Artificial Intelligence Review (2024) 57:150
https://doi.org/10.1007/s10462-024-10796-1

Anomaly detection and defense techniques in federated learning: a comprehensive review

Chang Zhang[1] · Shunkun Yang[2] · Lingfeng Mao[1] · Huansheng Ning[1]

Accepted: 6 May 2024 / Published online: 23 May 2024

Abstract

In recent years, deep learning methods based on a large amount of data have achieved substantial success in numerous fields. However, with increases in regulations for protecting private user data, access to such data has become restricted. To overcome this limitation, federated learning (FL) has been widely utilized for training deep learning models without centralizing data. However, the inaccessibility of FL data and heterogeneity of the client data render difficulty in providing security and protecting the privacy in FL. In addition, the security and privacy anomalies in the corresponding systems significantly hinder the application of FL. Numerous studies have been proposed aiming to maintain the model security and mitigate the leakage of private training data during the FL training phase. Existing surveys categorize FL attacks from a defensive standpoint, but lack the efficiency of pinpointing attack points and implementing timely defenses. In contrast, our survey comprehensively categorizes and summarizes detected anomalies across client, server, and communication perspectives, facilitating easier identification and timely defense measures. Our survey provides an overview of the FL system and briefly introduces the FL security and privacy anomalies. Next, we detail the existing security and privacy anomalies and the methods of detection and defense from the perspectives of the client, server, and communication process. Finally, we address the security and privacy anomalies in non-independent identically distributed cases during FL and summarize the related research progress. This survey aims to provide a systematic and comprehensive review of security and privacy research in FL to help understand the progress and better apply FL in additional scenarios.

✉ Shunkun Yang
ysk@buaa.edu.cn

Chang Zhang
zhangchang@xs.ustb.edu.cn

Lingfeng Mao
lingfengmao@ustb.edu.cn

Huansheng Ning
ninghuansheng@ustb.edu.cn

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Xueyuan Street, Beijing 100083, China

[2] School of Reliability and Systems Engineering, Beihang University, Xueyuan Street, Beijing 100191, China

Keywords Federated Learning · Security · Privacy · Defense · Anomaly detection

1 Introduction

Deep learning techniques employ data saved on a central server and are used for training and testing procedures to create complete machine learning models. They have been extensively applied in numerous sectors to successfully support progress in science and technology, such as in microbiology Qu et al. (2019), health monitoring Yuan et al. (2020), face recognition Sharma et al. (2020), and automatic driving. However, deep learning generally relies on uploading data to central servers, which increases the risk of data privacy breaches. Research on the application of data in various industries is affected by issues such as industry competition, privacy security, and legal regulations, resulting in the phenomenon of data islands where data is distributed among end users. Training a model within a single organization may result in suboptimal outcomes due to limited data types and susceptibility to data bias. For instance, in the scenario of training an X-ray imaging model, data from a single hospital may be limited, yet sharing data faces legal restrictions, posing challenges for collaborative model training.

Federated learning (FL) has been proposed to alleviate the aforementioned difficulties in traditional machine learning. In FL, a central server aggregates the model parameters of the client instead of the client's private data. This significantly alleviates the privacy problem of traditional machine learning; thus, FL is applicable in privacy-involved fields Yang et al. (2019), such as healthcare Antunes et al. (2022) and activity recognition Sozinov et al. (2018). Even though the above FL methods in several areas can bring tremendous value to many institutions, FL can introduce new attack surfaces at training time by enhancing an adversary's capabilities. FL remains vulnerable to abnormal attacks, including security and privacy attacks. Once the FL is attacked, it will cause privacy leakage, model damage, system robustness damage and other adverse effects, and even cause user trust. For instance, FL is vulnerable to poisoning attacks (which attempt to cause convergence to an incorrect model) and Byzantine attacks (which aim to prevent the model from converging). Moreover, relevant studies have been demonstrated that client-supplied models or updates can reveal particular qualities of a client's private data Enthoven and Al-Ars (2021). These techniques allow a malicious server to determine whether a specific data point is used in the training or whether a sample of data points from the training distribution is used (reconstruction attacks). For example, Melis et al. (2019) found that membership data and unintended feature leakages can be captured from shared gradients during a training process. Numerous methods have been proposed to defend against such attacks and address the security and privacy anomalies in FL Xie et al. (2019); Ma et al. (2022). For example, in the context of security attacks, Blanchard et al. (2017) checked for client updates in each iteration and discarded potentially malicious clients as the server aggregated updates. In terms of privacy attacks, Gao et al. (2021) proposed a defense against reconfiguration by searching for privacy-preserving transformation functions and preprocessing training samples with such functions to ensure an excellent performance from the trained model.

Several surveys Mothukuri et al. (2021); Blanco-Justicia et al. (2021); Lyu et al. (2020); Rodríguez-Barroso et al. (2022); Lyu et al. (2020) have summarized some of the threats and defenses in FL. However, these studies have certain limitations. First, the above reviews only consider certain specific branches of the security and privacy aspects in FL. Second, the classifications address the types of attacks and defenses but do not facilitate

linking the attacks in each part of FL, which in turn leads to an inability to facilitate timely actions. Third, most cases present non-independent identically distributed (non-IID) cases of FL; nonetheless, the abovementioned reviews do not consider attack and defense strategies for such cases. Based on the above limitations, this study summarizes the attacks and defenses in FL from the perspectives of the client, central server, and communication process involved in FL to help locate errors and promptly take appropriate defenses in the FL process. Moreover, this study summarizes the attack means and defense methods of FL in non-IID scenarios and improves the FL attack and defense situation. To this end, we summarize our contributions as follows.

• We summarize the attack types for common security and privacy exceptions and corresponding defense methods.

• A novel classification based on a review of anomaly detection and defense in FL is proposed for better locating the security and privacy anomalies in FL, thereby helping users promptly take appropriate protective measures.

• We summarize the FL attack means and defense methods in non-IID scenarios to improve the corresponding situations.

In this study, we summarize the attacks and defenses in the FL process from the perspectives of the client, central server, and communication process involved in FL. This can assist users in locating errors and taking appropriate defenses in the FL process. As non-IID cases present an essential factor affecting the anomaly detection performance in FL, this study provides a separate summary and introduction of the anomaly detection and defense in non-IID FL scenarios. In particular, we systematically introduce the FL systems in Sect. 2. Next, from the perspective of the composition and process of FL, we summarize the security and privacy abnormalities in FL in Sect. 3. We discuss existing abnormal attacks and corresponding defenses in Sect. 4. In addition, we present a state-of-the-art FL exception attack and detection in a non-IID case in Sect. 5, and discuss open problems in FL in Sect. 6. Finally, we conclude this study in Sect. 7.

2 Fedrated learning

Three key elements have contributed significantly to the success of machine learning: the availability of large amounts of data, improved machine processing capabilities, and excellent deep learning models. Despite the enormous success of machine learning Gheibi et al. (2021), the privacy of clients participating in training, scarcity of usable data, and associated laws and regulations have prevented machine learning from being more widely used. FL Li et al. (2020) has been suggested as a solution to these problems.

FL is a machine-learning approach Bonawitz et al. (2019) for training algorithms across a number of centralized or decentralized clients or servers storing local data samples but without transferring any actual data. FL prohibits the uploading of data to a server and cautions against the assumption that examples of local data are spread uniformly. FL alleviates important challenges such as data confidentiality and privacy by allowing several nodes to construct a general and strong machine learning model without exchanging data. Figure 1 shows a schematic of FL. The FL procedure is as follows.

