model weight are adjusted to ensure the backdoor injection.	1. The attacker needs to acquire sufficient information about the whole framework, including the number of participants, the training algorithm, the aggregation protocol, possible defense mechanism, and the convergence process of the global model. 2. The attack cost is the highest.	√

replace the suspicious data in the targeted area with its average among its neighbors. Once the suspicious data has been filtered, it is reintroduced into the training process to verify the successful elimination of the backdoor effect. Using the defender developed in [190], trigger-based and label-flipping backdoor attacks can be defended. After the blur operation, the suspicious data will be used again instead of being discarded. The data availability is restored. The filter also shows good defense ability against coordinated trigger-based backdoor attacks.

Limitations: However, training the backdoor filter on the server side requires prior information about the backdoor patterns and backdoored data sets. The assumptions are made that the attacker only conducts the black-box attack and the knowledge of the local model training process is not obtained by attackers. Moreover, the blur operation can only deal with the trigger in the pixel pattern. Thus, the proposed defense method can only be applied in image classification-oriented WFL.

2) Dropout mechanism during the SGD process: The stability-based defense method against backdoor attacks under the WFL framework is discussed in [191]. It has been discovered that the probability that the trained model correctly classifies the normal inputs (Clean ACC) is proportional to the probability that the trained model misclassifies the backdoored inputs (Backdoor ACC). Furthermore, a turning point exists after which the Clean ACC remains high and the Backdoor ACC reduces rapidly. The turning point is determined by manually injecting backdoor data sets into the original model to test the Clean ACC and Backdoor ACC, respectively. Once the turning point (or threshold) is determined, a dropout mechanism is developed during the SGD process. The gradient weight of the suspicious fraction is set to zero, and thus, the effect of the backdoor is eliminated. This work discusses that the turning points between Clean ACC and Backdoor ACC are similar among different types of backdoor attacks. Thus, the

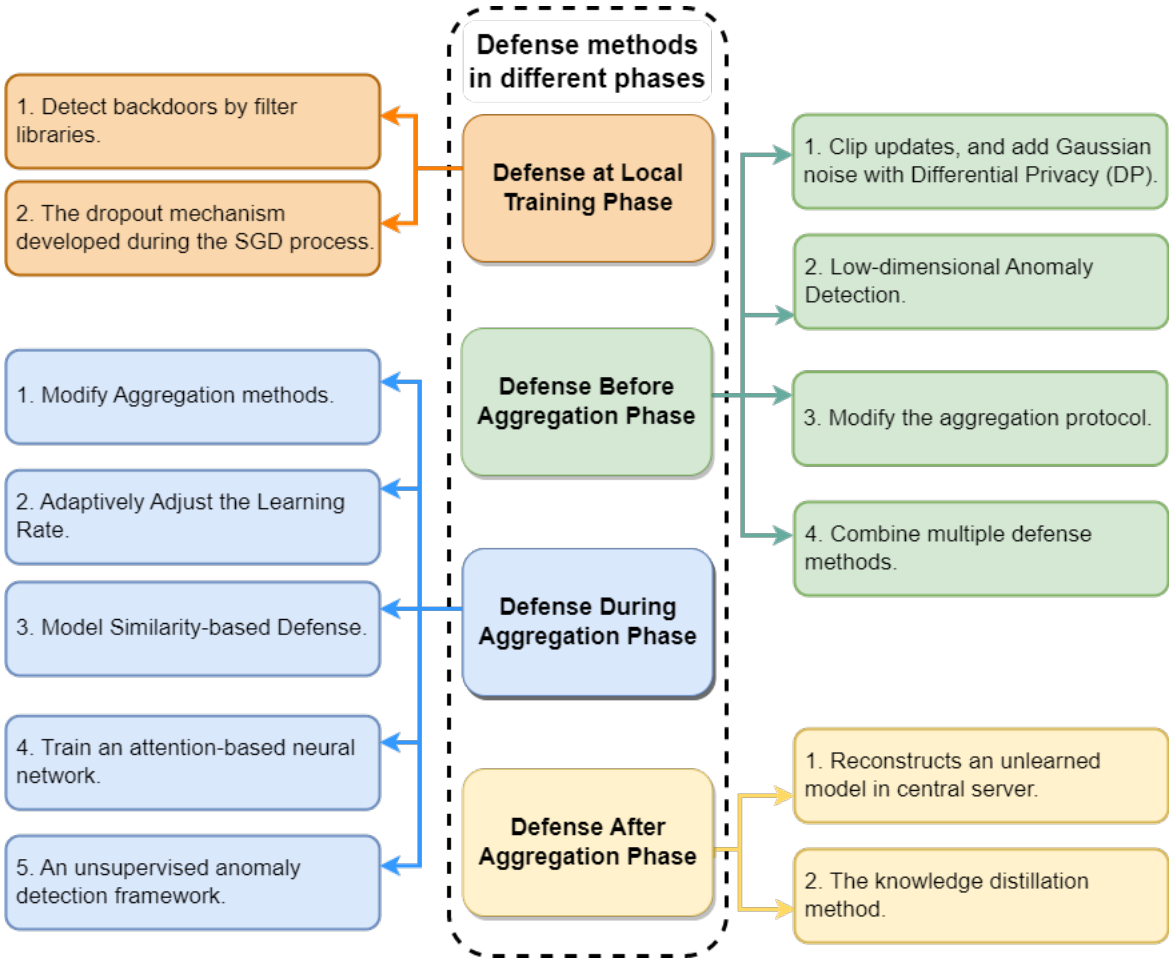


Fig. 12: A summary of defense methods for backdoor attacks on WFL.

defense algorithm proposed in [191] can be applied to multiple scenarios. The backdoor attack success rate under the defense mechanism reduces to zero, while the stability-based defense can ensure a high success rate for the main task.

Limitations: In the experimental setup, the number of participants is set to 10, and the training data are IID. The defense performance might degrade in the actual WFL, where the number of clients is larger, and the training data sets are typically non-IID. In the existing defense method during the local training phase, a local model without an injected backdoor is trained, and the black box backdoor attacks can be well defended. The defense mechanisms at this stage are also compatible with the SecAgg protocol. The remaining challenges are that the defender requires the proper knowledge of the backdoor patterns, and the trade-off between the defense performance and significant task accuracy must also be considered [192].

3) Remark: Wireless devices can be resource-constrained, e.g., in terms of energy supply and communication resources. Client selection and scheduling are indispensable in the context of WFL. To date, the client selection/scheduling has been designed in parallel to the backdoor defense studies. There is a significant opportunity to couple the designs of both in a holistic way to improve defense effectiveness and

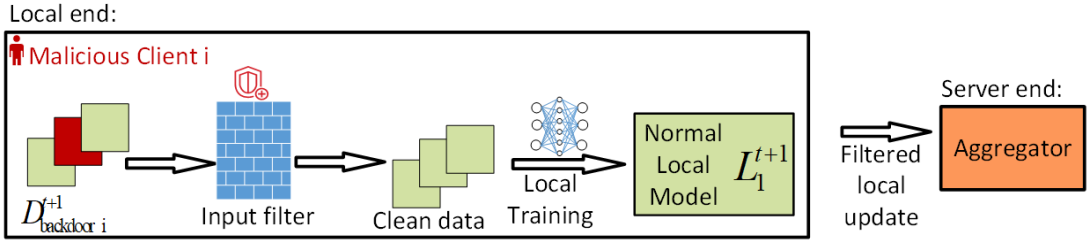


Fig. 13: Defense at the local training phase: The input dataset is filtered by the defender to exclude the injected backdoor pattern.

efficiency. For instance, clients with excellent instantaneous channel conditions and computing power availability can be selected to conduct local model training and detect backdoors effectively. Fairness measures, such as proportional fairness and max-min fairness, could also be considered when jointly designing the backdoor defense and WFL operations.

B. Defense Before Aggregation Phase

Another commonly adopted defense strategy against backdoor attacks is to detect the anomaly updates before sending them to the aggregator. The process of defending the backdoor attack before the aggregation phase can be seen in Fig. 14. The defense methods detect the suspicious local model updates before sending them to the server for aggregation. Many efforts have been carried out to deal with both data poisoning backdoor attack [193]–[195] and model poisoning backdoor attack [196]–[198]. The rest of the section reviews multiple widely studied defense methodologies and discusses their limitations.

1) Clip updates, and add Gaussian noises with Differential Privacy (DP): Compared with benign clients, the backdoor attackers are likely to upload updates with relatively larger norms. Thus, it is a reasonable defense that the central server to scale the client updates ${p_i}^{t+1}$ when its $L2$ norm $||{p_i}^{t+1}||_2$ is greater than the predetermined threshold C. However, the scaled local model updates ${p_i}^{t+1'}$ for each client can be formulated as:

$$p_i^{t+1'} = \begin{cases} p_i^{t+1}, & \text{if } \left\|p_i^{t+1}\right\|_2 < C; \\ \frac{C}{\left\|p_i^{t+1}\right\|_2} p_i^{t+1}, & \text{if } \left\|p_i^{t+1}\right\|_2 \geq C. \end{cases} \tag{9}$$

The scaling mechanism ensures that the norm of each model update is small [124].

After clipping updates, the differential privacy technique can be further applied to defend against backdoor attacks. Although DP-based methods primarily focus on protecting the privacy of individual client data, they still indirectly help to mitigate certain types of backdoor attacks. By adding properly crafted noises to the model updates, DP can make it more challenging for malicious clients to inject poisoned information into the updates without being detected.

Limitations: In a traditional model train with DP, the amount of noise added is relatively large to ensure privacy and security [199]. However, in the scenario of defending against backdoor attacks on WFL, although considerable noise injected can contribute to eliminating the effect of backdoors, it might also ruin the convergence of the main task. Thus, the trade-off between the primary task accuracy and attack defense rate must be considered when adding the noise.

Another limitation of this type of defense mechanism is that once the attacker knows the number or the pattern of the clipping threshold, the attacker can modulate the malicious updates to trick the defender and thus implement the backdoor. It also can be found that the simple norm clipping method shows less effectiveness against trigger-based backdoor attacks since the injected trigger pattern has limited influence on the uploaded norm compared with label-flipping-based backdoor attacks. It has also been pointed out in [38] that the DP-based defense method is less robust when the attacker conducts a multi-round attack.

In WFL, the impact of network communication protocols can be non-negligible. For instance, the TCP can ensure the reliable delivery of data across a network but requires one or multiple retransmissions in each communication cycle [200]. This can penalize the local models by reducing the local training time in a communication cycle. For this reason, the UDP is more commonly adopted in WFL [201]. However, the data distortion and packet loss caused by network communications need to be considered when designing the norm clipping threshold.

Due to the feature of WCN, the communication channels with better signals can be better selected in WFL. The clipping should filter not only the updates from clients but different communication channels as well. The ratio of information gathered from remote network nodes should be considered to prevent backdoor effects and ensure fairness.

2) Low-dimensional anomaly detection: Instead of simply using the norm clipping defense method, several low-dimensional anomaly detection methods have been proposed [202], [203]. The basic idea of low-dimensional anomaly detection methods is to project the compact computed local model onto its low-dimension space, as given by:

$$p_i^{t+1}(L) \rightarrow p_i^{t+1'}(l), \tag{10}$$

where L and l are the numbers of model features before and after the dimension reduction, respectively, and $L > l$. This considers the fact that the low-dimensional embeddings of the uploaded model updates still contain the crucial features

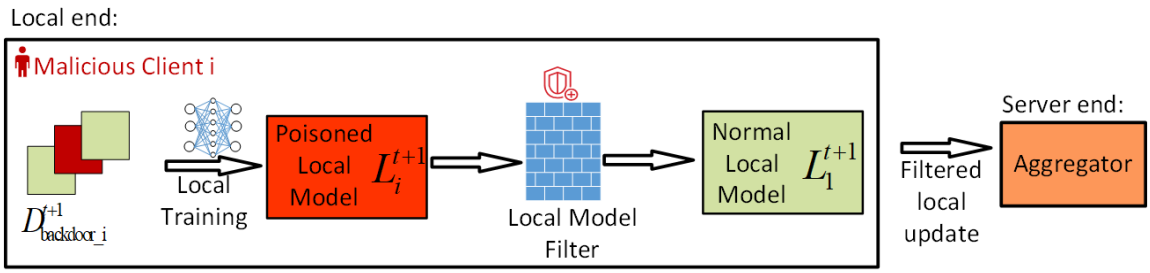


Fig. 14: Defense before the aggregation phase: Filter out the suspicious model updates at the local end before sending them to the aggregator.

that capture the essential variability in data instances [52]. After removing the noisy and redundant features of these data instances, the difference between benign and malicious updates becomes apparent.

One method is to train an encoder-decoder model to approximate low-dimensional embeddings [202]. The unbiased model updates are selected as the inputs of the encoder, and the output of the encoder is the low-dimensional embeddings. The output generated by the encoder is subsequently utilized as the input for the decoder, facilitating the process of reconstructing the original model updates. The encoder-decoder model is then trained to converge by minimizing the reconstruction error, and the converged encoder-decoder model can be adapted to test the practical model updates. The reconstruction error of the benign updates is much smaller than that of the malicious updates. In [203], a Principal Component Analysis (PCA) dimensionality reduction-based defense mechanism is proposed. In each round of a WFL process, a subset of the uploaded model updates is selected and pre-aggregated. The difference between the pre-aggregated and global models is computed, and the parameters in the computed difference corresponding to the predicted probabilities are extracted and stored in a list in the aggregator. After multiple rounds, the list is standardized by removing the mean and scaling to unit variance. The standardized list is then used to input the PCA to reduce the dimension of the updated data and visualize the class pattern. The biggest cluster is recognized as the benign update, while the malicious updates lie within a visibly different cluster. Once the updates from the attackers are detected, the server can ignore their updates and restrict their participation in future rounds.

Limitations: The encoder-decoder architecture requires extra unbiased updates to train the detector, and it is pretty hard to implement in real WFL applications. The PCA-based method requires prior knowledge of the targeted backdoor label, which is also hard to know in practical defense implementation. Meanwhile, the above-mentioned methods can only distinguish label-flipping-based backdoor attacks and cannot deal with trigger-based backdoor attacks. The low-dimensional methods are also incompatible with the SecAgg protocol, and the privacy of the participants in the WFL network cannot be guaranteed.