Step 1: A global model is downloaded by the clients from a central server. (**Communication**)

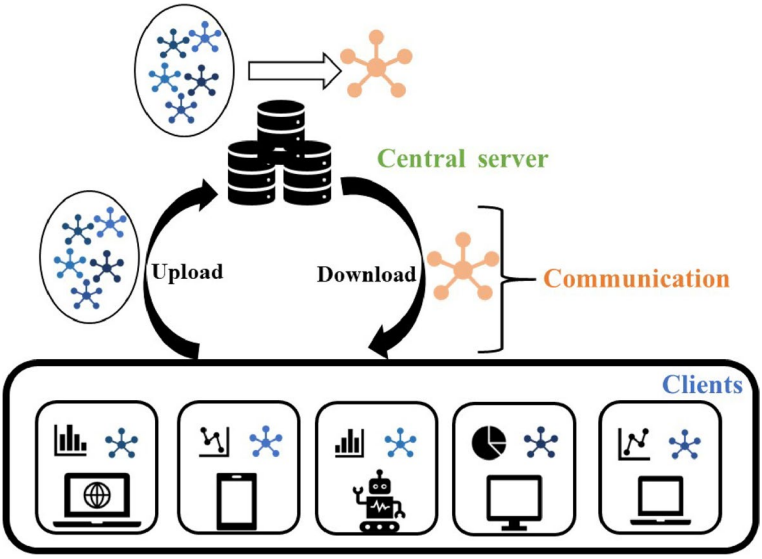

Fig. 1 Federated learning process framework

Step 2: Each client updates the model based on the local data from the training process. (**Clients**)

Step 3: The central server receives model updates from the clients. (**Communication**)

Step 4: The central server combined these updates to create a more accurate global model. (**Central server**)

Until the model converges, these procedures are repeated.

FL has been used in numerous fields Li et al. (2020). In healthcare, FL may be used to protect the privacy of patient data and enhance the ability of machine learning to assist healthcare practitioners. Other applications in healthcare include the use of FL to manage digital health records Brisimi et al. (2018) and to detect attacks in physical medical information systems (i.e., those maintaining sensitive information in patient health records Long et al. (2022)). In intelligent transportation, the application use cases include data exchanges between self-driving cars Liang et al. (2023), preventing vehicle information being stored in physical systems, and predicting traffic flows Lu et al. (2020). FL has achieved significant success in several areas; however, it still faces serious anomalies in security and privacy owing to the complexity of FL systems and unreliability of the client or central server. These security and privacy problems are worse than those in traditional machine learning because the underlying adversaries, which may have thousands of clients, are more difficult to detect and defend. Attacks on FL can have detrimental effects across various dimensions. Firstly, adversaries may engage in model poisoning, injecting malicious data during the training process, thereby compromising model integrity and reducing accuracy. Additionally, these attacks can lead to data privacy breaches, resulting in unauthorized access to sensitive user information and undermining trust in the system. Model inference attacks exploit model outputs to deduce sensitive data about individual

contributors, compromising privacy further. Crafted input data can deceive models, leading to incorrect predictions and severe consequences, especially in sensitive domains like healthcare or finance. Furthermore, denial of service (DoS) attacks disrupt the FL process, causing system unavailability and performance degradation. Certain attacks also lead to resource exhaustion, consuming significant computational resources and increasing operational costs. Lastly, successful attacks can tarnish the reputation of organizations deploying FL, resulting in loss of customer trust and potential legal repercussions. Mitigating these risks necessitates the implementation of robust defense mechanisms, such as secure communication, data encryption, anomaly detection, and adversarial robustness techniques.

According to Fig. 1 and the bolded fonts describing the process of FL Step 1-Step 4, the FL process is primarily divided into three parts: the central server, communication, and client(s). To locate the anomaly more conveniently and take the corresponding defensive measures in time, we describe the security exception, privacy exception, and corresponding defense measures in the FL process from three perspectives: those of the central server, communication process, and client. A summary diagram is shown in Fig. 2.

3 Security and privacy anomalies in FL

FL has made tremendous progress in several areas; however, essential security and privacy issues remain owing to the complexity of FL systems and unreliability of the client or central server. Because the underlying adversaries (with thousands of participants) are harder to discover and defend, these security and privacy issues are worse than those in standard machine learning. Here, we present the security anomalies, privacy anomalies, and related defense measures in the FL process from three perspectives: those of the central server, communication process, and client. Considering all three perspectives allows us to more readily detect the anomaly and implement the relevant defense measures in time. An overall summary is shown in Fig. 2.

From Fig. 2, we can see that a client may suffer from data poisoning, model poisoning, backdoor attacks, Byzantine tine, Sybil attacks, free-riding, and inference attacks. The central server is vulnerable to malicious servers, non-robust aggregation, and

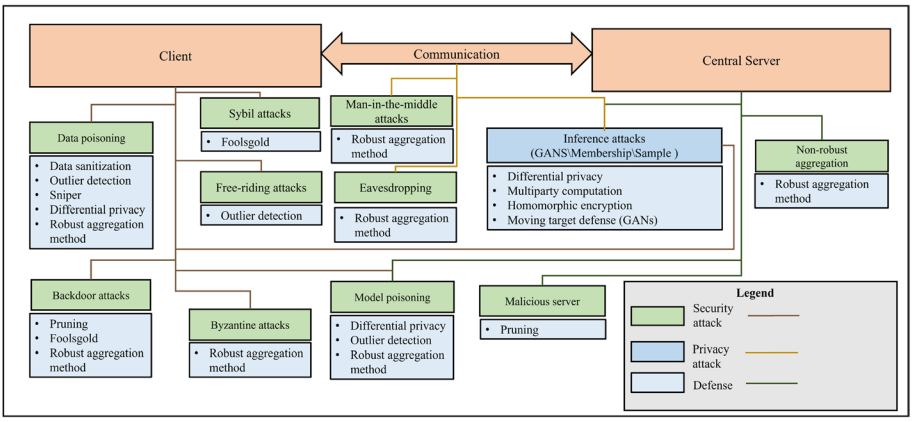

Fig. 2 Attack defense classification diagram for federated learning security and privacy

inference attacks. The communication process of uploading and downloading in FL is vulnerable to man-in-the-middle attacks and eavesdropping. Details are provided as follows.

3.1 FL security and privacy anomalies - clients

In practical application scenarios, the amount of client data in specific FL scenarios is significant, and attackers can use the parameters of the model and training data to gain access to the global model. In FL, client anomalies originate from both privacy and security aspects. Client anomaly attacks mainly include Data poisoning, Model poisoning, Backdoor attacks, Byzantine attacks, Sybil attacks, Free-riding attacks, and Inference attacks.

3.1.1 Data poisoning

In data poisoning attacks in traditional machine learning systems, the attacks attempt to blend hostile data points in the training phase and thereby increase classification mistakes. Biggio et al. initially established the term "data poisoning" Biggio et al. (2012). Corresponding strategies have since been proposed by researchers to defend against data poisoning assaults in traditional machine learning systems.

Due to the fact that during FL the model parameters are transmitted from the clients to the central server, malicious clients can contaminate the global model by uploading incorrect model parameters. Consequently, the data poisoning attacks in FL Tolpegin et al. (2020); Nuding et al. (2022) may be considered as the global model being trained using dirty samples; subsequently, the server receives the produced model parameters. A malicious client can also alter the global model by injecting harmful data into the client's local model and exploiting that data; this can be considered as a subtype of data poisoning Zhao et al. (2021); Lin et al. (2022). Data modification attacks Nasr et al. (2019) combine two classes in a dataset to deceive the machine model into incorrectly categorizing the target class all of the time, such as with feature conflicts Shafahi et al. (2018). The machine learning model can be confused by adding patterns from one class to the target class and performing random label swaps on the dataset (see the data poisoning diagram in Fig. 3).