3) Modify the aggregation protocol: In the WFL framework, the training data sets are non-IID, and the SecAgg protocol does not allow the server to check the individual updates from certain clients. However, the benign updates exhibit similar distributions on average, while the malicious updates exhibit outliers. It is also effective in developing defense mechanisms by modifying the aggregation protocol. In [204], a new aggregation protocol, PartFedAvg, is proposed to defend against data poisoning attacks. In every training round, each client needs to execute an additional random update selection step. The local model update ${p_i}^{t+1}$ has the same multi-layer structure as the global model, indicating the updates regarding different features of the model. Thus, in the PartFedAvg protocol, a fraction of features from the local update is set to zero and uploaded together with the remaining unchanged values. This can be formulated as:

$$p_i^{t+1'} = \begin{cases} p_i^{t+1}[j], & \text{if the } j\text{-th feature is selected;} \\ 0, & \text{if the } j\text{-th feature is not selected.} \end{cases} \tag{11}$$

Then, the server receives the updates provided by the clients and aggregates the global model update. Assuming in each round, m local clients are selected for computing model updates, and the total number of model features is l, the aggregation of the global model update can be expressed as:

$$p_g^{t+1}[j] = \begin{cases} \frac{\eta}{z^{t+1}[j]} \sum p_i^{t+1}[j], & \text{if } z^{t+}[j] \neq 0; \\ 0, & \text{if } z^{t+1}[j] = 0. \end{cases} \tag{12}$$

The values within each label are summed and divided by their respective count $z^{t+1}[j]$ as:

$$z^{t+1}[j] = z^{t+1}[j] + 1, \quad \text{if } p_i^{t+1'}[j] \neq 0 \quad \text{and} \quad 0 < i < m. \tag{13}$$

The calculated global update p_g^{t+1} is then employed to update the global model. The proposed method ensures the privacy of the participants involved. The malicious attacker cannot inject completed backdoor patterns in a short time, thus improving the robustness of the global model.

To better defend against both data poisoning backdoor attacks and model poisoning attacks, the Meta-FL is developed in [205]. In each round, the central server selects subsets of the participants. Each subset computer updates. Then, each subset follows the SecAgg protocol and sends the cryptography-masked updates. Although the server cannot access each client's update, it can migrate the effect of backdoor attacks by ignoring visibly different subset updates. In such a design, the participant's privacy is ensured, and it can be applied to defend against label-flipping-based, trigger-based, and model replacement-based backdoor attacks.

Limitations: The defense schemes proposed in [204] and [205] require modifying the local training process, and the backdoor effect can only be suppressed and cannot be entirely eliminated. The method requires the number of malicious attacks to be smaller than 50% of the total selected clients in each round.

4) Combine multiple defense methods: It is reasonable to combine multiple defense methods to address the limitations of individual defense mechanisms to improve reliability. The WFLAME framework proposed in [38] combines the model clustering, weight clipping approach, and the noise injection method is utilized to mitigate the backdoor effect while simultaneously preserving the benign performance of the aggregated model. In each round, the server receives the updates from the clients and uses HDBSCAN to identify and remove suspicious model updates. A dynamic weight-clipping approach is applied to deal with the boosted malicious model updates. The clipped model updates are then injected with adaptive noise to further improve robustness. At last, the modulated updates are aggregated with the same weight. In this design, the HDBSCAN clustering algorithm is dynamic and adaptive, and the clipping threshold and the amount of noise are also adaptively designed. It shows good defensive

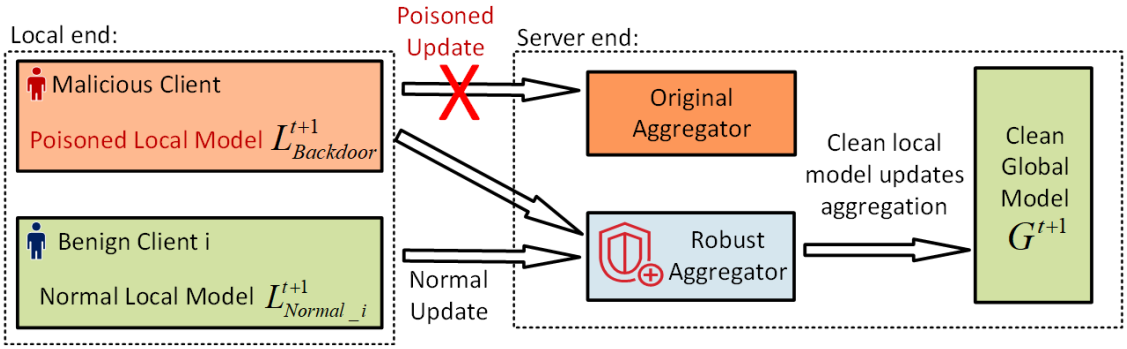


Fig. 15: Defense during the aggregation phase: A robust aggregator is designed at the server end to distinguish and reject malicious updates from participants.

performance against both data poisoning and model poisoning backdoor attacks.

Limitations: Implementing the WFLAME framework proposed in [38] requires massive and computationally expensive modification of the whole WFL network, which brings challenges in practical scenarios.

On the other hand, the broadcast nature of wireless channels gives rise to an opportunity for collaborative backdoor detection. For example, the clients are likely to overhear the local model updates from each other and can potentially evaluate those local model updates based on their own local datasets. Consensus [206] among the clients can be potentially used to select reliable local models. Byzantine fault tolerance techniques [207] can be potentially leveraged to identify misbehaved clients and maintain the integrity of WFL.

C. Defense During Aggregation Phase

In many recent works, the defense execution stage is moved from individual updates to the aggregation level. Multiple methodologies are carried out to filter out the poisoned updates during the aggregation phase. An overview of defense during the aggregation phase is presented in Fig. 15. Considering the fact that the aggregator, which simply combines the received updates, is vulnerable to being poisoned, defenders design a robust aggregator based on statistical methods [194], [208]. By calculating the statistical features of model updates, including average and median, one or more participants are selected as the reference to detect the malicious updates [193], [209]. The feasibility and effectiveness of defense during the aggregation phase have been proved in the existing works when facing model poisoning backdoor attacks [210]. In this section, multiple defense methods designed to take effect during the aggregation phase are studied, and their limitations are then analyzed.

1) Aggregation protocol modification methods: In [211], a classic repeated median regression algorithm is extended to the WFL applications. The repeated median regression is computed for each label when the server receives the model updates. The residuals at each dimension are also obtained. Standardizing these residuals yields the standard deviation, based on which the parameter confidence of each label is derived. Then, the extreme values are also corrected to eliminate the effect of boasted malicious parameters. The k-th updated models can then be re-weighted based on the feature importance:

$$W^{(k)} = \sum_{j=1}^{l} \omega_j^{(k)} \sigma(\omega_l), \tag{14}$$

where $\omega_j^{(k)}$ is the confidence of the j-th feature within the model, and $\sigma(\cdot)$ denotes the standard deviation. A feature with a large standard deviation implies that the value of this feature varies greatly among all participants, which needs to be more critical when sending for aggregation. Subsequently, the global model update can be aggregated as:

$$p_g^{t+1} = \sum_{k=1}^{m} \frac{W^{(k)}}{\sum_{i=1}^{m} W^{(i)}} p_k^{t+1}. \tag{15}$$

The proposed method is featured for defending label-flipping backdoor attacks and model replacement backdoor attacks.

Limitations: The aggregator requires the knowledge of individual model updates, which violates the SecAgg protocol. The proposed method performs poorly in the non-IID situation and cannot ensure privacy and security. Thus, it is hard to implement in practical WFL scenarios.

2) Adaptive adjustment of the learning rate: Different from modifying the aggregation method, the studies conducted in [212], [213] aim to adaptively adjust the learning rate to improve the robustness of the WFL framework. The distribution of the sign function of each label is obtained. A hyperparameter θ is then defined to determine the threshold of the learning rate. The adaptive learning rate for each model feature can be formulated as:

$$\eta_{\theta,i} = \begin{cases} \eta, & \text{if } \left| \sum_{k=1}^{m} \operatorname{sgn}\left(p_k^{t+1}[i]\right) \right| > \theta; \\ -\eta, & \text{otherwise.} \end{cases} \tag{16}$$

If the sum of signs of the i-th feature among all local model updates is less than θ, the sign of the learning rate is reversed to maximize the loss on that specific feature, instead of minimizing it. This method shows reasonable defense performance against data poisoning attacks [214].

Limitations: Similar to the aggregation modification method, the proposed defense mechanism requires the individual to

update information from the participant on the server side. Thus, it is not compatible with the SecAgg. The participants' privacy security cannot be guaranteed.

3) Model similarity-based defense: Another commonly adopted strategy of defense during the aggregation phase is the model similarity-based defense method [215]–[217]. The FoolsGold is proposed in [218], and the adversarial attack is detected based on the diversity of the model updates. In a WFL framework, the benign client's updates tend to have richer diversity, while the adversary normally behaves similarly. By comparing the uploaded update with its historic stored values, adversarial models with similar updates can be detected. In each round, each update is first compared with its historical value, and the similarity is computed. The similarities of the updates from different clients are also computed to reduce misdetection. The backdoor attack can be successfully detected by filtering out the outlier of the updates and adaptively adjusting the learning rate. This design has no constraints on the number of malicious clients, and it demonstrates reasonable performance against data poisoning attacks.

Limitations: The methods proposed in [215]–[218] assume that the attacker should be malicious all the time. The attacker needs to upload adversarial models in every round. In practical scenarios, the attacker can decide to inject the backdoor at a random round. The proposed method cannot detect it without sufficient accumulated historical malicious updates.

4) Defense using attention-based neural networks: In [219], the focus is on training an attention-based neural network. The central server simulates the WFL tasks under various types of backdoor attacks. Then, the simulated model updates are collected to train a separate neural network model, which learns the potential vulnerability of the global model against different backdoor attacks. Thus, the global model shows a self-supervised fashion.

The attention-based neural network is then applied to predict the actual updates received from the participants. The likelihood of a benign update received can be parameterized as:

$$s_i = \frac{Q\left(q_t\right) K\left(p_i\right)}{\|Q\left(q_t\right)\| \|K\left(p_i\right)\|}, \tag{17}$$

where $Q(q_t)$ is the query encoder from the trained defender, and updated with the actual model updates. If the alignment score s_i is closer to +1, p_i is more likely to be a benign update. On the other hand, if s_i is close to -1, p_i is likely to be an adversary update. This method shows better defense performance than traditional median-based aggregation and residual-based re-weighting aggregation. By training the neural network with new backdoor attack data sets, the self-supervised defense mechanism can defend it properly [107].

Limitations: The major problem of this defense method is that to distinguish the malicious updates, all possible backdoor attack patterns must be trained in the server before implementation. During the detection, the aggregator requires access to single local updates. Thus, the privacy of the participant is at risk of leaking.

5) Unsupervised anomaly detection-based defense: Considering the fact that the backdoor attack effect is discernible in terms of model weight, an unsupervised anomaly detection framework, ARIBA, is proposed in [220]. In this design, the model weights are pre-processed to obtain the discernible patterns of filters. Multiple filters are then collected and fed into an unsupervised anomaly detection algorithm to detect the suspicious filters. For example, when the server receives m local model updates, a set of ϕ filters, denoted as $f_{i,j}$, $0 < i < m$, $0 < j < \phi$, are used for the detection via the Mahalanobis-distance-based algorithm [221]. The detected suspicious filters are applied to compute the anomaly score of each client, as given by:

$$s_{i,j} = \begin{cases} 0, & \text{if } f_{i,j} \text{ is benign;} \\ 1, & \text{if } f_{i,j} \text{ is adversary.} \end{cases} \tag{18}$$

As a result, the malicious client can then be identified accordingly [222]. It is assumed that the attacker knows all possible defense patterns, can conduct attacks at any time, and multiple attackers can conduct coordinated attacks. The proposed method can defend against multiple types of backdoor attacks with properly trained weight filters without degrading the benign performance.

Limitations: The unsupervised anomaly detection methods proposed in [220] and [222] are incompatible with the SecAgg protocol. Thus, data privacy and security can hardly be ensured.

In addition to the above-mentioned limitations, the emerging OTA-WFL may pose new challenges to the existing defense mechanisms designed for the aggregation phase due to the fact that the FL server would only access the aggregated global model under OTA-WFL [14], [101]. The local models of individual clients are obscured. While this helps improve the privacy of the individuals' local models to a great extent, it allows backdoor attacks to take place unnoticed. Few studies have been concerned about this particular issue. For example, the authors of [192] attempt to restore all the individual models using serial interference cancellation. To what extent can the restored local models be assessed to detect backdoor attacks remains unclear, as the local model restoration process may distort the local models and affect the backdoor detection accuracy.

D. Defense After Aggregation Phase

Several defense methodologies have been proposed to deal with the injected backdoor effect that remained in the global model after the aggregation phase —- The idea of defense after the aggregation phase can be seen in Fig. 16. The server will store the historic value gathered from participants to train a reference global model, based on which the trained converged WFL global model is examined to filter the anomaly contribution from suspicious clients. By excluding the influence of suspicious clients, the backdoor effect will be entirely eliminated. The defender at this stage can well handle the model poisoning backdoor attacks, including model replacement backdoor attack [37], [208], [223] and weight scaling backdoor attack [38], [194].

The post-aggregation backdoor defense mechanisms, i.e., performed upon the aggregated global models, can be particularly important for the emerging OTA-WFL systems to

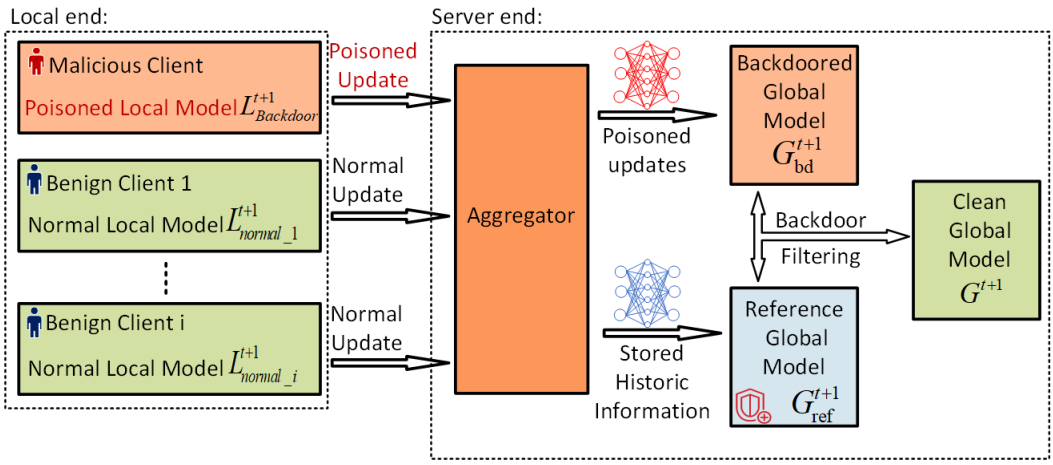


Fig. 16: Defense after the aggregation phase: The server will store the historical value gathered from participants to train a reference global model for anomaly detection.

circumvent the obscurity of the local models in the OTA-WFL processes. This section reviews two typical defense designs after the aggregation phase. Their limitations are then discussed.

1) Reconstruct an unlearned model in a central server: In [224], the federated unlearning methodology, namely, FedEraser, is proposed. In this scheme, the central server stores the historical updates received from clients during the global model training process. Once the global model converges, the FedEraser reconstructs an unlearned model in the central server based on the stored historical parameters. During the unlearned model training process, the benign clients are defined as the calibrating clients. In each round, the central server collects the updates from calibrating clients concerning the calibrated global model. Then, the FedEraser aggregates these updates and renews the global model to be an unlearned WFL global model. By comparing the performance of the unlearned model on clients' data with that of the original global model, the backdoor effects can be well eliminated [225]. In the FedEraser design, no substantial modification is required to the existing WFL architecture and the training process of the participants. Thus, the proposed method is easy to implement. In addition, the FedEraser trains the unlearned model using the parameters stored during the Federated global model training process. It is claimed that the training speed of constructing the unlearned model in FedEraser is four times faster than training from scratch.

Limitations: The calibration process during the unlearned model training relies on the historical updates of the participating clients. The limited store capability of the hardware makes it hard to deploy in practice. In addition, in each round of federated model training, the user's local gradient is randomly selected and uploaded to the server to train the global model. Thus, the reliability of the calibration of the global model using the unlearned model trained by the historical parameters remains discussed. On the other hand, although the training data and the training process are not required, the FedEraser must target the attacker client, which is also pretty hard in actual implementation.

2) Knowledge distillation method: The knowledge distillation method is proposed in [216], [226] to restore the global model's performance. In this design, the central server stores the historical parameters of the participating clients. Once the global model is obtained. The contribution of the targeted malicious clients is subtracted, and a skewed unlearning model is derived. Then, the knowledge distillation is performed. The original model is defined as the teacher model, while the skewed unlearning model is defined as the student model. The teacher model produces class predictions that are utilized to label the dataset employed for training the student model. The concept of temperature is introduced in this design. The higher temperature produces a more ambiguous probability distribution, while the lower temperature produces a more discrete probability. The student model is trained under high temperatures to compensate for the skew. The knowledge distillation training process is executed on the server side. Thus, no extra information about the client and communication between the client and the server are required. Due to the fact that the data set used in the distillation does not contain backdoor patterns, the student model will not learn the backdoors injected in the teacher model. It is also testified that the proposed knowledge distillation-based unlearning method can improve the robustness of the global model and defend against all types of backdoor attacks [227].

Limitations: The unlearning mechanism requires the knowledge of the target clients, and the attackers are normally stealthy in practical scenarios. Meanwhile, the proposed method is not compatible with SecAgg, and the privacy of the participants is at risk of leakage.

E. Summary

Overall, the existing defense schemes are analyzed in each type and summarized in Table VIII. The existing defense methods can be broadly divided into four categories: defense at the local training phase, defense before the aggregation phase, defense during the aggregation phase, and defense after the aggregation phase. Defense methods at the local training

phase are implemented on the client side, aiming at the data poisoning attacks. The defense cost is relatively low, and the defense mechanism at this stage is compatible with SecAgg. Defense before the aggregation phase is also mounted on the client side. The defender is keen to exclude suspicious updates before sending them to the aggregator. Some methodologies employ the modification of SecAgg, thus making it hard to be configured in practical WFL networks. The defense during the aggregation is capable of identifying the backdoored updates from the malicious clients and excluding them from rejecting backdoor injection. The defenders are located on the server side, and normally, they are not compatible with SecAgg. The defense after the aggregation phase will evaluate the converged global model to test whether it has been polluted by the backdoor pattern or not. The defense mechanism is also at the server side, and the historical information gathered from the participant is necessarily needed.

V. Performance Evaluation and Comparison

In this section, the performance of backdoor attack strategies and defense methods in each phase are presented, followed by a comparative analysis.

A. Comparison Among Backdoor Attack Strategies

Four representative backdoor attack strategies [29], [32], [37], [116] on WFL in each phase (as shown in Table IV) are firstly compared. The experimental setting is given in Table IX. The global model is trained on the FMNIST dataset. It is assumed that there are a total of 100 participants, and in each training iteration, the server randomly picks 10 participants for the global model update. The number of compromised clients controlled by malicious attackers increases with relatively more complex attack methods. It is also assumed that the malicious clients can be continuously selected in multiple rounds. The trigger pattern used in this case follows the original method, as discussed in [116]. To make it easier to follow, a GitHub entry has been created for the reader's information[3].

The evaluation metrics used in this case are BAR and MAR, where BAR denotes the probability that the global model is misled to classify backdoor inputs into the target label, and MAR evaluates the prediction accuracy of the backdoor-injected global model on clean inputs. Generally, a higher BAR means the backdoor effect is better performed. In contrast, for backdoor attacks on WFL, the MAR has been commonly used to evaluate the stealthiness of the injected backdoor pattern. The results are shown in Figs. 17 and 18, and Table X.

The experimental results further verify the conclusions in Sections III and IV. It can be seen that the global model replacement backdoor attack [37] launched during the aggregation process achieves the highest BAR while also maintaining a reasonable MAR. However, the attacker must obtain sufficient knowledge of the whole WFL network, introducing a considerable attack cost, as discussed in Section III-C. On the other hand, the local model replacement backdoor attack

[3]https://github.com/DrQu89757/backdoor_attack_in_WFL

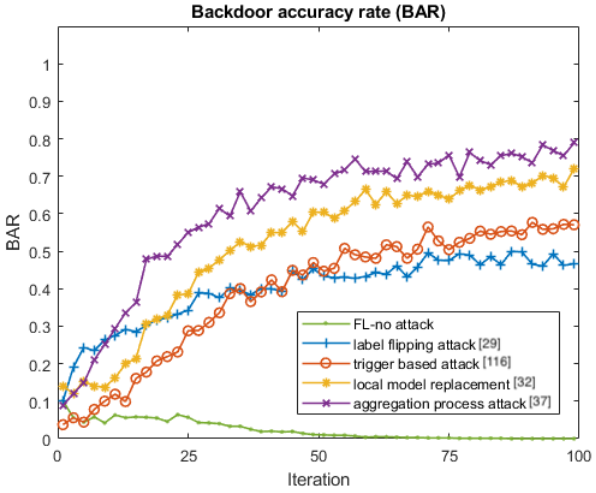


Fig. 17: Backdoor accuracy rate of [29], [32], [37], [116].

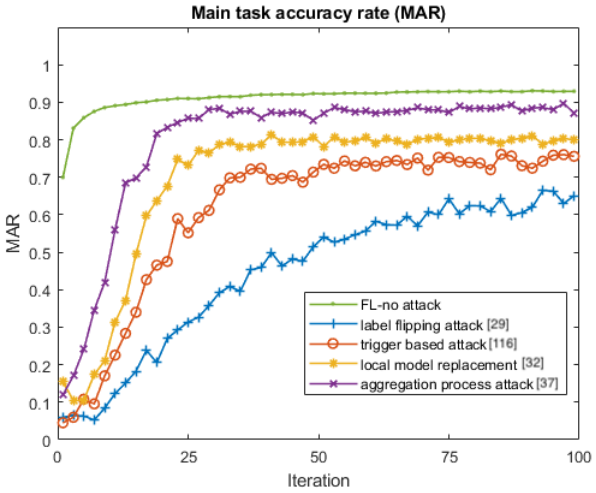


Fig. 18: Global model main task accuracy rate of [29], [32], [37], [116].

during the local model training phase shows the second-best performance. In general, the model poisoning-based backdoor attacks outperform data poisoning-based backdoor attacks while yielding a higher complexity. The low BAR of trigger-based backdoor attacks could be due to insufficient CCR. Since the completed trigger pattern is distributed in multiple compromised clients, low level of malicious participation leads to low BAR. The performance of the label-flipping-based backdoor attacks depends on the selection of the target label. Based on [33], this type of backdoor attacks is anticipated to perform better when targets are on the edge label. Consequently, it is more unlikely to be detected, and the attack strength is relatively weaker.

To further evaluate the longevity of different attack strategies, the lifespan of the injected backdoor pattern is considered. With the backdoored global model obtained in the previous step, they are further trained with the clean dataset and honest participants for 50 more iterations, and the change of MAR under each attack method is recorded in Table XI.

TABLE VIII: Defense methods for backdoor attacks on WFL

Defense Phase	Ref.[1]	Strategy	Detail	Exp. Env.[2]	Comp. with SecAgg?	Main Task	Dataset	Exacerbation by WCN
Local Training	[39]	Poisoned Data Filtering Method.	Federated backdoor filter designed to identify backdoor inputs and restore the data to availability by the blur-label flipping strategy.	Client	√	Image classification	MNIST; CIFAR-10	√
	[228]	Model Stability based Method.	The stability-inducing operations are introduced to improve the generalization of the testing data.	Client	√	Intelligent Edge Computing Service	MNIST	√
Before Aggregation	[124]	Differential Privacy based Defense.	The local updates are norm-clipped, and Gaussian noise is added to the global model.	Client	×	Tensor Flow WFL applications	FMNIST	√
	[202]	Reduced Dimension based Defense.	Spectral anomaly detection mechanism is designed to remove the malicious model updates in a low dimensional latent space.	Client	×	Image classification; Sentiment Analysis	MNIST; FMNIST; Senti-ment140	√
	[115]	Reduced Dimension based Defense.	The defense extracts parameters from the high-dimensional updates and applies PCA for dimensionality reduction to filter out malicious updates.	Client	×	Image classification	CIFAR-10; Fashion-MNIST	√
	[205]	SecAgg Procotol Modula-tion based Defense.	The local clients follow the SecAgg protocol and send the cryptography-asked updates. The backdoor attacks are mitigated on the aggregation level.	Client	√	Image classification	SVHN; GTSRB	√
	[204]	SecAgg Procotol Modulation-based Defense	The partial aggregation protocol is designed to improve the privacy and robustness of WFL.	Client	√	Image classification; Word prediction	MNIST; CIFAR-10; LOAN	√
	[38]	Mulit-Methods Combined Defense	The defense method combines the model clustering, weight clipping approach, and noise injection to achieve adaptive clipping and eliminate the adversarial backdoors.	Client	×	Word prediction; NIDS; Image classification	Reddit; IoT-Traffic; CIFAR-10; MNIST; Tiny-ImageNet	√
During Aggregation	[218]	Model Similarity-based Defense	The defense mechanism can identify attacks based on the diversity of client updates.	Server	×	Classifications	MNIST; VGGFace; KDDCup; Amazon	√
	[219]	Model Similarity-based Defense	The defense mechanism trains an NN with an attention mechanism to learn the vulnerability of WFL models from a set of plausible attacks.	Server	×	Image classification; IMDb sentiment analysis	MNIST; CIFAR; ImageNet	√
	[220]	Model Similarity-based Defense	The defense employs unsupervised anomaly detection to evaluate the pre-processed filters. The malicious clients can then be identified according to their anomaly scores.	Server	×	Image classification	MNIST; CIFAR-10; Fashion-MNIST	√
	[229]	Statistical Method	The defense proposes one coordinate-wise median-based and one coordinate-wise trimmed mean-based gradient descent algorithm to filter out backdoors	Server	×	Tensor-Flow tasks	MNIST	√
	[211]	Statistical Method	A novel aggregation algorithm with residual-based reweighting is proposed, which combines repeated median regression with the reweighting scheme in iteratively least squares.	Server	×	Image classification; Word prediction; Loan prediction	MNIST; CIFAR-10; Amazon; LOAN	√
	[212]	Statistical Method	The defense adjusts the aggregation server's learning rate per dimension and round based on the sign information of local client updates.	Server	×	Image classification	MNIST; FMNIST	√
After Aggregation	[224]	WFL with Forgetting Mecha-nism	The FedReaser is proposed to store the historical updates of federated clients and re-train the unlearned model in the central.	Server	×	Image classification; Word prediction	UCI Adult Purchase; MNIST; CIFAR-10	√
	[226]	WFL with Forgetting Mecha-nism	A Federated unlearning method is proposed to subtract the accumulated client updates and leverage the knowledge distillation method to restore the model performance without local privacy information.	Server	×	Image classification	MNIST; CIFAR-10; GTSRB	√
	[230]	Global Model Test Method	The defense proposed one feedback-based WFL to detect backdoors by leveraging data of multiple clients.	Server	√	Image classification	CIFAR-10; FMNIST	√

[1] Reference paper that belongs to the specific group.
[2] Experiment environments, including server or clients.

TABLE IX: Experimental settings for the attacks

Attack strategies	Attack type	Attack phase	No. of participants	CCR	No. of training iterations	Dataset
Label flipping based attack [29]	Data poisoning backdoor attack	Local data collection phase	100	0.1	100	FMNIST
Coordinated trigger based attack [116]	Data poisoning backdoor attack	Local data collection phase	100	0.2	100	FMNIST
Local model replacement attack [32]	Model poisoning backdoor attack	Local model training phase	100	0.5	100	FMNIST
Global model replacement attack [37]	Model poisoning backdoor attack	Aggregation phase	100	0.5	100	FMNIST

TABLE X: Backdoor attack performance

Attack strategies	BAR(%)	MAR(%)
Label flipping based attack [29]	46.6	65.03
Coordinated trigger based attack [116]	57.06	75.57
Local model replacement attack [32]	71.96	88.91
Global model replacement attack [37]	79.86	91.68
Original WFL-no attack	0	92.92

It is observed that the backdoor effect of data poisoning-based backdoor attacks is quickly diluted as the attacker can only manipulate the training dataset in the local client, so the attack strength is limited. In contrast, model poisoning-based backdoor attacks can inject relatively longer-lasting backdoor effects into the global model. Specifically, the global model replacement proposed in [37] is conducted when the global model is about to converge. The backdoor pattern is more likely to remain in the global model for a longer time.

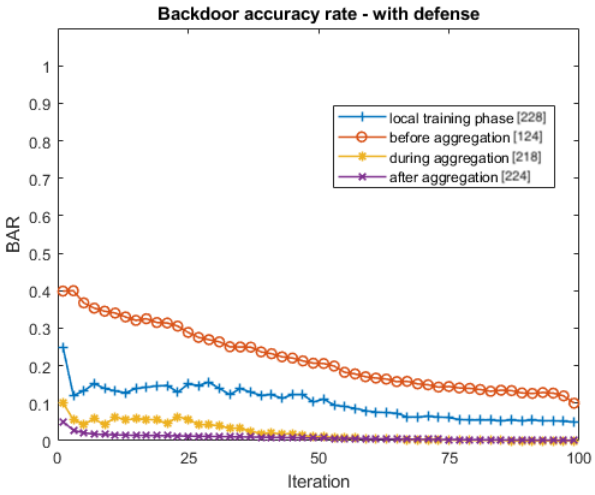


Fig. 19: Backdoor accuracy rate with defense methods [124], [218], [224], [228].

B. Comparison Among Backdoor Attack Defense Methods

In this section, we present representative experimental results to evaluate and compare the performances of existing defense methodologies for backdoor attacks on WFL. The defense methods can be categorized into four phases, as shown in Fig. 12. In each phase, one featured defense method is selected to defend against backdoor attacks of different types, as illustrated in Section V-A. The settings of each defense method are shown in Table XII, where M denotes the norm threshold of the norm clipping, and σ represents the Gaussian noise parameter. The same FMNIST dataset is adopted to train the global model while the settings of each attack method remain consistent. The performance of each defense method is illustrated in Figs. 19 and 20, and Table XIII.