3.1.2 Model poisoning

In contrast to data poisoning, model poisoning involves a hostile client directly targeting the global model Zhou et al. (2021); Bhagoji et al. (2018). Research Bhagoji et al. (2019) has shown that model positioning attacks significantly impact the model more compared with data poisoning assaults. As shown in Fig. 4, the global model is directly affected by the malicious client's modifications to the client's model, which are often made in FL before the new model is uploaded to the central server Cao et al. (2019). Model poisoning uses gradient manipulation technologies to alter local model gradients, thereby harming the global model's performance and decreasing its overall accuracy Jere et al. (2020); Hong et al. (2020). For instance, when using FL for image recognition, in Li et al. (2021), the classifier of an image model may be altered such that it applies attacker-selected labels to certain aspects of the picture. Model poisoning may also be accomplished by training rule modification strategies to provide attackers access to the trained model. By altering the model's output, the attacker makes the attack invisible; thus, the trained model may be updated as usual Jere et al. (2020); Kairouz et al. (2021). To decrease the inaccuracies

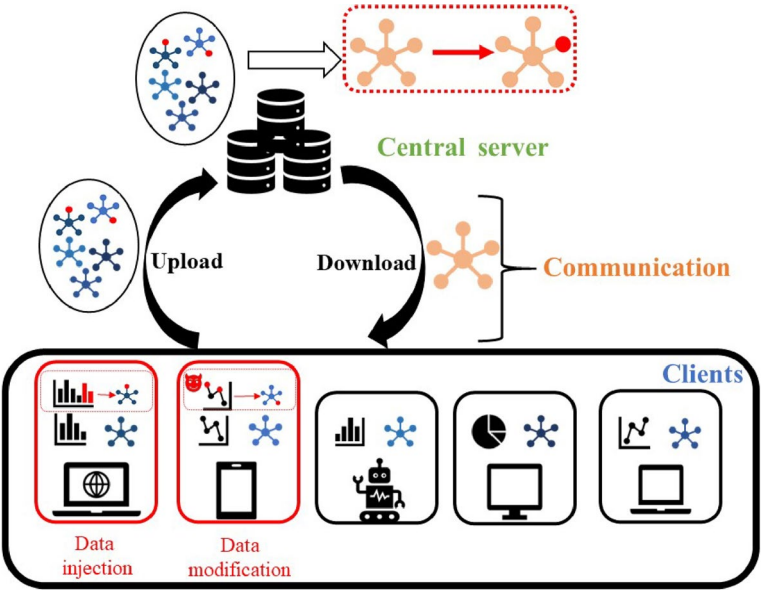

Fig. 3 Data poisoning diagram

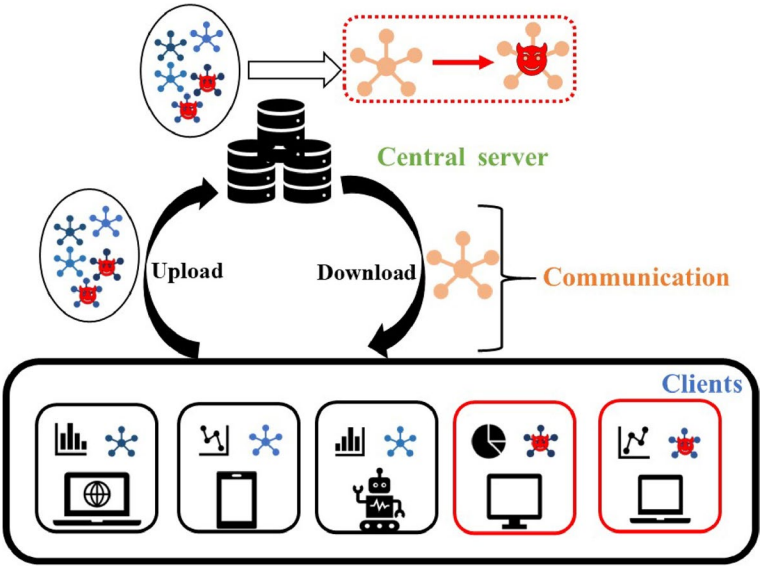

Fig. 4 Model poisoning diagram

between the goal functions and proper weight update distribution, Wahab et al. (2021) introduced penalty terms. The effective deployment of an undetectable targeted model poisoning resulted from this improvement.

3.1.3 Backdoor attacks

One type of model poisoning attack is a backdoor attack. Backdoor attacks Bagdasaryan et al. (2020); Liu et al. (2018) introduce dangerous tasks into already running models while maintaining the integrity of the original tasks. Detecting the abnormalities caused by backdoor assaults is more challenging because the accuracy of the machine learning activities may not be instantly impacted Sun et al. (2019); Wang et al. (2020); Xie et al. (2019). In FL, backdoor attacks operate by maintaining the precision of the primary job while introducing covert backdoors into the global model, as shown in Fig. 5. By training the model on the selected backdoor data, compromised machines participating in the FL can add a backdoor. Backdoor assaults have a catastrophic effect because they can confidently predict false positives. Additionally, clever backdoor assaults in FL can successfully overcome the disastrous forgetting issue Li et al. (2020), thus preventing the backdoor from being omitted when training is being conducted. A Trojan horse threat, which seeks to maintain the current duties of the machine learning model while covertly conducting destructive activities, is an example of a class of backdoor threats comparable to this one Bagdasaryan et al. (2020); Koloskova et al. (2019). In a previous study Bagdasaryan et al. (2020), the specific steps of the backdoor attack were presented.

3.1.4 Byzantine attacks

The phrase "Byzantine fault" is derived from a problem faced by a Byzantine general, and broadly refers to the difficulty in reaching a consensus in a distributed system

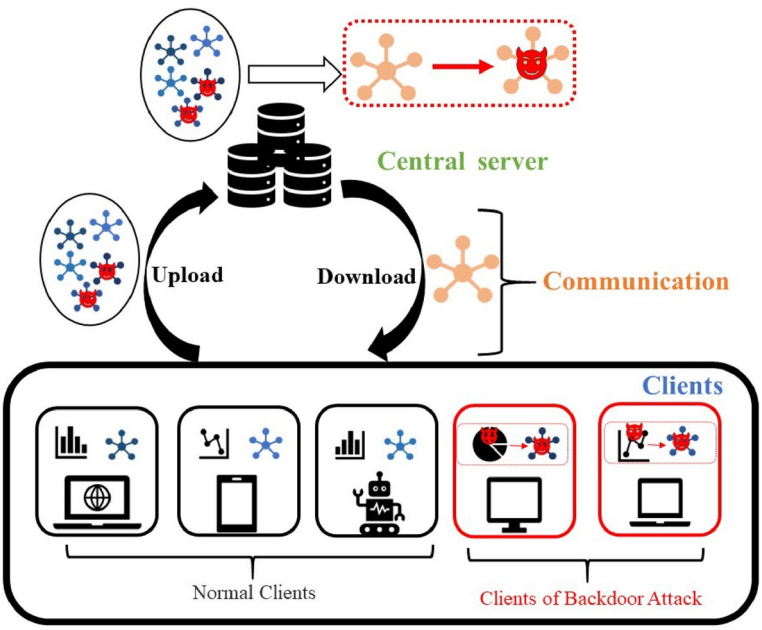

Fig. 5 Backdoor attacks diagram. The attacker compromises one or more clients, trains on the backdoor data, and submits the updating model, which replaces the result from the local training