It is noted that all defense methods in different phases can reduce backdoor effects while maintaining a steady MAR. Specifically, the defense methods equipped during and after the aggregation phase have the best performance against model poisoning attacks. It is reasonable to assume that the defense mechanisms working on the server side are more powerful than those on the client side. However, such defense methods [218], [224] require the historical information of local model updates

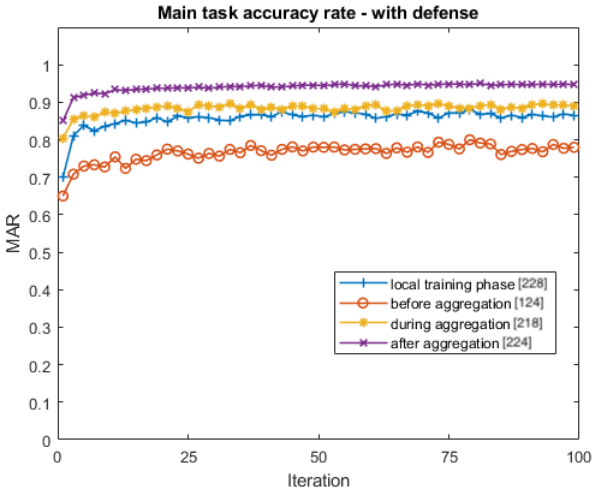


Fig. 20: Main task accuracy rate with defense methods [124], [218], [224], [228].

TABLE XI: Lifespan comparison of attacks

Attack strategies	BAR		
	after 0 iteration	after 10 iterations	after 50 iterations
Label flipping based attack [29]	46.6	15.04	0
Coordinated trigger based attack [116]	57.06	22.82	0
Local model replacement attack [32]	71.96	43.17	12.95
Global model replacement attack [37]	79.86	64.88	49.51

TABLE XII: Experimental settings for the defense methods

Defense type	Defense phase	Defense target	Aggregation Protocol	M	σ	Dataset
Model stability based method [228]	Local training phase	Data poisoning backdoor attack	FedAvg	0	0	FMNIST
Norm clipping and DP based method [124]	Local training phase	Data poisoning backdoor attack	Normclip	1	0.0025	FMNIST
Model similarity-based method [218]	During aggregation phase	Model poisoning backdoor attack	FoolsGold	0	0	FMNIST
WFL with forgetting mechanism [224]	After aggregation phase	Model poisoning backdoor attack	RLR	0.5/1	0.0001	FMNIST

TABLE XIII: Defense performance

Defense type	BAR with defense	MAR with defense
Model stability based method [228]	9.72	78.17
Norm clipping and DP based method [124]	6.59	86.12
Model similarity-based method [218]	1.12	89.53
WFL with forgetting mechanism [224]	0.24	94.76

and are not compatible with SecAgg, which brings more difficulty in the practical implementation. It is also worth mentioning that the unsatisfactory performance of the norm clipping-based defense methods before the aggregation phase heavily depends on the improper selection of the threshold M and noise parameter σ. As pointed out in [124], the norm clipping-based methods inevitably introduce overkill or over-fitting problems. Meanwhile, the Gaussian noise introduced by differential privacy also adds extra disturbance to the global model training process. Thus, it is essential to properly design the defense algorithm's thresholds.

C. Summary

Experiments have been conducted to evaluate and compare the strength of backdoor attacks in different types and the performance of defense methodologies in different phases. The experimental results verify the discussions in Section III and IV. It can be seen that the model poisoning-based backdoor attacks show higher BAR when compared to data poisoning-based backdoor attacks. However, the model poisoning-based backdoor attack algorithms have relatively higher complexity and normally require much more information about the entire network. Many existing defense methodologies have been proven to be capable of defending both data poisoning-based backdoor attacks and model poisoning-based backdoor attacks. Specifically, the defense mechanisms working on the local training phase and before the aggregation phase can address data poisoning-based backdoor attacks. The defense mechanisms deployed during and after the aggregation phase can handle model poisoning-based backdoor attacks well. Although many powerful backdoor attack strategies, such as global model replacement [143], can already be settled, the major dilemma in developing a more convenient defense methodology is how to better defend against all types of backdoor attacks while ensuring the privacy and security of participants (i.e., to be compatible with SecAgg). In the following sections, the lessons learned from existing works are concluded, and further potential research directions are also discussed.

VI. Lessons Learned

In the previous sections, the working mechanisms of existing backdoor attack designs and the defense methods have been discussed in detail. The limitations of each of the attack and defense methods have been discussed. The major findings from the existing works are summarized in this section.

A. Backdoor Attack Methods

In this section, the lessons learned from the existing backdoor attack strategies on WFL are analyzed.

1) Data poisoning-based backdoor attacks: Earlier works on data-based backdoor attacks focused on un-targeted attacks that aimed to degrade the accuracy of the global model by injecting "poisoned" data into the training set [231], [232]. A simple label-flipping method was often used to replace the targeted label [123]. These types of attacks do not require any prior knowledge about the learning model or mechanism. However, the effect of the injected backdoor tends to be diluted over multiple training rounds. Some studies have shown that

the attack success rate (ASR) of label-flipping attacks can be increased when the attacker controls more compromised clients [233], [234]. In contrast, other researchers have focused on injecting the poisoned data into the "tail", or labels with relatively smaller weights [33]. This type of "tail backdoor" can remain undetected for a longer duration but only affects the classification of a subset of the training samples. There is a trade-off between the ASR and the strength of the backdoor effect in this case.

In addition to degrading the performance of the global model, there has also been a growing interest in targeted backdoor attacks for WFL. These attacks aim to have the global model learn a trigger pattern embedded in a poisoned data sample, which can either be imperceptible (e.g., a watermark) or perceptible but innocuous (e.g., glasses on a face). The goal of a targeted backdoor attack is to have the global model classify the poisoned sample into the target class without affecting the accuracy of the primary task. The trigger can be randomly generated or specifically designed for the target label [29], [33], [122]. Depending on the number of compromised clients controlled by the attacker, targeted backdoor attacks can be further divided into two types: centralized attacks [235] and coordinated attacks [116], [122].

One advantage of data-based backdoor attacks is that the attacker does not need to have access to the entire training dataset or knowledge of the training process and mechanism. Additionally, the ASR of these attacks can be improved by increasing the number of compromised clients controlled by the attacker. However, because these attacks only target the training data, their success depends on the sufficient number of attack rounds to avoid the dilution of the backdoor effect [236].

2) Model poisoning-based backdoor attacks: Model-based attacks, on the other hand, target the training process of the local models. In these attacks, the attacker injects a poisoned data sample into the training process along with benign samples and designs a loss function to minimize the difference between the poisoned model and the benign model [32], [126], [191]. This allows the malicious model to be aggregated into the global model without detection. Research has shown that these attacks can achieve a high ASR in edge cases. Another approach to model-based backdoor attacks is to adjust the weight of the poisoned local model to a sufficiently large value before aggregation, ensuring that the backdoor is injected into the global model [32]. The optimal weight must be carefully designed to achieve the highest ASR without detection.

Compared with a backdoor attack for data, a model-based backdoor attack shows better ASR and longer duration [164]. In return, the attack cost of model-based backdoor attacks is relatively higher. The attacker needs to gain complete information about the training data set and the training mechanism. The attack exhibits better performance when the global model is approaching convergence. Thus, the training process and progress are also necessary for the attacker [49].

3) Insights: Data poisoning happens during the local data collection phase, while model poisoning happens afterward. Compared to data poisoning with model poisoning, data poisoning backdoor attack shows a relatively lower attack cost, and model poisoning attack has a better attack accuracy rate and longer backdoor life span. Within the data poisoning attack, label flipping is one of the easiest attack methods, resulting in a lower success rate, and can be easily detected. A trigger-based backdoor is the most adopted one, and with the proper design trigger pattern, the attack performance will be greatly improved.

Once the attack obtains the ability to manipulate the training algorithm or aggregation protocol, one can conduct a model poisoning attack. The compromised client generated the backdoor inject local updates via ether label flipping or trigger pattern. With a higher knowledge level known by the attack, the attack success rate is higher as well. In return, the attack cost will also dramatically increase.

B. Backdoor Defense

There are many defense mechanisms proposed to detect and defend the backdoor attack. Based on the complexity and different defense phases, which can be categorized into four types as shown in Fig. 12.

The defender usually clips the contribution from suspicious clients to get rid of the backdoor effect. In this process, some honest information will be miss-clipped, and the main task accuracy will be influenced. Meanwhile, some defense methods attempt to train a unique classifier that can distinguish the adversarial input. Such a newly trained classifier requires information on all possible backdoor patterns, and the defense cost is high. Third, the WFL framework is featured for its non-IID property. The defender is hard to trace back to the compromised client and leaves it out of the network. And there is no proven solution to deal with the injected backdoor pattern after the aggregation phase.

It is also worth mentioning that there is a common trade-off among all existing defense methodologies: the defense success rate and the convergence rate of the global model. Since all defense methods inherently eliminate the suspicious backdoor effect by clipping the possible malicious updates, inevitably, a portion of honest updates will be neglected by mistake. Thus, it's always a top priority in designing a defense mechanism: the backdoor effect should be appropriately removed, and the global model can quickly converge at the same time.

VII. Challenges, Open Problems, and Future Research Directions

It is foreseeable that Wireless Federated Learning (WFL) techniques will play a pivotal role as one of the core components in the development of next-generation wireless communication networks, including Beyond 5G (B5G) [237] and 6G [238]. To deliver intelligent network operation and efficient resource management with practical WFL applications in the wireless communication network, it is of great significance to develop a WFL framework that is secure and robust to various kinds of backdoor attacks. Thus, the study of backdoor attacks and defense mechanisms in WFL becomes essential. In recent years, this area has been a hot research topic in both academics and industries.

Many research efforts have been devoted to the development of both backdoor attack methodologies and backdoor attack

defense mechanisms. The objective of backdoor attacks is to dig deeper into the information flow process in the WFL network and amplify the impact of any potential insecure vulnerabilities within the WFL framework. In terms of backdoor attack designs, researchers focus on how to improve the attack success rate. These efforts aim at emphasizing the severity of the problem. More research attention can then be addressed to design the targeted defense mechanism. The ultimate goal of defense mechanism design in WFL is the ability to better deal with the actual malicious behaviors that occur in practical applications. Via a secure and robust WFL network, many WFL applications can be deployed with confidence.

In the previous sections, the review of existing breakthroughs in backdoor attack strategies and defense methods has been presented, based on which many critical challenges and open issues in WFL have been revealed. In this section, these challenges and open issues are analyzed from the viewpoints of both backdoor attack methods and backdoor attack defense mechanisms. The potential solutions are discussed afterward. The discussions are summarized in Fig. 21.

A. Backdoor Attacks on WFL

As reviewed in Section III, backdoor attacks on WFL can be classified into three categories: those that occur during the local data collection phase, the local model training phase, and the server aggregation phase. These attacks are designed based on the attackers' prior knowledge, and each has its own limitations. The open issues related to these attack methods are discussed in this section.

1) How to decrease the data-dependency of data poisoning backdoor attacks: The backdoor attack during the local data collection phase requires the attacker to control or have access to part of the local training data set [239]. In order to guarantee both the attack success rate and the local model convergence, the backdoor pattern injected in the poisoned dataset shows dependency on the benign dataset, i.e., the poisoned dataset cannot have a huge difference from the honest ones. It has been proved that in the edge-case (or low-possibility sample), the backdoor-injected poisoned dataset can be irrelevant to the label description. For example, in the image classification application, the images of "Southwest airplane" are fed to the label of "truck" [33], [240]. Such implementation does decrease the dependency on the backdoored input design. The attack strength is also limited. A more proper way of crafting the poisoned dataset to perform a data-agnostic attack remains an open discussion.

2) How to improve the success rate of data poisoning backdoor attacks: Backdoor attacks during the local data collection phase target specific parts of the local training data [49], [51], [117], [241]. These attacks are conducted using a black model and are relatively low-cost. However, current designs for these attacks assume the ratio of compromised clients in each round to be high, which is unrealistic in practical WFL deployments where the number of participants is typically quite large. Consequently, the ASR is low, and the backdoor effect is quickly diluted. To increase the ASR and effect at this stage, attackers need to find a way to increase their participation rate during the global model training process, which would increase the attack efficiency and success rate.

Challenges of backdoor attacks during the local data collection phase can be classified into two types: label-flipping-based and trigger-based attacks. Label-flipping-based attacks have low attack costs, but the flipped labels can cause noticeable outliers in the uploaded updates, making them easy to detect and eliminate [123]. To achieve a better attack effect, attackers must consider the trade-off between injection rate and stealthiness. Trigger-based attacks also create outliers, but these can be designed based on the target label to decrease the likelihood of detection. For example, if the target label is a picture of human faces, the trigger could be designed as glass. This targeted trigger design can decrease the probability of detection by the defender.

3) How to improve the backdoor attack effect during local model training: Backdoor attacks during the local model training phase require attackers to have complete control over compromised clients [39], [49], [242]. These attacks involve modifying the training process to inject a backdoor pattern, which can be stealthier and more accurate, but also more expensive to carry out than attacks conducted during the local data collection phase. The local model training process is difficult to tamper with, and the effects of a backdoor can be quickly diluted if compromised clients are not frequently selected. To increase the effectiveness of a local training phase attack, it is important to find a balance between injecting the backdoor and remaining stealthy. Even if an attacker is able to get the local backdoor pattern uploaded to the server aggregator, the defense mechanism on the server may still be able to detect and eliminate the backdoor, preventing a global injection [51].

4) How to decrease the backdoor attack cost during model aggregation: Backdoor attacks during the model aggregation phase have the highest success rate but also the highest cost [212]. To carry out these attacks, attackers need to have complete knowledge about the target WFL framework, including the number of participants, local training algorithm, aggregation protocol, and global model convergence process [51], [243]. In practice, it can be difficult for attackers to obtain this information, and the attack may not be easily replicated in a different WFL task due to the unique architecture of each task. To make the attack more feasible, attackers may try to minimize modifications to the original WFL framework and reduce the cost of the attack.

5) How to organize a coordinated backdoor attack in a more effective way: Coordinated-trigger-based data poisoning attack can be launched during the local data collection phase by multiple attackers, with each attack member only knowing part of the trigger pattern. Compared with the centralized-trigger-based attack, the injected trigger exhibits better stealthiness, making it harder to be fully defended. However, such a coordinated attack is still based on the black box model, resulting in limited attack strength. There exists a possibility that multiple attackers can initialize a joint backdoor attack via a white-box model. By acquiring more information from attack members and the WFL network training process, the backdoor pattern can be better crafted. By distributing backdoor patterns

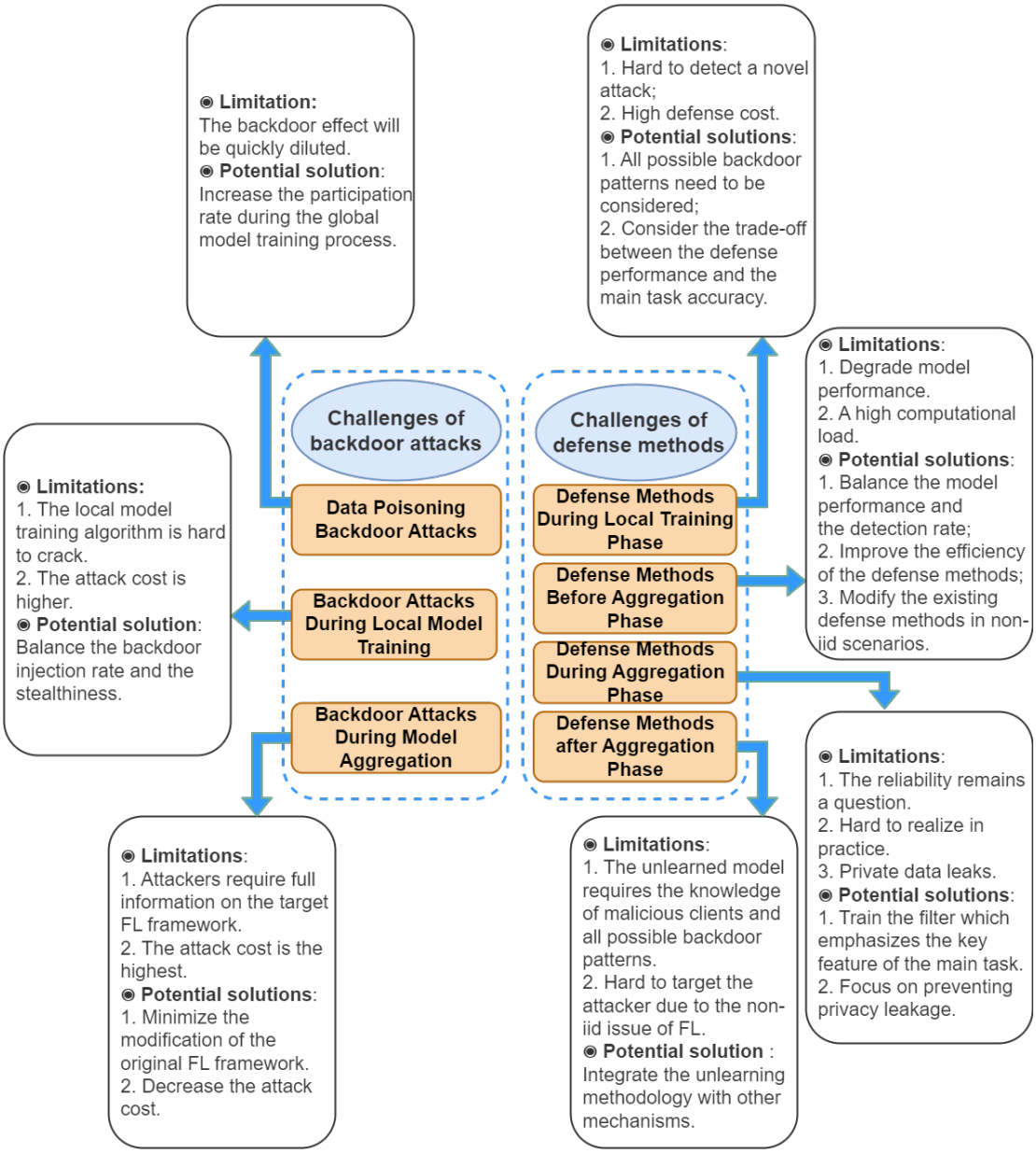


Fig. 21: A summary of challenges and future research directions for backdoor attacks on WFL.

in multiple compromised clients, the attackers are also more likely to trick the defender.

6) How to ensure the trigger rate when conducting backdoor attacks on WFL in WCNs: The trigger rate refers to the percentage of compromised clients selected in each training round. Therefore, a higher trigger rate can lead to a more severe attack. It is noted that in the WCN, the data distribution among different devices is normally imbalanced. Thus, the data received by a certain participant from the server could be much more(or less) than other participants in the same network. Such imbalanced data distribution will have a significant impact on model update calculation and upload. If the attacker could make use of such imbalances and intentionally control compromised clients to acquire more data from the WCN, it is reasonable to assume that the trigger rate will also increase. Consequently, the attack strength could be further improved.

B. Defense Methods in WFL

Defending against backdoor attacks on WFL can be divided into four categories based on the different training phases. These defense strategies may require different levels of knowledge about the model. In this section, we discuss the open issues of these defense methodologies and potential solutions.

1) How to achieve reasonable defense performance during the local training phase: The existing defense methods for the local training phase involve using a clean model trained with honest data as a filter to detect injected backdoors [228], [228], [242], [244], [245]. This type of defense is effective in defending against black-box backdoor attacks and can be used with a secure aggregation protocol. However, considering all possible backdoor patterns to achieve good defense performance can be costly, and the defense may not be able to detect novel attacks. Moreover, there is a trade-off between defense

performance and main task accuracy when designing this type of defense mechanism.

2) How to balance the model performance and the detection rate before the aggregation phase: Defending before the aggregation phase is a common strategy in which the server aggregator detects and ignores malicious updates and restricts participation from suspicious clients [212], [246]–[248]. This can significantly improve the robustness of the overall framework. However, there are still some design challenges to be addressed. Defense methods based on differential privacy (DP) can add noise to the model, which can degrade its performance. It is important to find the right balance between model performance and detection rate when adding noise. Additionally, using DP-GAN can be computationally intensive, so finding ways to improve the efficiency of these methods is an active area of research. Furthermore, the non-IID structure of WFL frameworks can make it challenging to defend against backdoor attacks while preserving the privacy and security of participants. Modifying the existing defense methods to work in non-IID scenarios is a key research direction.

3) How to improve the defense performances during the aggregation phase: The defense during the aggregation phase involves filtering out poisoned updates after they have been received by the aggregator [230], [249], [250]. Some defense methodologies at this stage use historical information to train a filter that can identify the backdoor in the original global model. However, the training process of WFL is random, which raises questions about the reliability of historic values. Additionally, the defense requires the attacker to participate in every training round, which may not be practical. To improve the defense performance, one solution could be to train a filter that emphasizes the key features of the main task, which may increase the filter's reliability. Privacy is also a major concern at this stage, and future research should focus on preventing privacy leakage.

4) How to remove the injected backdoor effect after the aggregation phase: After the aggregation phase, defense mechanisms can be employed to address any remaining injected backdoors in the global model [251]–[255]. One such approach is the use of an "unlearning" model, which aims to eliminate the residual effects of the backdoor by excluding the influence of suspicious participants and has been shown to be effective [225], [256], [257]. However, this approach requires the knowledge of the malicious clients and all possible backdoor patterns, which can be difficult to obtain in a WFL setting where the data is non-IID. It may be beneficial to combine the unlearning methodology with another mechanism to help identify malicious clients and remove backdoors while preserving client privacy.

5) How to optimize the processing load of the deployed defender: WFL is featured for its large number of participating devices and different types of devices, and devices might have limited computation abilities. While for the defenders that are deployed at the client end, the processing capability becomes an inevitable problem [50], [193]. It is essential to find a way to implement the defender, which can detect the anomaly input while keeping the local training process smooth. On the other hand, the defense mechanisms on the remote side also cause computing burdens to the server. The trade-off between processing efficiency and defense performance deserves further investigation.

6) How to defend backdoor attacks on decentralized WFL: In decentralized WFL, each node acts as both a contributor and a server. The model difference among nodes within the decentralized network could be significant. The existing similarity-based norm clipping methods may not function properly. On the other hand, there is no central server in the decentralized WFL framework. Each model can only obtain part of the network information from its neighbors or selected remote nodes. It is hard to broadcast a global defender to monitor all the nodes. Backdoor patterns are easier to parasite within a network and gradually propagate and spread throughout the network. A proper way to detect backdoor effects in decentralized WFL is a promising research direction.

7) How to compensate the transmission delay in the global model training process: During the training process of WFL in the WCN, it is challenging to ensure synchronization and consistency among all devices. The transmission delay or package loss during the wireless communication can cause uploaded local model updates to drop and further influence the global model convergence. Moreover, the attacker can take advantage of such unfairness and increase the attack success rate. It is of great significance to develop a defense mechanism that is capable of re-constructing the dropped transmission message and, at the same time, monitoring over-active clients to further increase the robustness of the entire network.

VIII. Conclusion

Owing to the inherent features of wireless networks and systems such as a large number of participants, massive data, and geographically distributed deployment, WFL is a rapidly growing technology with potential applications for various intelligent computing tasks in WCNs. However, as WFL advances, so do the techniques for launching backdoor attacks, posing a threat to the robustness of WFL. This survey paper has thoroughly examined the existing attack strategies and defense mechanisms to safeguard against all types of attacks in WFL. The backdoor attack methods have been systematically classified into two main categories: data poisoning attacks and model poisoning attacks, with the latter further divided into local training phase attacks and aggregation phase attacks. The defense methods have been categorized into four types based on their application stages: during local model training, before aggregation, during aggregation, and after aggregation. The key characteristics of existing backdoor attack and defense methods in WFL over WCN have been highlighted to provide insights for future research and development efforts in this significant domain.

- In backdoor attacks during the local data collection phase, also known as data poisoning-based backdoor attacks, an attacker can change only a part of the training data. Since limited prior information on the entire model is required, it contributes to a decreasing attack cost. Such a black-box model attack has a relatively weaker attack strength, and the backdoor effect is easier to detect and eliminate.

- For the model poisoning-based backdoor attacks, the attacker needs to fully control one or more participants. With more knowledge of the model, e.g., the training algorithm and aggregation protocol acquired by the attacker, the attack success rate and the backdoor stealthiness can be improved. However, the attack cost is higher compared to a data poisoning-based backdoor attack. It is worth mentioning that, in most of the existing attack methods, a typical assumption is that, malicious participants are more likely to be selected in each training round, and the proportion of clients controlled by the attacker is typically high.
- Many defense mechanisms have been proposed to settle robust threats. Based on different defense stages, the defender can filter the malicious updates from the compromised clients controlled by the attacker, either systematically or statistically.
- In the existing defense method design, there is one common trade-off between the detection rate and the main task success rate. In fact, backdoor detection relies on threshold design in different defense mechanisms. Inevitably, some information in the genuine model will be clipped, which impacts the model's performance. Meanwhile, several defense strategies are not compatible with the security aggregation protocols.
- The existing defense mechanisms are commonly designed for one or several specific backdoor attack methodologies. The training process of the defender requires historical information about the attacks. Thus, they are still vulnerable when facing a novel type of attack.

This survey has provided a clear and concise overview of the current state of backdoor attacks and defenses in WFL, and guidelines on enhancing the robustness of WFL against backdoor attacks.

ACKNOWLEDGMENTS

This research is supported by the National Key R&D Program of China Grant No. 2022ZD0116800, Taishan Scholars Program No. TSQNZ20230621 and TSQN202211214, Shandong Excellent Young Scientists Fund Program (Overseas) No. 2023HWYQ-113.

REFERENCES

[1] Q. V. Khanh, N. V. Hoai, L. D. Manh, A. N. Le, and G. Jeon, "Wireless communication technologies for iot in 5g: Vision, applications, and challenges," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–12, 2022.

[2] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang, and R. P. Liu, "Energy-efficient admission of delay-sensitive tasks for mobile edge computing," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2603–2616, 2018.

[3] Q. Cui, Z. Gong, W. Ni, Y. Hou, X. Chen, X. Tao, and P. Zhang, "Stochastic online learning for mobile edge computing: Learning from changes," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 63–69, 2019.

[4] X. Lyu, W. Ni, H. Tian, R. P. Liu, X. Wang, G. B. Giannakis, and A. Paulraj, "Distributed online optimization of fog computing for selfish devices with out-of-date information," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7704–7717, 2018.

[5] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 60–69, 2019.

[6] Y. Wang, T. Li, S. Li, X. Yuan, and W. Ni, "New adversarial image detection based on sentiment analysis," *IEEE Transactions on Neural Networks and Learning Systems*, 2023, early access.

[7] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[8] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep Q-network for UAV-assisted online power transfer and data collection," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 215–12 226, 2019.

[9] S. Hu, X. Yuan, W. Ni, X. Wang, and A. Jamalipour, "RIS-assisted jamming rejection and path planning for uav-borne iot platform: A new deep reinforcement learning framework," *IEEE Internet of Things Journal*, 2023, early access.

[10] K. Li, Y. Cui, W. Li, T. Lv, X. Yuan, S. Li, W. Ni, M. Simsek, and F. Dressler, "When Internet of Things meets metaverse: Convergence of physical and cyber worlds," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4148 – 4173, Dec. 2022.

[11] M. S. Yousefpoor, E. Yousefpoor, H. Barati, A. Barati, A. Movaghar, and M. Hosseinzadeh, "Secure data aggregation methods and countermeasures against various attacks in wireless sensor networks: A comprehensive review," *Journal of Network and Computer Applications*, vol. 190, p. 103118, 2021.

[12] X. Zhang, G. Klevering, X. Lei, Y. Hu, L. Xiao, and G.-h. Tu, "The security in optical wireless communication: A survey," *ACM Computing Surveys*, 2023.

[13] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1759–1799, 2021.

[14] X. Yu, B. Xiao, W. Ni, and X. Wang, "Optimal power control for over-the-air federated edge learning using statistical channel knowledge," in *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2022, pp. 232–237.

[15] L. Pu, Q. Cui, X. Li, B. Zhao, W. Ni, M. Ai, X. Tao *et al.*, "Federated learning-based heterogeneous load prediction and slicing for 5G systems and beyond," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 166–172.

[16] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Federated learning for online resource allocation in mobile edge computing: A deep reinforcement learning approach," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[17] Z. Wang, Z. Zhang, Y. Tian, Q. Yang, H. Shan, W. Wang, and T. Q. Quek, "Asynchronous federated learning over wireless communication networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 6961–6978, 2022.

[18] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[19] W. Li, T. Lv, Y. Cao, W. Ni, and M. Peng, "Multi-carrier NOMA-empowered wireless federated learning with optimal power and bandwidth allocation," *IEEE Transactions on Wireless Communications*, 2023, early access.

[20] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[21] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[22] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021.

[23] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[24] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[25] S. Zawad, A. Ali, P.-Y. Chen, A. Anwar, Y. Zhou, N. Baracaldo, Y. Tian, and F. Yan, "Curse or redemption? how data heterogeneity affects the robustness of federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10 807–10 814.

[26] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[27] S. Rajput, H. Wang, Z. Charles, and D. Papailiopoulos, "Detox: A redundancy-based framework for faster and more robust gradient aggregation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] X. Cao, J. Jia, and N. Z. Gong, "Provably secure federated learning against malicious clients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6885–6893.

[29] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "Poisongan: Generative poisoning attacks against federated learning in edge computing systems," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3310–3322, 2020.

[30] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[31] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to {Byzantine-Robust} federated learning," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1605–1622.