Lamport et al. (2019). The consensus issue may be caused by malicious clients, attackers who change the information broadcast in the communication channel, or transmission faults between clients. FL is extremely sensitive to Byzantine attacks, in which malicious users modify reliable models or gradients to hinder the learning process or intentionally taint the training data, leading the global model to learn inaccurate information. Blanchard et al. (2017) examined the resistance of distributed implementations of stochastic gradient descent (a fundamental component of FL) to Byzantine assaults. The concept of this attack includes clients who provide incorrect updates to the server, even if they are not necessarily harmful. These flawed updates might result from communication problems, software faults, clients whose data differs from the distribution of the other clients' data, malevolent customers that alter their data, or updates to diverge the global model. Numerous Byzantine assault defense solutions have been proposed to manage this issue and protect FL Fang et al. (2020); Cao et al. (2020); Zhao et al. (2021); Muñoz-González et al. (2019). Although these research efforts have shown some initial progress in thwarting Byzantine assaults, such success remains uncommon. When the data is IID, the gradient updates of benign clients typically fall near the right gradient. However, the Byzantine clients' gradient updates might be random and must be managed using robust estimation techniques to aggregate them. For instance, iterative filtering was suggested in Su and Xu (2018) for robust gradient aggregation, whereas Yin et al. Yin et al. (2018) recommended element-wise median and marginal trimmed operations on gradient updates from clients. Blanchard et al. (2017) proposed a general but computationally expensive algorithm called Krum that, in each iteration, chose the gradient update that, in essence, had the smallest sum of distances from the nearest gradient updates. Chen et al. (2017) propose a geometric median rule for strongly convex settings. Bagdasaryan et al. Bagdasaryan et al. (2020) suggested a powerful model replacement attack for FL with a non-IID data distribution across clients that can manipulate the training process to ensure that the model performs well on the backdoor task of the attacker's choice while also performing well on the primary FL goal. Another study Prakash and Avestimehr (2020) showed that even when each client utilized a mini-batch for computing the gradient, a noisy media technique improved the convergence performance with non-IID data while training with small neural networks and the Modified National Institute of Standards and Technology dataset. Research on Byzantine assaults may concentrate on providing an efficient defense in the client's non-IID scenarios, as FL is applied to a broader range of domains where the client is non-IID.

3.1.5 Sybil attacks

Douceur et al. proposed a Sybil attack in 2002 Douceur (2002); it comprises an attack on peer-to-peer (P2P) networks. To degrade a P2P network's robustness and redundancy, and observe or obstruct regular operations, the attacker utilizes a single node to create numerous false identities. This is because the attacker can defeat trustworthy nodes in the network by a majority of votes if they generate a sufficient number of phony identities. The sequence of transactions may be readily modified in a large-scale assault, i.e., by blocking confirmation of transactions or even reversing them, which might result in issues such as multiple payments. A Sybil attack may also occur in FL. The FL model can be attacked by an attacker posing as a client using FL, thereby reducing the FL efficacy Fun et al. (2018). Fung et al. (2020) considered the vulnerability of FL to Sybil assaults, and presented a taxonomy of Sybil goals and tactics. Jiang et al. (2020) built initial Sybil attacks on differential privacy-based FL architectures and demonstrated how they affected the model

convergence. FL is often used in real-world settings to safeguard the privacy of client data. Studies on assaults in FL environments are currently being conducted in the context of edge computing Jiang et al. (2020). In the future, when FL is used, assaults may occur in communication, healthcare, education, and other vital aspects of human existence.

3.1.6 Free-riding attacks

"Free-riding" attacks Lin et al. (2019) occur when specific clients connect to the FL environment only to use the global model, i.e., without participating in its training. They also insert their subpar models into the global model update without utilizing the local data for training. Such attacks may have a more damaging effect in smaller FL scenarios. Because FL is quickly becoming the de facto training scheme in current cooperative learning programs, free-riding may play an important role in future machine learning applications. FL is reformulated as a stochastic process for characterizing the evolution of aggregated parameters over iterations, Fraboni et al. 2021) created a theoretical framework for the analysis of free-rider attacks in FL systems. Assuring that the training process converges to the desired aim, as represented by the aggregated model of fair customers, is essential for preventing opportunistic free-rider assaults. Their study showed that explicit conditions for ensuring the success of the attack can be derived using the suggested framework.

Free-riding attacks are another concern in FL where some participating clients may intentionally withhold or provide low-quality updates to the central server, exploiting the collective effort of other clients while avoiding their fair share of computational or communication costs. Here's an elaboration on the harm caused by free-riding attacks and the principle of preventing them:

Free-riding Attacks: In FL, free-riding attacks occur when certain clients exploit the collaborative learning process by either providing suboptimal model updates or abstaining from participation altogether, thereby benefiting from the improved global model without contributing proportionally to its improvement. This undermines the fairness, efficiency, and effectiveness of FL, as it distorts the aggregation process and reduces the quality of the global model.

Preventing Free-riding Attacks: The principle of preventing free-riding attacks in FL revolves around incentivizing active and honest participation while deterring malicious or opportunistic behavior. One approach is to implement reputation systems or incentive mechanisms that reward clients for contributing high-quality updates and penalize those engaging in free-riding behavior. Additionally, FL frameworks can employ techniques such as model validation, differential rewards, and client selection strategies to mitigate the impact of free-riding attacks and promote cooperative behavior among participating clients. By incentivizing active engagement and ensuring fairness in the contribution process, FL systems can effectively prevent free-riding attacks and foster collaborative learning environments.

3.2 FL security and privacy anomalies - server

In FL, the central server anomalies originate from privacy and security aspects. As can be seen from Fig. 2, central server anomaly attacks primarily include malicious servers, non-robust aggregation, model poisoning, generalized adversarial networks (GANs), and inference attacks. In general, a central server must be robust, secure, reliable, and safe. In particular, the initial model parameters, local model aggregation, and global model updates

are shared with all clients via the central server. Thus, the physical or cloud-based server chosen for this work should be examined to ensure that no attackers can take advantage of the server's security flaws.

3.2.1 Malicious server

In FL, the concept of a malicious server refers to a scenario where the central server coordinating the learning process is compromised or controlled by an attacker. This poses significant risks to the integrity, privacy, and security of the FL system.

The central server's responsibility in cross-client FL is to aggregate the submitted client models. If the server is malevolent, it can influence the global model by building malicious task-acquired models in the global model via shared aggregation processes. This can have a substantial negative impact on the FL training process. Such as: A malicious server Han et al. (2024) can inject poisoned updates into the FL process. These updates may contain intentionally crafted gradients or parameters designed to undermine the integrity of the global model. By injecting biased or misleading updates, the attacker can manipulate the learning process and compromise the accuracy and reliability of the trained model. Since the central server often aggregates model updates from participating clients, a malicious server can intercept and access sensitive data transmitted during the FL process Guo et al. (2023). This can lead to unauthorized access to user data, violating privacy regulations and compromising the confidentiality of user information. A malicious server may exploit the model's outputs to infer sensitive information about individual data contributors. By analyzing the model's predictions or gradients, the attacker can deduce information about the training data of individual clients, compromising their privacy and confidentiality. A malicious server may tamper with the model aggregation process, manipulating the weights or parameters of the global model to favor certain clients or objectives. This can lead to biased or unfair model updates, undermining the fairness and equity of the FL process Tang et al. (2024). Decentralized and traceable blockchain technology has been employed in several studies to offer a mitigation strategy for malicious server assaults in FL Zhou et al. (2020); Ma et al. (2022).

Mitigating the risks posed by a malicious server in FL requires robust security measures and defense mechanisms. These may include secure communication protocols, data encryption techniques, anomaly detection mechanisms, Byzantine fault-tolerant algorithms, and decentralized governance structures to ensure the integrity, privacy, and security of the FL system. Additionally, continuous monitoring and auditing of the FL process can help detect and mitigate any malicious activities or attacks perpetrated by a compromised central server.