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

[33] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos, "Attack of the tails: Yes, you really can backdoor federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 070–16 084, 2020.

[34] H. Eilertsen, "Backdoor found in themes and plug-ins from accesspress themes," https://jetpack.com/blog/backdoor-found-in-themes-and-plugins-from-accesspress-themes/.

[35] K. Li, J. Zheng, X. Yuan, W. Ni, O. B. Akan, and H. V. Poor, "Data-agnostic model poisoning against federated learning: A graph autoencoder approach," *arxiv, 2311.18498*, 2023.

[36] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[37] N. Rodríguez-Barroso, E. Martínez-Cámara, M. V. Luzón, and F. Herrera, "Backdoor attacks-resilient aggregation based on robust filtering of outliers in federated learning for image classification," *Knowledge-Based Systems*, vol. 245, p. 108588, 2022.

[38] T. D. Nguyen, P. Rieger, R. De Viti, H. Chen, B. B. Brandenburg, H. Yalame, H. Möllering, H. Fereidooni, S. Marchal, M. Miettinen *et al.*, "FLAME: Taming backdoors in federated learning," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1415–1432.

[39] B. Hou, J. Gao, X. Guo, T. Baker, Y. Zhang, Y. Wen, and Z. Liu, "Mitigating the backdoor attack by federated filters for industrial iot applications," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3562–3571, 2021.

[40] J. Sun, A. Li, L. DiValentin, A. Hassanzadeh, Y. Chen, and H. Li, "FL-WBC: Enhancing robustness against model poisoning attacks in federated learning from a client perspective," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 613–12 624, 2021.

[41] R. Shokri *et al.*, "Bypassing backdoor detection algorithms in deep learning," in *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2020, pp. 175–183.

[42] Y. Dong, X. Yang, Z. Deng, T. Pang, Z. Xiao, H. Su, and J. Zhu, "Black-box detection of backdoor attacks with limited information and data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 482–16 491.

[43] L. Zhang, G. Ding, Q. Wu, Y. Zou, Z. Han, and J. Wang, "Byzantine attack and defense in cognitive radio networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1342–1363, 2015.

[44] N. M. Jebreel and J. Domingo-Ferrer, "Fl-defender: Combating targeted attacks in federated learning," *Knowledge-Based Systems*, vol. 260, p. 110178, 2023.

[45] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[46] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[47] H. S. Sikandar, H. Waheed, S. Tahir, S. U. Malik, and W. Rafique, "A detailed survey on federated learning attacks and defenses," *Electronics*, vol. 12, no. 2, p. 260, 2023.

[48] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *arXiv preprint arXiv:2007.10760*, 2020.

[49] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[50] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *IEEE Open Journal of Signal Processing*, 2022.

[51] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[52] X. Gong, Y. Chen, Q. Wang, and W. Kong, "Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions," *IEEE Wireless Communications*, 2022.

[53] T. Dung Nguyen, T. Nguyen, P. Le Nguyen, H. H. Pham, K. Doan, and K.-S. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *arXiv e-prints*, pp. arXiv–2303, 2023.

[54] M. S. Jere, T. Farnan, and F. Koushanfar, "A taxonomy of attacks on federated learning," *IEEE Security & Privacy*, vol. 19, no. 2, pp. 20–28, 2020.

[55] S. Hu, X. Chen, W. Ni, E. Hossain, and X. Wang, "Distributed machine learning for wireless communication networks: Techniques, architectures, and applications," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1458–1493, 2021.

[56] X. Lyu, C. Ren, W. Ni, H. Tian, R. P. Liu, and E. Dutkiewicz, "Optimal online data partitioning for geo-distributed machine learning in edge of wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2393–2406, 2019.

[57] G. Yu, X. Wang, P. Yu, C. Sun, W. Ni, and R. P. Liu, "Dataset obfuscation: Its applications to and impacts on edge machine learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 5, pp. 1 – 15, 2023.

[58] M. Gupta and S. Singh, "A survey on the zigbee protocol, it's security in internet of things (iot) and comparison of zigbee with bluetooth and wi-fi," in *Applications of artificial intelligence in engineering: proceedings of first global conference on artificial intelligence and applications (GCAIA 2020)*. Springer, 2021, pp. 473–482.

[59] S. Kaffash, A. T. Nguyen, and J. Zhu, "Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis," *International Journal of Production Economics*, vol. 231, p. 107868, 2021.

[60] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang, and H. Du, "Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8810–8824, 2022.

[61] Q. Cui, X. Hu, W. Ni, X. Tao, P. Zhang, T. Chen, K.-C. Chen, and M. Haenggi, "Vehicular mobility patterns and their applications to Internet-of-Vehicles: A comprehensive survey," *Science China Information Sciences*, vol. 65, no. 11, pp. 1–42, 2022.

[62] B. Vlačić, L. Corbo, S. C. e Silva, and M. Dabić, "The evolving role of artificial intelligence in marketing: A review and research agenda," *Journal of Business Research*, vol. 128, pp. 187–203, 2021.

[63] J. Wang, C. Xu, J. Zhang, and R. Zhong, "Big data analytics for intelligent manufacturing systems: A review," *Journal of Manufacturing Systems*, vol. 62, pp. 738–752, 2022.

[64] M. A. Khan, S. Abbas, A. Rehman, Y. Saeed, A. Zeb, M. I. Uddin, N. Nasser, and A. Ali, "A machine learning approach for blockchain-based smart home networks security," *IEEE Network*, vol. 35, no. 3, pp. 223–229, 2020.

[65] Z. Lv, W. Kong, X. Zhang, D. Jiang, H. Lv, and X. Lu, "Intelligent security planning for regional distributed energy internet," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3540–3547, 2019.

[66] A. A. Abdellatif, A. Mohamed, C. F. Chiasserini, M. Tlili, and A. Erbad, "Edge computing for smart health: Context-aware approaches, opportunities, and challenges," *IEEE Network*, vol. 33, no. 3, pp. 196–203, 2019.

[67] X. Chen, W. Dai, W. Ni, X. Wang, S. Zhang, S. Xu, and Y. Sun, "Augmented deep reinforcement learning for online energy minimization of wireless powered mobile edge computing," *IEEE Trans. Comm.*, vol. 71, no. 5, pp. 2698–2710, 2023.

[68] S. Bi, C. Wang, B. Wu, S. Hu, W. Huang, W. Ni, Y. Gong, and X. Wang, "A comprehensive survey on applications of AI technologies to failure analysis of industrial systems," *Engineering Failure Analysis*, vol. 148, p. 107172, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350630723001267

[69] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020.

[70] J. Odirichukwu, P. Asagba, and F. Onuodu, "Interoperable protocols of the internet of things and internet of robotic things: A review." *International Journal of Computing, Intelligence and Security Research*, vol. 1, no. 1, pp. 101–123, 2021.

[71] X. Yuan, W. Ni, M. Ding, K. Wei, J. Li, and H. V. Poor, "Amplitude-varying perturbation for balancing privacy and utility in federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1884–1897, 2023.

[72] B. Vaseghi, S. S. Hashemi, S. Mobayen, and A. Fekih, "Finite time chaos synchronization in time-delay channel and its application to satellite image encryption in ofdm communication systems," *IEEE Access*, vol. 9, pp. 21 332–21 344, 2021.

[73] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Towards dual-functional wireless networks for 6g and beyond," *IEEE journal on selected areas in communications*, 2022.

[74] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When machine learning meets privacy in 6g: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2694–2724, 2020.

[75] H. Huang, A. V. Savkin, and W. Ni, "Decentralized navigation of a UAV team for collaborative covert eavesdropping on a group of mobile ground nodes," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 4, pp. 3932–3941, 2022.

[76] H. Huang, A. Savkin, and W. Ni, "Navigation of a UAV team for collaborative eavesdropping on multiple ground transmitters," *IEEE Transactions on Vehicular Technology*, 2021.

[77] X. Yuan, Z. Feng, W. Ni, R. Liu, J. A. Zhang, and W. Xu, "Secrecy performance of terrestrial radio links under collaborative aerial eavesdropping," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 604–619, 2020.

[78] X. Yuan, Z. Feng, W. Ni, Z. Wei, R. P. Liu, and J. A. Zhang, "Secrecy rate analysis against aerial eavesdropper," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7027–7042, 2019.

[79] X. Yuan, S. Hu, W. Ni, R. P. Liu, and X. Wang, "Joint user, channel, modulation-coding selection, and RIS configuration for jamming resistance in multiuser ofdma systems," *IEEE Transactions on Communications*, vol. 71, no. 3, p. 1631 – 1645, March 2023.

[80] T. Altaf, X. Wang, W. Ni, G. Yu, R. P. Liu, and R. Braun, "A new concatenated multigraph neural network for IoT intrusion detection," *Internet of Things*, p. 100818, 2023.

[81] T. Altaf, X. Wang, W. Ni, R. P. Liu, and R. Braun, "NE-GConv: A lightweight node edge graph convolutional network for intrusion detection," *Computers & Security*, vol. 130, p. 103285, 2023.

[82] K. Li, W. Ni, A. Noor, and M. Guizani, "Employing intelligent aerial data aggregators for the Internet of Things: Challenges and solutions," *IEEE Internet of Things Magazine*, vol. 5, no. 1, pp. 136–141, 2022.

[83] S. Wang, C. Yuen, W. Ni, Y. L. Guan, and T. Lv, "Multiagent deep reinforcement learning for cost- and delay-sensitive virtual network function placement and routing," *IEEE Transactions on Communications*, vol. 70, no. 8, pp. 5208–5224, 2022.

[84] Q. Cui, X. Zhao, W. Ni, Z. Hu, X. Tao, and P. Zhang, "Multi-agent deep reinforcement learning-based interdependent computing for mobile edge computing-assisted robot teams," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 5, pp. 6599–6610, 2023.

[85] M. A. Raza, M. Abolhasan, J. Lipman, N. Shariati, W. Ni, and A. Jamalipour, "Statistical learning-based grant-free access for delay-sensitive internet of things applications," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 5, pp. 5492–5506, 2022.

[86] M. Raza, M. Abolhasan, J. Lipman, N. Shariati, W. Ni, and A. Jamalipour, "Statistical learning-based adaptive network access for the industrial Internet-of-Things," *IEEE Internet of Things Journal*, 2023.

[87] B. Wu, T. Chen, W. Ni, and X. Wang, "Multi-agent multi-armed bandit learning for online management of edge-assisted computing," *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8188–8199, 2021.

[88] N. H. Tran, W. Bao, A. Zomaya, M. N. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *IEEE INFOCOM 2019-IEEE conference on computer communications*. IEEE, 2019, pp. 1387–1395.

[89] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[90] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[91] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, "Exploring deep-reinforcement-learning-assisted federated learning for online resource allocation in privacy-preserving EdgeIoT," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 099–21 110, 2022.

[92] J. Liu, J. Huang, Y. Zhou, X. Li, S. Ji, H. Xiong, and D. Dou, "From distributed machine learning to federated learning: A survey," *Knowledge and Information Systems*, vol. 64, no. 4, pp. 885–917, 2022.

[93] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[94] J. Zhang, J. Zhang, D. W. K. Ng, and B. Ai, "Federated learning-based cell-free massive mimo system for privacy-preserving," *IEEE Transactions on Wireless Communications*, 2022.

[95] S. Chen, D. Yu, Y. Zou, J. Yu, and X. Cheng, "Decentralized wireless federated learning with differential privacy," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 6273–6282, 2022.

[96] M. Lee, G. Yu, and H. Dai, "Privacy-preserving decentralized inference with graph neural networks in wireless networks," *IEEE Transactions on Wireless Communications*, 2023.

[97] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170–185, 2020.

[98] S. H. Haji and A. M. Abdulazeez, "Comparison of optimization techniques based on gradient descent algorithm: A review," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 4, pp. 2715–2743, 2021.

[99] S. Messaoud, A. Bradai, and E. Moulay, "Online GMM clustering and mini-batch gradient descent based optimization for industrial IoT 4.0," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1427–1435, 2019.

[100] K. P. Rakshitha and N. Naveen, "Op-RMSprop (optimized-root mean square propagation) classification for prediction of Polycystic Ovary Syndrome (PCOS) using hybrid machine learning technique," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022.

[101] X. Yu, B. Xiao, W. Ni, and X. Wang, "Optimal adaptive power control for over-the-air federated edge learning under fading channels," *IEEE Trans. Comm.*, vol. 71, no. 9, p. 5199 – 5213, Sep. 2023.

[102] B. Xiao, X. Yu, W. Ni, X. Wang, and H. V. Poor, "Over-the-air federated learning: Status quo, open challenges, and future directions," 2023.

[103] B. C. Tedeschini, S. Savazzi, R. Stoklasa, L. Barbieri, I. Stathopoulos, M. Nicoli, and L. Serio, "Decentralized federated learning for healthcare networks: A case study on tumor segmentation," *IEEE Access*, vol. 10, pp. 8693–8708, 2022.

[104] G. Yu, X. Wang, C. Sun, Q. Wang, P. Yu, W. Ni, R. P. Liu, and X. Xu, "IronForge: An open, secure, fair, decentralized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023, early access.

[105] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2512–2520.

[106] M. Song, Z. Wang, Z. Zhang, Y. Song, Q. Wang, J. Ren, and H. Qi, "Analyzing user-level privacy attack against federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2430–2444, 2020.

[107] M. Fan, Z. Si, X. Xie, Y. Liu, and T. Liu, "Text backdoor detection using an interpretable rnn abstract model," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4117–4132, 2021.

[108] S. Hu, X. Yuan, W. Ni, X. Wang, E. Hossain, and H. V. Poor, "Ofdma-f^2l: Federated learning with flexible aggregation over an OFDMA air interface," *IEEE Transactions on Wireless Communications*, 2023, to appear.

[109] Y. Guan, S. Zou, H. Peng, W. Ni, Y. Sun, and H. Gao, "Cooperative UAV trajectory design for disaster area emergency communications: A multi-agent PPO method," *IEEE Internet of Things Journal*, 2023, early access.

[110] A. Hbaieb, S. Ayed, and L. Chaari, "Federated learning based IDS approach for the IoV," in *Proceedings of the 17th International Conference on Availability, Reliability and Security*, ser. ARES '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3538969.3544422

[111] X. Lin, D. Ai, B. Ma *et al.*, "Federated learning-based intrusion detection system for in-vehicle network using statistics of controller area network messages," in *Proc. 2nd International Conference on Network Simulation and Evaluation*, 22 - 24 Nov. 2023, pp. 1 – 1.

[112] Y. Guan, S. Zou, K. Li *et al.*, "MAPPO-based cooperative UAV trajectory design with long-range emergency communications in disaster areas," in *Proc. 2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2023, pp. 376–381.

[113] S. Rani, A. Kataria, S. Kumar, and P. Tiwari, "Federated learning for secure IoMT-applications in smart healthcare systems: A comprehensive review," *Knowledge-Based Systems*, vol. 274, p. 110658, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123004082

[114] K. Li, X. Yuan, J. Zheng, W. Ni, and M. Guizani, "Exploring adversarial graph autoencoders to manipulate federated learning in the Internet of Things," in *Proc. 2023 International Wireless Communications and Mobile Computing (IWCMC)*, 2023, pp. 898–903.