3.2.2 Non-robust aggregation

In FL, the role of the central server is to federate the models from the clients using an aggregation algorithm. However, FL's aggregation techniques have been demonstrated to be subject to adversarial assaults. Yu and Wu (2020); Fu et al. (2019). Using label flipping, backdoor assaults, and noisy updates as examples, a non-robust aggregation technique can result in models that are unavailable and compromised. The central server aggregation algorithm's resilience may be reduced by current protection techniques. The existence of relatively significant differences in the data distributions of clients is inconsistent with the underlying assumptions in the FL environment. In addition, other defense methods (such

as differential privacy algorithms Wei et al. (2020)) may affect the performance of FL and, in turn, lead to anomalous behavior from the global model. Therefore, studying robust FL algorithms for the central servers in FL a is very important task.

3.2.3 Inference attacks

The performance of an FL model is determined by the client's training dataset as a well-trained FL model successfully predicts unseen data using characteristics learned from its training dataset. As such, the types of samples are included in the training data for an FL model can be feasibly determined based on how well the model performs. This poses a serious risk to the client's dataset. After examining the FL model developed using the dataset, an attacker can obtain samples from a commercially accessible dataset. FL is more susceptible to inference assaults compared with traditional machine learning according to certain studies Lee et al. (2021); Hu et al. (2021); Luo et al. (2021) because the training topology discloses parameters when communicating. Next, we discuss inference attacks in detail from the perspectives of membership inference attacks, property inference attacks, and GAN inference attacks.

• Membership Inference attacks. Attacks using a membership inference attempt to determine whether client data are utilized for training the model Truex et al. (2019). For example, in the medical context, an attacker can determine Shokri et al. (2017) if a particular patient profile is used to build a classifier connected to a particular illness Nasr et al. (2019). An attacker can only see the final target model from one participant training in traditional machine learning. Some researchers divide such attacks into passive and active member reasoning attacks. In FL, the attacker can be the parameter server or any of the client nodes Hayes et al. (2017), and the adversary can determine if a certain sample is a part of the private training data of a specific participant or of any participant. Each participant may manage its parameter uploads while keeping track of global parameter updates on the central server, which also controls how each participant views changes to the global parameters over time. Consequently, compared with attacks in classical machine learning, membership inference attacks are simpler to execute on the central server and clients as they have more knowledge regarding the modifications of each iteration.

• Sample Inference attacks. FL systems use a gradient or model parameter-sharing framework to avoid data leakages from participants; however, some studies have shown poor federated results recently launched sample inference attacks. Zhu et al. (2019) reported that sharing gradients can lead to leakages of private training data. TThey devised a gradient depth leakage (DLG) method for quickly acquiring training inputs and labels. DLG is capable of recovering pixel-accurate original pictures as well as label-matching original text. As the original DLG cannot reliably extract ground truth labels and generate high-quality training samples, in Zhao et al. (2020), the authors proposed an analysis method called "improved gradient depth leakage" (iDLG). iDLG extracts labels from shared gradients by exploiting the relationships between labels and corresponding gradient symbols. iDLG works for any differentiable model that has been trained with a cross-entropy loss. To extract high-quality training samples, the attacker can greatly simplify the DLG attack. The sample inference attack described above is built around two key components: a Euclidean cost function and optimization using a limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm. For example, for recovering a single picture from an average gradient, Yin et al. (2021) suggested "GradInvision." GradInvision, in particular, calculates the label from the gradient of the fully connected layer and then optimizes a

random input to match the goal gradient using fidelity regularization to provide a higher-quality reconstructed picture. Dang et al. (2021) reconstructed a gradient using only the last layer of the model to reconstruct the RLG of the training sample (revealing the label from the gradient). Chen et al. (2021) studied a significant batch-size data leakage problem in vertical joint learning. To boost the efficacy of the training samples and labels, bulk sample assaults were utilized. Geiping et al. (2020) proposed the concept of cosine similarity, that is, capturing only information regarding the training state. When asymptotic to another data point, the angle quantifies the change in the prediction over a data point.

• GANs Inference attacks. GANs have been proposed in the field of deep learning, with the goal being that the GANs generate similar samples with the same distribution as in the training data by learning from deep neural networks without accessing the original data samples. For example, in the image domain, GANs initialize the network with random values and then simulate images in discriminative network training data. For FL GANs, we can describe the attacks in terms of both GAN-based client and server attacks. Potential rivals among FL clients are possible; these clients can contribute only to outdated local data in exchange for the global model. They can utilize inference techniques to determine the information of other clients once they have the global model. The adversary generates prototype samples of the target training set by training the GANs in a GAN-based client attack, thereby taking advantage of the real-time nature of the training process and compromising the client's privacy by training the GANs. The adversary engages in FL as an honest client and attempts to extract information regarding a class of data to which he does not belong, as defined in Hitaj et al. (2017), for the GANs to operate. GANs are a subset of active attacks. In an active assault, the antagonist may also affect how the FL causes the afflicted client to show information regarding the intended class. However, GAN-based client assaults have several drawbacks Melis et al. (2019), as follows. To introduce a harmful impact into the learning process, they require alterations in the model design. After several repetitions of the process, the detrimental influence produced by the rogue client may fade into insignificance. Moreover, the attack cannot replicate the precise samples from the victim; it can only imitate the input data used for training. The authors of Wang et al. (2019) suggested multitasking GANs for aided recognition (called mGANAI) to address the shortcomings of client-side GANs-based attack; these operated on the FL server side, and did not interfere with learning. In general, attack enhancements are based on performing extra tasks throughout the GANs' training period. These enhancements raise the caliber of the samples created without interfering with the collaborative learning or changing the shared model, thereby opening the door for covert assaults. Insufficient knowledge exists in the parameter server to properly train a collaborative machine learning model. A previous study Ha et al. (2022) assumed that the participant was the one who was subject to the privacy leakage attack, and compared the success rates of inference attacks from model parameters using GANs models.

3.3 FL security and privacy anomalies - communication

FL uses randomly selected clients to implement an iterative learning process involving a large amount of communication over a given network (e.g., uploading and downloading of the model parameters multiple times). As such, learning the insecure communications in FL is challenging. The communication process-based attacks include man-in-the-middle attacks and eavesdropping.

3.3.1 Man-in-the-middle attacks

Man-in-the-middle interception refers to intercepting model updates between clients and central servers in FL and replacing good models with malicious ones Wang et al. (2020). Man-in-the-middle attacks are performed primarily by jamming operations on communication networks or with artificial networks. The attacker can re-encrypt a hijacked channel to the designed destination by observing it after saving or modifying it; as such, this attack is not easily detectable. Karapanos et al. Karapanos and Capkun (2014) developed "Server Invariance with Strong Client Authentication" (SISCA) by considering transport layer security (TLS) man-in-the-middle attacks in the context of online applications, and used channel ID-based authentication in conjunction with server invariance. No matter how an attacker effectively impersonates a server, SISCA can prevent user impersonation via TLS man-in-the-middle attacks. Wong et al. (2020) presented a plan for a man-in-the-middle attack on Internet of Things (IoT) devices communicating via the MQTT protocol. The two components of this attack methodology are a unique bidirectional encoder representations from transformers-based adversarial model for producing malicious messages using a method inspired by a GAN, and an MQTT parser designed to analyze and modify the MQTT messages at the bit level. As a result of the above research, the current research on man-in-the-middle attacks is focused on communication, and a migration of this research to FL communication will be valuable in the future.