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

[116] X. Gong, Y. Chen, H. Huang, Y. Liao, S. Wang, and Q. Wang, "Coordinated backdoor attacks against federated learning with model-dependent triggers," *IEEE network*, vol. 36, no. 1, pp. 84–90, 2022.

[117] A. Schwarzschild, M. Goldblum, A. Gupta, J. P. Dickerson, and T. Goldstein, "Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9389–9398.

[118] F. A. Yerlikaya and Ş. Bahtiyar, "Data poisoning attacks against machine learning algorithms," *Expert Systems with Applications*, vol. 208, p. 118101, 2022.

[119] M. Jagielski, G. Severi, N. Pousette Harger, and A. Oprea, "Subpopulation data poisoning attacks," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 3104–3122.

[120] Y. Yang, T. Y. Liu, and B. Mirzasoleiman, "Not all poisons are created equal: Robust training against data poisoning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 154–25 165.

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

[122] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *International Conference on Learning Representations*, 2019.

[123] E. Rosenfeld, E. Winston, P. Ravikumar, and Z. Kolter, "Certified robustness to label-flipping attacks via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8230–8241.

[124] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, 2019.

[125] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks against deep learning systems in the physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6206–6215.

[126] Z. Zhang, A. Panda, L. Song, Y. Yang, M. Mahoney, P. Mittal, R. Kannan, and J. Gonzalez, "Neurotoxin: durable backdoors in federated learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 429–26 446.

[127] Y. Feng, B. Ma, J. Zhang, S. Zhao, Y. Xia, and D. Tao, "FIBA: Frequency-injection based backdoor attack in medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 876–20 885.

[128] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 957–11 965.

[129] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

[130] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 443–14 452.

[131] K. Yoshida and T. Fujino, "Disabling backdoor and identifying poison data by using knowledge distillation in backdoor attacks on deep neural networks," in *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security*, 2020, pp. 117–127.

[132] K. Doan, Y. Lao, and P. Li, "Backdoor attack with imperceptible input and latent modification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 944–18 957, 2021.

[133] Y. Sun, T. Zhang, X. Ma, P. Zhou, J. Lou, Z. Xu, X. Di, Y. Cheng, and L. Sun, "Backdoor attacks on crowd counting," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5351–5360.

[134] R. Mayerhofer and R. Mayer, "Poisoning attacks against feature-based image classification," in *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, 2022, pp. 358–360.

[135] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 2043–2059.

[136] J. Xu, R. Wang, S. Koffas, K. Liang, and S. Picek, "More is better (mostly): On the backdoor attacks in federated graph neural networks," in *Proceedings of the 38th Annual Computer Security Applications Conference*, 2022, pp. 684–698.

[137] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang, "BADNL: Backdoor attacks against NLP models with semantic-preserving improvements," in *Annual Computer Security Applications Conference*, 2021, pp. 554–569.

[138] H.-y. Lu, C. Fan, J. Yang, C. Hu, W. Fang, and X.-j. Wu, "Where to attack: A dynamic locator model for backdoor attack in text classifications," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 984–993.

[139] H. Kwon and S. Lee, "Textual backdoor attack for the text classification system," *Security and Communication Networks*, vol. 2021, pp. 1–11, 2021.

[140] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, "Hidden trigger backdoor attack on NLP models via linguistic style manipulation," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 3611–3628.

[141] K. Shao, Y. Zhang, J. Yang, X. Li, and H. Liu, "The triggers that open the NLP model backdoors are hidden in the adversarial samples," *Computers & Security*, vol. 118, p. 102730, 2022.

[142] Y. Liu, G. Shen, G. Tao, S. An, S. Ma, and X. Zhang, "PICCOLO: Exposing complex backdoors in NLP transformer models," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 2025–2042.

[143] K. Doan, Y. Lao, W. Zhao, and P. Li, "LIRA: Learnable, imperceptible and robust backdoor attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 966–11 976.

[144] J. Xu, M. Xue, and S. Picek, "Explainability-based backdoor attacks against graph neural networks," in *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*, 2021, pp. 31–36.

[145] K. Varma, Y. Zhou, N. Baracaldo, and A. Anwar, "LEGATO: A layerwise gradient aggregation algorithm for mitigating byzantine attacks in federated learning," in *2021 IEEE 14th international conference on cloud computing (CLOUD)*. IEEE, 2021, pp. 272–277.

[146] A. Gouissem, K. Abualsaud, E. Yaacoub, T. Khattab, and M. Guizani, "Federated learning stability under byzantine attacks," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 572–577.

[147] X. Ma, Q. Jiang, M. Shojafar, M. Alazab, S. Kumar, and S. Kumari, "DisBezant: secure and robust federated learning against byzantine attack in iot-enabled mts," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[148] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," in *ECAI 2012*. IOS Press, 2012, pp. 870–875.

[149] K. Aryal, M. Gupta, and M. Abdelsalam, "Analysis of label-flip poisoning attack on machine learning based malware detector," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 4236–4245.

[150] T. Liu, X. Hu, and T. Shu, "Facilitating early-stage backdoor attacks in federated learning with whole population distribution inference," *IEEE Internet of Things Journal*, 2023.

[151] Z. Zhao, X. Chen, Y. Xuan, Y. Dong, D. Wang, and K. Liang, "Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 213–15 222.

[152] C. Zhao, Y. Wen, S. Li, F. Liu, and D. Meng, "FederatedReverse: A detection and defense method against backdoor attacks in federated learning," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021, pp. 51–62.

[153] X. Li, Z. Chen, Y. Zhao, Z. Tong, Y. Zhao, A. Lim, and J. T. Zhou, "PointBA: Towards backdoor attacks in 3d point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 492–16 501.

[154] Z. Wei, J. Shi, Y. Duan, R. Liu, Y. Han, and Z. Liu, "Backdoor filter: Mitigating visible backdoor triggers in dataset," in *2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI)*. IEEE, 2021, pp. 102–105.

[155] Y. Li, J. Hua, H. Wang, C. Chen, and Y. Liu, "DeepPayload: Black-box backdoor attack on deep learning models through neural payload injection," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 263–274.

[156] J. Cao and l. Zhu, "A highly efficient, confidential, and continuous federated learning backdoor attack strategy," in *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, 2022, pp. 18–27.

[157] K. Chen, H. Zhang, X. Feng, X. Zhang, B. Mi, and Z. Jin, "Backdoor attacks against distributed swarm learning," *ISA transactions*, 2023.

[158] S. Hong, N. Carlini, and A. Kurakin, "Handcrafted backdoors in deep neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8068–8080, 2022.

[159] S. Garg, A. Kumar, V. Goel, and Y. Liang, "Can adversarial weight perturbations inject neural backdoors," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2029–2032.

[160] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, 2021.

[161] C. Ren, X. Lyu, W. Ni, H. Tian, and R. P. Liu, "Distributed online learning of fog computing under nonuniform device cardinality," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1147–1159, 2019.

[162] M. Zhang, L. Hu, C. Shi, and X. Wang, "Adversarial label-flipping attack and defense for graph neural networks," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 791–800.

[163] G. Shen, Y. Liu, G. Tao, S. An, Q. Xu, S. Cheng, S. Ma, and X. Zhang, "Backdoor scanning for deep neural networks through k-arm optimization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9525–9536.

[164] C. Xie, M. Chen, P.-Y. Chen, and B. Li, "CRFL: Certifiably robust federated learning against backdoor attacks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 372–11 382.

[165] K. Zhang, Y. Zhang, R. Sun, P.-W. Tsai, M. U. Hassan, X. Yuan, M. Xue, and J. Chen, "Bounded and unbiased composite differential privacy," *arXiv preprint arXiv:2311.02324*, 2023.

[166] Y. Wang, M. Zhao, S. Li, X. Yuan, and W. Ni, "Dispersed pixel perturbation-based imperceptible backdoor trigger for image classifier models," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3091–3106, 2022.

[167] D. Ye, M.-M. Chen, and H.-J. Yang, "Distributed adaptive event-triggered fault-tolerant consensus of multiagent systems with general linear dynamics," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 757–767, 2018.

[168] Y. Chen, Z. Zheng, and X. Gong, "MARNet: Backdoor attacks against cooperative multi-agent reinforcement learning," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[169] Z. Zhang, J. Jia, B. Wang, and N. Z. Gong, "Backdoor attacks to graph neural networks," in *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*, 2021, pp. 15–26.

[170] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022, pp. 703–718.

[171] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 463–16 472.

[172] Z. Qin, L. Yao, D. Chen, Y. Li, B. Ding, and M. Cheng, "Revisiting personalized federated learning: Robustness against backdoor attacks," *arXiv preprint arXiv:2302.01677*, 2023.

[173] J. H. Anajemba, C. Iwendi, I. Razzak, J. A. Ansere, and I. M. Okpalaoguchi, "A counter-eavesdropping technique for optimized privacy of wireless industrial iot communications," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 6445–6454, 2022.

[174] X. Jia, Y. Zhang, B. Wu, J. Wang, and X. Cao, "Boosting fast adversarial training with learnable adversarial initialization," *IEEE Transactions on Image Processing*, vol. 31, pp. 4417–4430, 2022.

[175] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2488–2524, 2019.

[176] X. Lyu, W. Ni, H. Tian, R. P. Liu, X. Wang, G. B. Giannakis, and A. Paulraj, "Optimal schedule of mobile edge computing for Internet of Things using partial information," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2606–2615, 2017.

[177] M. Shafieinejad, N. Lukas, J. Wang, X. Li, and F. Kerschbaum, "On the robustness of backdoor-based watermarking in deep neural networks," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, 2021, pp. 177–188.

[178] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, and D. Huba, "Federated learning with buffered asynchronous aggregation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3581–3607.

[179] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *IEEE INFOCOM 2022-IEEE conference on computer communications*. IEEE, 2022, pp. 1739–1748.

[180] A. Mahmoudi, H. S. Ghadikolaei, J. M. B. D. S. Júnior, and C. Fischione, "FedCau: A proactive stop policy for communication and computation efficient federated learning," *arXiv preprint arXiv:2204.07773*, 2022.

[181] V. Shejwalkar and A. Houmansadr, "Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning," in *NDSS*, 2021.

[182] S. Mahloujifar, E. Ghosh, and M. Chase, "Property inference from poisoning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2022, pp. 1569–1569.

[183] Y. Zeng, W. Park, Z. M. Mao, and R. Jia, "Rethinking the backdoor attacks' triggers: A frequency perspective," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 473–16 481.

[184] Y. Wang, J. Li, H. Liu, Y. Wang, Y. Wu, F. Huang, and R. Ji, "Black-box dissector: Towards erasing-based hard-label model stealing attack," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 2022, pp. 192–208.

[185] C. Chen, L. Wei, L. Zhang, Y. Peng, J. Ning *et al.*, "DeepGuard: Backdoor attack detection and identification schemes in privacy-preserving deep neural networks," *Security and Communication Networks*, vol. 2022, 2022.

[186] E. Hallaji, R. Razavi-Far, M. Saif, and E. Herrera-Viedma, "Label noise analysis meets adversarial training: A defense against label poisoning in federated learning," *Knowledge-Based Systems*, vol. 266, p. 110384, 2023.

[187] S. Gajbhiye, P. Singh, and S. Gupta, "Data poisoning attack by label flipping on splitfed learning," in *Recent Trends in Image Processing and Pattern Recognition: 5th International Conference, RTIP2R 2022, Kingsville, TX, USA, December 1-2, 2022, Revised Selected Papers*. Springer, 2023, pp. 391–405.

[188] H. Zheng, H. Xiong, J. Chen, H. Ma, and G. Huang, "Motif-Backdoor: Rethinking the backdoor attack on graph neural networks via Motifs," *IEEE Transactions on Computational Social Systems*, 2023.

[189] Y. Chen, Z. Ye, H. Zhao, Y. Wang *et al.*, "Feature-based graph backdoor attack in the node classification task," *International Journal of Intelligent Systems*, vol. 2023, 2023.

[190] H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. Thuraisingham, "DEEPSWEEP: An evaluation framework for mitigating dnn backdoor attacks using data augmentation," in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 363–377.

[191] T. D. Nguyen, P. Rieger, M. Miettinen, and 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.

[192] J. Zheng, H. Tian, W. Ni, W. Ni, and P. Zhang, "Balancing accuracy and integrity for reconfigurable intelligent surface-aided over-the-air

federated learning," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 964–10 980, 2022.

[193] Y. Wei, H. Gao, Y. Wang, Y. Gao, and H. Liu, "A lightweight backdoor defense framework based on image inpainting," *Neurocomputing*, vol. 537, pp. 22–36, 2023.

[194] Y. Wang, D.-H. Zhai, Y. He, and Y. Xia, "An adaptive robust defending algorithm against backdoor attacks in federated learning," *Future Generation Computer Systems*, vol. 143, pp. 118–131, 2023.

[195] Y.-C. Lai, J.-Y. Lin, Y.-D. Lin, R.-H. Hwang, P.-C. Lin, H.-K. Wu, and C.-K. Chen, "Two-phase defense against poisoning attacks on federated learning-based intrusion detection," *Computers & Security*, vol. 129, p. 103205, 2023.

[196] X. Qi, T. Xie, Y. Li, S. Mahloujifar, and P. Mittal, "Revisiting the assumption of latent separability for backdoor defenses," in *The eleventh international conference on learning representations*, 2023.

[197] S. Zhai, Q. Shen, X. Chen, W. Wang, C. Li, Y. Fang, and Z. Wu, "NCL: Textual backdoor defense using noise-augmented contrastive learning," *arXiv preprint arXiv:2303.01742*, 2023.

[198] E. Soremekun, S. Udeshi, and S. Chattopadhyay, "Towards backdoor attacks and defense in robust machine learning models," *Computers & Security*, p. 103101, 2023.

[199] Z. Chen, S. Wang, A. Fu, Y. Gao, S. Yu, and R. H. Deng, "LinkBreaker: Breaking the backdoor-trigger link in dnns via neurons consistency check," *IEEE Transactions on Information Forensics and Security*, 2022.

[200] X. Su, Y. Zhou, L. Cui, and J. Liu, "On model transmission strategies in federated learning with lossy communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 4, pp. 1173–1185, 2023.

[201] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 487–500, 2022.

[202] S. Li, Y. Cheng, W. Wang, Y. Liu, and T. Chen, "Learning to detect malicious clients for robust federated learning," *arXiv preprint arXiv:2002.00211*, 2020.

[203] Z. Gu and Y. Yang, "Detecting malicious model updates from federated learning on conditional variational autoencoder," in *2021 IEEE international parallel and distributed processing symposium (IPDPS)*. IEEE, 2021, pp. 671–680.