3.3.2 Eavesdropping

In FL, the learning process is iterated through rounds of communication between the clients and central server. Attackers may intercept data if the communication channel is weak. Black-box models are often difficult to attack; thus, eavesdropping can be seen as a medium-severity threat when used to assault FL models. A client with less robust security can be taken over by attackers, thus giving them easy access to a white-box global model and model parameters. Accordingly, the covert communication-based FL (CCFL) strategy was proposed by Yuan et al. (2021). The newly developed CCFL security technique conceals the existence of wireless communication activities. This reduces the attacker's ability to extract useful information from the FL network training protocol (a crucial step in the majority of current attacks), ultimately improving the privacy of FL networks. By using eavesdropping techniques to gather privacy records and deduce a client's identity, Yuan et al. (2021) investigated the effects of data leakage on the effectiveness of natural language processing. Poor communication techniques typically lead to eavesdropping in FL, which is considered a medium-level hazard Mothukuri et al. (2021).

4 Defensive techniques for security and privacy anomalies in FL

Defense technique research may effectively lower the likelihood of danger and assist FL residents in averting security and privacy abnormalities. Defenses are of two types: active and passive. Figure 6 reviews existing FL defensive strategies and the dangers they are designed to counter. We categorize the FL defenses into two groups: FL security assault strategies and privacy defense strategies. The defense mechanisms against

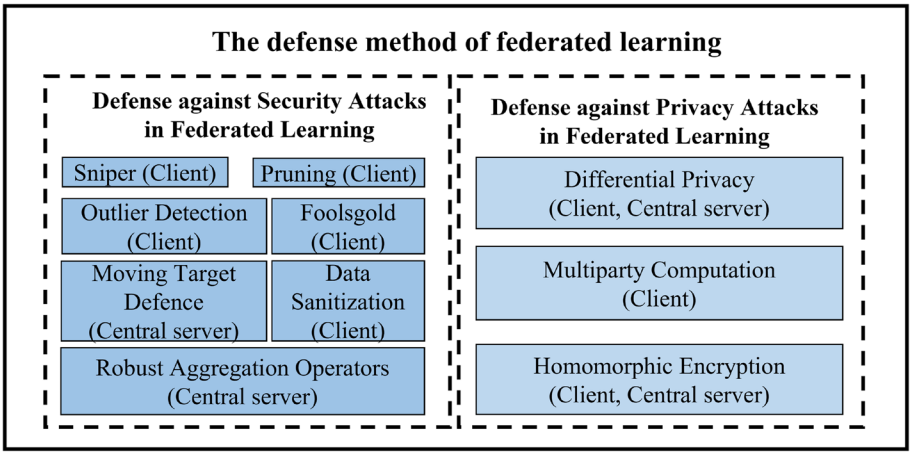

Fig. 6 Federated learning defense method classification

security threats include sniper, pruning, data sanitization, FoolsGold, outlier detection, robust aggregation operators, and moving target protection strategies. The privacy defense techniques include differential privacy, homomorphic encryption, and multiparty computation.

4.1 Defense against security attacks in FL

4.1.1 Sniper

Few studies have explored distributed poisoning attacks in FL, where multiple malicious participants inject poisoned training samples into the training procedures with the same attack goal. It remains unclear whether distributed poisoning with multiple attackers is more effective than traditional poisoning with a single attacker, given the same total number of poisoned training samples. Additionally, the effect of the number of attackers and poisoned samples on the attack success rate is not well-understood. To mitigate such attacks and ensure the integrity of the global model, the server needs to identify a set of honest local models. One approach, such as in Sniper, involves constructing a graph where vertices represent local models collected from participants during an update. If two local models are sufficiently close (i.e., with a relatively small Euclidean distance), an edge exists between them in the graph.The server then identifies honest local models by solving a maximum clique problem in the graph. The global model is obtained by aggregating only those local models contained within the resulting clique. In Cao et al. (2019), Cao et al. suggested the "Sniper" approach for FL, in which dispersed poisoning assaults may pose a more significant threat. This method recognizes legitimate clients and lowers the success rate of poisoning by removing toxic client models from malicious clients. In particular, Sniper creates a benign local model by removing a toxic client's model from a malicious client based on the solution to a maximum-cluster issue. Subsequently, it updates the global model without considering the toxic client's local model.

4.1.2 Pruning

In FL, the majority of model weights contain information regarding the initial job; the weights sensitive to poisoning attacks make up only a small portion thereof. Wu et al. (2020) create a federated pruning approach based on the description provided above by deleting unnecessary neurons from a neural network and modifying outliers in the model. When utilizing the federated pruning method, the accuracy loss brought on by the pruning methodology can be compensated by fine-tuning the validation set. The backdoor attack issue in FL can be successfully mitigated using a federated pruning strategy. Additionally, by limiting the model complexity, the pruning strategy can increase the accuracy. Pruning strategies have been utilized by some researchers Jiang et al. (2022) to increase the effectiveness of FL with regard to edge devices. The authors of Rodríguez-Barroso et al. (2022) summarized their pruning technique as "less is more"; this also represents the pruning strategy's ability to use non-redundant neurons to modify outliers and fight against security assaults in FL.

4.1.3 Data sanitization

In Cretu et al. (2008), the authors proposed the idea of data sanitization. This was achieved primarily using anomaly detectors for identifying data and filtering out problematic data points. For example, the authors in Islam et al. (2022) combined data sanitization techniques with differential privacy techniques in both interactive and non- interactive environments by adding noise to an input function based on a patient's genomic data, thereby obtaining better applications. Some recent works have used data sanitization techniques to achieve better results. Additionally, strong statistical models are used to enhance data sanitization methods; this was shown to be an effective approach in FL Tran et al. (2018) to protect against security assaults (backdoor attacks). Therefore, a data sanitization method can help prevent FL poisoning assaults. However, other studies have created assaults that can bypass various widely used data sanitization safeguards Koh et al. (2018). In a future study, a data sanitization technique can be used with additional protection techniques (such as homomorphic encryption or robust aggregation processes) to improve the defensive outcomes.

4.1.4 FoolsGold

In FL, when multiple attackers participate in the training of the global model, they may provide the same backdoor attack target during the training process. However, this similarity does not appear from benign clients because the data is unique and not shared per client. Thus, after detecting this anomaly, FoolsGold reduces backdoor attacks by protecting the learning rates of benign clients and reducing the learning rates of malignant clients. A malicious client sends a forged updated model to a central server by creating multiple fake identities. This attack can compromise the security of the FL environment. FoolsGold is a defense scheme based on contribution similarity and, thus, adjustment of participant rates; FoolsGold does not require other auxiliary information and makes fewer assumptions regarding the participants. In a previous study Fung et al. (2020), researchers proposed a FoolsGold approach for countering effective Sybil-based

tag-flipping attacks and backdoor poisoning attacks. However, using FoolsGold to defend against adaptive attacks in FL Bagdasaryan et al. (2020) is challenging.

FoolsGold is a FL technique designed to address the challenges of free-riding attacks, where some participating clients may withhold or provide low-quality updates to exploit the collective effort of other clients while avoiding their fair share of computational or communication costs. Here's a detailed analysis of the advantages and disadvantages of the FoolsGold method, along with suitable situations for its application:

FoolsGold offers a range of advantages in FL. It effectively mitigates free-riding attacks by incentivizing honest participation and penalizing opportunistic behavior through a trust-based reward mechanism, thereby improving the overall accuracy and reliability of the model. Additionally, FoolsGold promotes fairness and equity by ensuring proportional client contributions to model improvement, enhancing the integrity and representativeness of the trained model. Furthermore, by incentivizing active engagement and collaboration among clients, FoolsGold leads to higher-quality updates and improved model performance and convergence speed, thereby enhancing the effectiveness and efficiency of FL algorithms. Moreover, FoolsGold incorporates robust defense mechanisms to detect and mitigate malicious behavior, such as free-riding attacks, employing trust scores and adaptive reward mechanisms to enhance system resilience against adversarial manipulation and ensure the integrity of the learning process.

Despite its advantages, the FoolsGold method presents several challenges. Implementation may introduce complexity to FL systems, requiring careful management of trust score computation, reward allocation, and client selection strategies. This complexity may pose deployment challenges in real-world scenarios, while the computational and communication overhead associated with FoolsGold could impact scalability and efficiency, particularly in large-scale deployments. Furthermore, FoolsGold's reliance on trust scores may lead to challenges in accurately estimating them, potentially resulting in unfair reward allocation or ineffective defense against free-riding attacks. However, FoolsGold technology finds suitable application in collaborative learning environments, such as healthcare consortiums or financial institutions, where it promotes fairness and cooperation. It is also well-suited for scenarios involving sensitive data, ensuring privacy while facilitating effective collaboration. Moreover, FoolsGold is apt for large-scale FL deployments, dynamically managing reward mechanisms and client selection to maintain robustness and integrity in the learning process.

4.1.5 Outlier detection

Analytical and statistical techniques are used in outlier detection to identify occurrences deviating from expected patterns or activity. Anomaly detection algorithms can be utilized to spot problematic clients in FL environments. To identify assaults, such as poisoning attacks, the FL server evaluates individual changes and their effects on the global shared model. However, targeted backdoor assaults provide the most significant risk of these defenses failing. In the outlier detection method according to Chen et al. (2017), the central server rebuilds the updated models from the clients and evaluates the model performance metrics against a validation dataset created by combining all updates minus those from the clients. Following that, any client changes that (by some criteria or threshold) reduce the model performance are labeled as outliers.

"AUROR," a defensive mechanism against poisoning attempts in collaborative learning based on K-means clustering, was proposed by Shen et al. Shen et al. (2016) for FL,

essentially aiming to differentiate between benign and suspicious clusters. Sattler et al. (2020) suggested clustering model updates according to the cosine distance. Tolpegin et al. (2020) suggested utilizing an independent component analysis for dimensionality reduction prior to applying anomaly detection as neural network models may be high-dimensional in actual applications. Li et al. (2020) suggested a method for detecting spectral anomalies that involves embedding both original and poisoned samples in a low-latitude space and then finding the samples with significant deviations. Wu et al. (2022) developed the "FadMan" algorithm, i.e., a vertical FL framework proven using five real-world datasets on two tasks (correlated anomaly detection on several attributed networks and anomaly detection on an attributeless network). It was designed for public nodes aligned with numerous private nodes with various features. Nguyen et al. (2019) presented DIoT, an autonomous self-learning distributed system for spotting hacked IoT devices. The system used an FL strategy for anomaly-based intrusion detection. Generally, in FL, the client uploads a local model to the central server. Therefore, we can infer from FL's overall process whether the client significantly impacts the global model. If the client's model is abnormal, it will impact the global model directly or indirectly, and in extreme circumstances, it may compromise the client's privacy. We predict that future studies will focus on the connections between client-side outliers, data dimensionality, and data distributions, based on a combination of the above-mentioned research.

4.1.6 Robust aggregation operators

Defensive approaches typically use statistical and analytical methods to identify unintended patterns. In certain federated aggregation algorithms, outliers can make the results inaccurate or prolong the convergence time. Different robust aggregation techniques can be used in the FL context to detect anomalous attacks. Numerous related studies have been produced. For example, before the aggregation phase, do a clustering action on each client update, the authors in Shen et al. (2016) proposed AUROR (mentioned above) as protection against rogue client updates. By doing this, malicious client updates can be identified. As described in Blanchard et al. (2017), the Krum model employs Euclidean distance to identify client-specific input parameter variations. In [98], the authors spotted unusual updates from customers in FL. In Li et al. (2019), an autoencoder-based anomaly detection defense was developed and aided in spotting fraudulent local model modifications. Variational self-encoders Li et al. (2020) and spectral anomaly detection have also been used Kingma and Welling (2019). In Shafahi (2018), aiming to recognize negative effect updates from clients, loss function-based rejection and error rate-based rejection (ERR) defenses were proposed; these were influenced by current machine learning defenses, such as "RONI" (hostile impact rejection) Barreno et al. (2010) and "TRIM" Jagielski et al. (2018). Table 1 provides a detailed description of these additional techniques.

4.1.7 Moving target defence

The 2009 US National Cyber Leap Year Summit introduced the idea of shifting the target defense. In this context, one proactive defense technique for stopping attacks is called moving-target defense (MTD). MTD provides the best security against server, network, and application infiltration. MTD is a preventative defensive architecture designed to hide sources of vulnerability from attackers. Moving target defense strategies include the use of protection mechanisms, such as IP hopping and virtual IP pooling, for domain name

Table 1 The detailed description of additional techniques

Methods	Description	References
Median	The Median technique is a reliable aggregation approach that chooses the value that symbolizes the center of the entire distribution and switches the original arithmetic mean with the updated model's median	Pillutla et al. (2019) Fu et al. (2019) Hu et al. (2021)
Bulyan	Combining the MultiKrum federated aggregation operator and the trimmed-mean algorithm, it is developed as a federated aggregation operator to guard against poisoning attempts	Fang et al. (2020) Guo et al. (2022) Xie et al. (2019)
Sageflow	A defense that handles both stragglers and enemies at once using staleness-aware grouping, entropy-based filtering, and loss-weighted averaging	Park et al. (2021) Park et al. (2021)
Norm thresholding	It is a robust-aggregation operator, which effectively limits the contribution of each individual update to the aggregated model by clipping the norm of the model updates to a set value	Zhu and Ling (2021) (Zhu and Ling 2021)
Geometric-mean	By employing the product of their values, it indicates the central tendency or the usual value of the data distribution	Zhu et al. (2022) Stripelis et al. (2022) Gouissem et al. (2022)
Trimmed-mean	It is a variation on the arithmetic mean that involves removing a predetermined proportion of extreme values that are both below and above the distribution of the data	Zizzo et al. (2020) Yin et al. (2018)

Table 1 (continued)

Methods	Description	References
Krum and Multi krum	Its foundation is the filtering out of client model changes that exhibit excessive behavior. In order to do so, it ranks the clients based on the geometric distances between the distributions of their model updates, then selects the client that is most similar to the majority as the aggregated model. Multikrum has a parameter that determines how many clients will be combined to create the aggregated model. Under the aggregate model, the majority Multikrum has a parameter that determines how many clients will be combined to create the aggregated model	Wang et al. (2020) Jebreel et al. (2020)
Game-theory approach	The aggregation process is set up as a mixed-strategy game between the server and each client, with the server having the option to accept or reject each client's valid actions of sending good or bad model updates	Chen et al. (2022) Lim et al. (2020)
Residual-based Reweighting	Iteratively Reweighted Least Squares (IRLS), which bases its reweighting technique on reweighting each parameter by its vertical distance (residual) from a robust regression line, is a method for enhancing the median-based aggregation operator	Fu et al. (2019)
Adaptive Federated Averaging (AFA)	AFA technology use the cosine similarity to measure the quality of model updates during training	Reddi et al. (2020) Muñoz-González et al. (2019)

system pings. Another study suggested a novel converter-based moving target defense strategy for defending against deception attacks Liu et al. (2021). This was achieved by deliberately perturbing the primary control gains. The observation that the primary control law of the power converter device in DC microgrids is typically programmable was the foundation of this strategy. To ensure that the MTD remains hidden while optimizing the composite rank of its matrix, efficacy, and coverage of all required buses Tan et al. (2021), a depth-first search-based distribution-flexible alternative current transmission system placement approach was presented. MTDs have been frequently employed in the last five years to thwart eavesdropping Xu et al. (2022); Ghourab et al. (2018, 2022). This has caused researchers in FL to consider the possibility of listening to the client's uploading and downloading of the models. MTDs are also employed to defend against eavesdropping attempts during communication between an FL client and central server. To proactively foil multiple threats during the training process and provide robust security performance for general FL systems, Zhou et al. (2021) recommended an augmented dual-shuffle-based moving-target defense framework. Future FL systems could use moving-target protection in combination with other defense techniques to boost security and privacy.