[204] J. Gao, B. Zhang, X. Guo, T. Baker, M. Li, and Z. Liu, "Secure partial aggregation: Making federated learning more robust for industry 4.0 applications," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 6340–6348, 2022.

[205] O. Aramoon, P.-Y. Chen, G. Qu, and Y. Tian, "Meta federated learning," *arXiv preprint arXiv:2102.05561*, 2021.

[206] I. Makhdoom, M. Abolhasan, H. Abbas, and W. Ni, "Blockchain's adoption in IoT: The challenges, and a way forward," *Journal of Network and Computer Applications*, vol. 125, pp. 251–279, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804518303473

[207] X. Wang, X. Zha, W. Ni, R. P. Liu, Y. J. Guo, X. Niu, and K. Zheng, "Survey on blockchain for Internet of Things," *Computer Communications*, vol. 136, pp. 10–29, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366418306881

[208] F. Elhattab, S. Bouchenak, R. Talbi, and V. Nitu, "Robust federated learning for ubiquitous computing through mitigation of edge-case backdoor attacks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–27, 2023.

[209] Y. Gao, Y. Li, L. Zhu, D. Wu, Y. Jiang, and S.-T. Xia, "Not all samples are born equal: Towards effective clean-label backdoor attacks," *Pattern Recognition*, vol. 139, p. 109512, 2023.

[210] P. Rieger, T. D. Nguyen, M. Miettinen, and A.-R. Sadeghi, "DeepSight: Mitigating backdoor attacks in federated learning through deep model inspection," *arXiv preprint arXiv:2201.00763*, 2022.

[211] S. Fu, C. Xie, B. Li, and Q. Chen, "Attack-resistant federated learning with residual-based reweighting," *arXiv preprint arXiv:1912.11464*, 2019.

[212] M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9268–9276.

[213] G. Liu, I. Khalil, A. Khreishah, and N. Phan, "A synergetic attack against neural network classifiers combining backdoor and adversarial examples," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 834–846.

[214] Z. Xiang, D. J. Miller, H. Wang, and G. Kesidis, "Detecting scene-plausible perceptible backdoors in trained dnns without access to the training set," *Neural computation*, vol. 33, no. 5, pp. 1329–1371, 2021.

[215] R. Hou, S. Ai, Q. Chen, H. Yan, T. Huang, and K. Chen, "Similarity-based integrity protection for deep learning systems," *Information Sciences*, vol. 601, pp. 255–267, 2022.

[216] Y. Zhang, T. Zhang, Q. Liu, G. Sun, and H. Wu, "Increasing depth, distribution distillation, and model soup: erasing backdoor triggers for deep neural networks," *Journal of Electronic Imaging*, vol. 31, no. 6, p. 063005, 2022.

[217] S. Hu, Z. Zhou, Y. Zhang, L. Y. Zhang, Y. Zheng, Y. He, and H. Jin, "BadHash: Invisible backdoor attacks against deep hashing with clean label," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 678–686.

[218] C. Fung, C. J. Yoon, and 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.

[219] C. P. Wan and Q. Chen, "Robust federated learning with attack-adaptive aggregation," *arXiv preprint arXiv:2102.05257*, 2021.

[220] Y. Mi, J. Guan, and S. Zhou, "ARIBA: Towards accurate and robust identification of backdoor attacks in federated learning," *arXiv preprint arXiv:2202.04311*, 2022.

[221] K. Dashdondov and M.-H. Kim, "Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction," *Neural Processing Letters*, pp. 1–13, 2021.

[222] X. Li, Z. Xiang, D. J. Miller, and G. Kesidis, "Test-time detection of backdoor triggers for poisoned deep neural networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3333–3337.

[223] S. Lu, R. Li, W. Liu, and X. Chen, "Defense against backdoor attack in federated learning," *Computers & Security*, vol. 121, p. 102819, 2022.

[224] G. Liu, X. Ma, Y. Yang, C. Wang, and 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.

[225] D. M. Sommer, L. Song, S. Wagh, and P. Mittal, "Athena: Probabilistic verification of machine unlearning," *Proceedings on Privacy Enhancing Technologies*, vol. 3, pp. 268–290, 2022.

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

[227] J. Ye, Y. Mao, J. Song, X. Wang, C. Jin, and M. Song, "Safe distillation box," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3117–3124.

[228] Y. Zhao, K. Xu, H. Wang, B. Li, and R. Jia, "Stability-based analysis and defense against backdoor attacks on edge computing services," *IEEE Network*, vol. 35, no. 1, pp. 163–169, 2021.

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

[230] S. Andreina, G. A. Marson, H. Möllering, and 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.

[231] A. Saha, A. Tejankar, S. A. Koohpayegani, and H. Pirsiavash, "Backdoor attacks on self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 337–13 346.

[232] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1354–1371.

[233] F. Tahmasebian, J. Lou, and L. Xiong, "RobustFed: a truth inference approach for robust federated learning," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1868–1877.

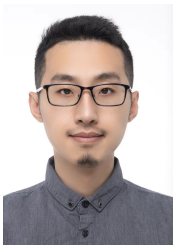
[234] X. Xiao, Z. Tang, C. Li, B. Xiao, and K. Li, "SCA: Sybil-based collusion attacks of iiot data poisoning in federated learning," *IEEE Transactions on Industrial Informatics*, 2022.

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

[236] S. Awan, B. Luo, and F. Li, "CONTRA: Defending against poisoning attacks in federated learning," in *European Symposium on Research in Computer Security*. Springer, 2021, pp. 455–475.

[237] Y. Wan, Y. Qu, L. Gao, and Y. Xiang, "Privacy-preserving blockchain-enabled federated learning for b5g-driven edge computing," *Computer Networks*, vol. 204, p. 108671, 2022.
[238] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Low-latency federated learning and blockchain for edge association in digital twin empowered 6g networks," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 5098–5107, 2020.
[239] N. Furth, A. Khreishah, G. Liu, N. Phan, and Y. Jararweh, "Un-fair trojan: Targeted backdoor attacks against model fairness," in *2022 Ninth International Conference on Software Defined Systems (SDS)*. IEEE, 2022, pp. 1–9.
[240] J. Yang, J. Zheng, Z. Zhang, Q. Chen, D. S. Wong, and Y. Li, "Security of federated learning for cloud-edge intelligence collaborative computing," *International Journal of Intelligent Systems*, vol. 37, no. 11, pp. 9290–9308, 2022.
[241] L. Truong, C. Jones, B. Hutchinson, A. August, B. Praggastis, R. Jasper, N. Nichols, and A. Tuor, "Systematic evaluation of backdoor data poisoning attacks on image classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 788–789.
[242] S. Li, S. Ma, M. Xue, and B. Z. H. Zhao, "Deep learning backdoors," in *Security and Artificial Intelligence*. Springer, 2022, pp. 313–334.
[243] F. Xie, Y. Gao, J. Wang, and W. Zhao, "Defending local poisoning attacks in multi-party learning via immune system," *Knowledge-Based Systems*, vol. 238, p. 107850, 2022.
[244] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 900–14 912, 2021.
[245] C.-L. Chen, S. Babakniya, M. Paolieri, and L. Golubchik, "Defending against poisoning backdoor attacks on federated meta-learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 5, pp. 1–25, 2022.
[246] C. Wu, X. Yang, S. Zhu, and P. Mitra, "Toward cleansing backdoored neural networks in federated learning," in *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2022, pp. 820–830.
[247] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4574–4588, 2021.
[248] F. Nuding and R. Mayer, "Data poisoning in sequential and parallel federated learning," in *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, 2022, pp. 24–34.
[249] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[250] J. Zhou, Z. Zhong, J. Wang, X. Zhang, F. Chen, and C. Yan, "Robust federated learning with adaptable learning rate," in *2021 7th International Conference on Big Data and Information Analytics (BigDIA)*. IEEE, 2021, pp. 485–490.
[251] X. Liu, F. Li, B. Wen, and Q. Li, "Removing backdoor-based watermarks in neural networks with limited data," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 10 149–10 156.
[252] W. Aiken, H. Kim, S. Woo, and J. Ryoo, "Neural network laundering: Removing black-box backdoor watermarks from deep neural networks," *Computers & Security*, vol. 106, p. 102277, 2021.
[253] J. Hayase, W. Kong, R. Somani, and S. Oh, "Defense against backdoor attacks via robust covariance estimation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4129–4139.
[254] C. Chen and J. Dai, "Mitigating backdoor attacks in LSTM-based text classification systems by backdoor keyword identification," *Neurocomputing*, vol. 452, pp. 253–262, 2021.
[255] Z. Xiang, D. J. Miller, and G. Kesidis, "Reverse engineering imperceptible backdoor attacks on deep neural networks for detection and training set cleansing," *Computers & Security*, vol. 106, p. 102280, 2021.
[256] Z. Ma, Y. Liu, X. Liu, J. Liu, J. Ma, and K. Ren, "Learn to forget: Machine unlearning via neuron masking," *IEEE Transactions on Dependable and Secure Computing*, 2022.
[257] H. Hu, Z. Salčić, G. Dobbie, J. Chen, L. Sun, and X. Zhang, "Membership inference via backdooring," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, vol. 23, 2022, pp. 3832–3838.

Yichen Wan is currently pursuing the Ph.D. degree in Information Technology at Deakin University. She received her B.S. degree in Information Technology at RMIT University in 2016, her M.S. degree in Networking at Melbourne Institute of Technology in 2018, and her B.S.(First-Class Hons) degree in Information Technology at Deakin University in 2021. Her research interests focus on Machine Learning, Edge Computing, the Internet of Things, and corresponding security and privacy issues.



Youyang Qu is currently a research scientist of data61, Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. Before joining CSIRO, he served as a research fellow at Deakin University. He received his B.S. degree in Mechanical Automation in 2012 and M.S. degree in Software Engineering in 2015 from the Beijing Institute of Technology, respectively. He received his Ph.D. degree at the School of Information Technology, Deakin University, in 2019. His research interests focus on Machine Learning, Big Data, IoT, blockchain, and corresponding security and customizable privacy issues. He has over 50 publications, including high-quality journals and conference papers such as IEEE TII, IEEE TNSE, ACM Computing Surveys, IEEE IOTJ, etc. He is active in the research society and has served as an organizing committee member in SPDE 2020, BigSecuirty 2021, and Tridentcom 2021/2022.

Wei Ni (Fellow, IEEE) received the B.E. and Ph.D. degrees in communication science and engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. He was a Post-Doctoral Research Fellow at Shanghai Jiao Tong University, Shanghai, from 2005 to 2008; the Deputy Project Manager of the Bell Laboratories, Alcatel/Alcatel-Lucent, Shanghai, from 2005 to 2008; and a Senior Researcher with Devices Research and Development, Nokia, from 2008 to 2009. He is currently the Principal Research Scientist of the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, NSW, Australia; a Conjoint Professor with the University of New South Wales, Sydney; an Adjunct Professor with the University of Technology Sydney, Sydney; and an Honorary Professor with Macquarie University, Sydney. He has authored 8 book chapters, more than 300 journal articles, 100 conference papers, 26 patents, and 10 standard proposals accepted by IEEE. His research interests include machine learning, online learning, stochastic optimization, and their applications to system efficiency and integrity. He has won several research awards, including the 2022 IEEE IWCMC Best Paper Award, the 2022 Elsevier Best Review Paper Award, and the 2021 Elsevier YJNCA Best Review Paper Award, as well as the 2021 IEEE Vehicular Technology Society (VTS) Chapter of the Year Award. Dr. Ni has served first as the Secretary, Vice-Chair, and then Chair for IEEE New South Wales (NSW) VTS Chapter from 2015 to 2023, the Track Chair for VTC-Spring 2017, the Track Co-Chair for IEEE VTC-Spring 2016, the Publication Chair for BodyNet 2015, and the Student Travel Grant Chair for WPMC 2014. He has been an Editor of IEEE Transactions on Wireless Communications since 2018, an Editor of IEEE Transactions on Vehicular Technology since 2022, and an Editor of IEEE Communications Surveys and Tutorials and IEEE Transactions on Information Forensics and Security since 2024.

Yong Xiang (Senior Member, IEEE) received the Ph.D. degree in Electrical and Electronic Engineering from The University of Melbourne, Australia. He is a Professor at the School of Information Technology, Deakin University, Australia. His research interests include distributed computing, cybersecurity and privacy, machine learning and AI, and communications technologies. He has published 7 authored books, over 230 refereed journal articles, and over 100 conference papers in these areas. Professor Xiang is the Senior Area Editor of IEEE Signal Processing Letters, the Associate Editor of IEEE Communications Surveys and Tutorials, and the Associate Editor of Computer Standards and Interfaces. He was the Associate Editor of IEEE Signal Processing Letters and IEEE Access, and the Guest Editor of IEEE Transactions on Industrial Informatics, IEEE Multimedia, etc. He has served as Honorary Chair, General Chair, Program Chair, TPC Chair, Symposium Chair and Track Chair for many conferences, and was invited to give keynotes at numerous international conferences.

Longxiang Gao (SM17) received his PhD in Computer Science from Deakin University, Australia. He is currently a Professor in Shandong Computer Science Center at Qilu University of Technology (Shandong Academy of Sciences). He was a Senior Lecturer at School of Information Technology, Deakin University and a post-doctoral research fellow at IBM Research & Development, Australia. His research interests include Fog/Edge computing, Blockchain, data analysis and privacy protection. Dr. Gao has over 150 publications, including patent, monograph, book chapter, journal and conference papers. Some of his publications have been published in the top venue, such as IEEE TMC, IEEE TPDS, IEEE TSC, IEEE TKDE, IEEE TDSC, IEEE TVT, IEEE TCSS, IEEE TII, IEEE TNSE, IEEE TWC and ACM Computing Surveys. He has being Chief Investigator (CI) for more than 30 research projects (the total awarded amount is over $6 million), from pure research project to contracted industry research. He is a Senior Member of IEEE and ACM.

Ekram Hossain (Fellow, IEEE) is a Professor and the Associate Head (Graduate Studies) of the Department of Electrical and Computer Engineering, University of Manitoba, Canada. He is a Member (Class of 2016) of the College of the Royal Society of Canada. He is also a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada. His current research interests include design, analysis, and optimization beyond 5G/6G cellular wireless networks. He was elevated to an IEEE fellow, for contributions to spectrum management and resource allocation in cognitive and cellular radio networks. He was an Elected Member of the Board of Governors of the IEEE Communications Society for the term 2018–2020. He received the 2017 IEEE ComSoc TCGCC (Technical Committee on Green Communications and Computing) Distinguished Technical Achievement Recognition Award, for outstanding technical leadership and achievement in green wireless communications and networking. He has won several research awards, including the 2017 IEEE Communications Society Best Survey Paper Award and the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award. He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2017-2023. Previously, he served as the Editor-in-Chief (EiC) for the IEEE Press (2018–2021) and the IEEE Communications Surveys and Tutorials (2012–2016). He was a Distinguished Lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society. He served as the Director of Magazines (2020-2021) and the director of Online Content (2022-2023) for the IEEE Communications Society(2020–2021).