4.2 Defense against privacy attacks in FL

Although FL defense approaches can address data privacy concerns at the client, the analysis in Sect. 3 suggests that the client may experience privacy leakage risks during local training, client aggregation operations, and throughout the entire FL communication process. At present, differential privacy, homomorphic encryption, and secure multiparty computation techniques are often used to defend against privacy attacks in FL. Figure 7 shows the relationships between the three. From Fig. 7, we can conclude that differential privacy, homomorphic encryption, and secure multiparty computation techniques can be used separately as encryption means in FL. The differential privacy and homomorphic encryption techniques can also be used separately in combination with secure multiparty computation to alleviate the privacy problems in FL. Moreover, differential privacy, homomorphic encryption, and secure multiparty computation techniques can be used together to solve FL privacy problems. In this section, we outline these approaches to mitigating such issues.

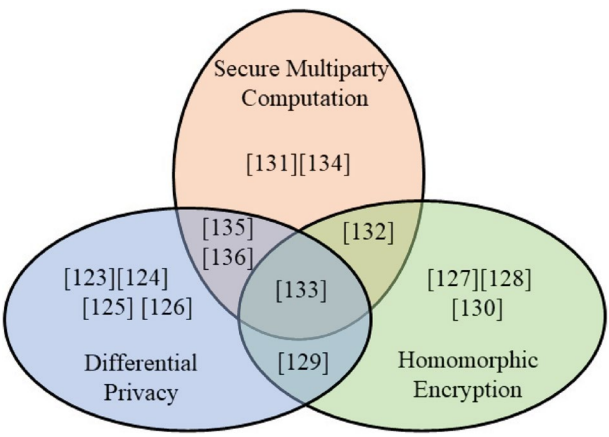

Fig. 7 Federated learning privacy protection technology relationship diagram

4.2.1 Differential privacy

The FL procedure puts clients' privacy at risk at every step. Client information may be revealed when the server communicates, trains, or aggregates. The images illustrate how differential privacy Dwork (2008) preserves privacy by introducing noise into sensitive nformation. Before the model aggregation in FL, noise is added to the client information to preserve privacy Wei et al. (2020). In one study Zhao et al. (2020), the authors combined FL with differential privacy to address the lack of privacy (an approach that has been applied in several ways, such as in the Internet of Vehicles). Subsequently, they experimentally verified their system using a large number of real datasets, as guaranteed by an algorithm. Evidently, the algorithm ensured the system worked while protecting privacy. In a study on electronic medical records in healthcare Choudhury et al. (2019), the authors obtained a higher degree of thorough maintenance. In other studies Wei et al. (2020); Girgis et al. (2021), the authors considered that the inclusion of noise may impact the accuracy and communication efficiency in FL. With a solid balance between accuracy and privacy, differential privacy may also be used in conjunction with tailored FL Hu et al. (2020). As the effectiveness of communication in FL may be impacted by privacy preservation by the addition of noise, Girgis et al. (2021) proposed a distributed communication-efficient and locally differentially private stochastic gradient descent algorithm, and subsequently investigated its communication, privacy, and convergence trade-offs.

Differential privacy serves as a framework for assessing the privacy assurances of statistical databases and algorithms, offering a structured approach to quantifying the privacy safeguarding of individuals' data within datasets or algorithmic processes. Its advantages include robust privacy protection by minimizing the influence of individual data points on query outcomes, flexibility for integration into diverse data processing pipelines and analytical methodologies, and provision of quantifiable privacy assurances, empowering stakeholders to navigate the privacy-utility trade-offs effectively. Additionally, adaptive mechanisms within differential privacy facilitate privacy level adjustments tailored to data sensitivity and application requirements. However, there are challenges, including potential utility trade-offs due to the introduction of noise, complexity in implementation and comprehension, computational overhead from privacy mechanisms, and limitations in protecting against all privacy threats, particularly auxiliary information or side-channel attacks. Despite these challenges, differential privacy remains a potent tool for balancing data analysis needs with individual privacy protection imperatives. The schematic of FL combined with differential privacy approach is shown in Fig. 8.

4.2.2 Homomorphic encryption

The FL process involves multiple client–server communications; for example, the client needs to upload the local model to the server, and the server needs to pass the aggregated model to the client without class participation. If the communication channel is not secure, an attacker can eavesdrop on the FL model information and embed toxic models or benign poison models, thus directly affecting the effectiveness of FL. Therefore, providing communication security is essential in FL. In FL, homomorphic encryption is typically used when the server and client perform model updates, and it is intended to protect the privacy of client data. A framework for FL based on partial homomorphic encryption was proposed by Fang and Qian (2021); it aimed to transmit encryption gradients using only

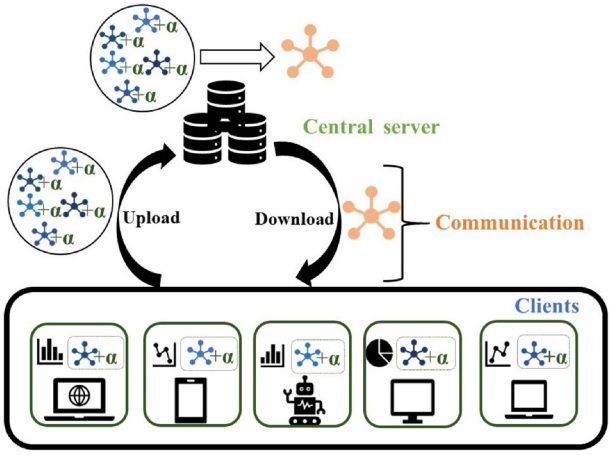

Fig. 8 Schematic of FL combined with differential privacy approach. α represents the random noise

homomorphic encryption for all participants. To address the membership inference problem in FL, Park and Lim (2022) trained a model using fully homomorphic encryption and reported that the encrypted FL model and unencrypted FL model performed slightly differently. To address the hacking problem in an industrial IoT, Jia et al. (2021) proposed a distributed K-means clustering approach based on differential privacy and homomorphic encryption to achieve multiple protection levels for shared models. However, homomorphic encryption increases communication and computational costs. Zhang et al. (2020) addressed this drawback by encoding a batch of quantized gradients using a long integer method instead of encrypting individual gradients for one-time encryption. Furthermore, adding a gradient cropping technique alleviated the problem of the high communication and computational costs for homomorphic encryption. In general, homomorphic encryption requires additional communication and computation overhead, and future research work can explore methods for reducing such overhead. At present, the homomorphic encryption technology in FL has been applied in healthcare Zhang et al. (2022); Wibawa et al. (2022), Internet of things Hijazi et al. (2023), blockchain Jia et al. (2021)and other field Madi et al. (2021), and has achieved good results. The schematic of FL combined with homomorphic encryption approach is shown in Fig. 9.

4.2.3 Secure multiparty computation

The concept of secure multiparty computation was first introduced in Canetti et al. (1996) to protect the inputs of the multiple participants in a centrally computed function or model. One drawback of secure multiparty computation is that it adds communication overhead for the client, which may burden some clients. Unlike traditional secure multiparty computation, the secure multiparty computation in FL needs to encrypt only the client parameters. It does not need to transmit a large amount of client data; this makes the computation more efficient and has led to this approach being widely used in FL. One study Aono et al. (2017) investigated the information leakage problem during central server and client updates in FL. They combined asynchronous stochastic gradient descent with encryption to prevent data leakage from the central server. The client updates were also encrypted to prevent leakage of client data. However, protecting client data using encryption in certain application domains may adversely affect the model. Therefore, efficiency and privacy

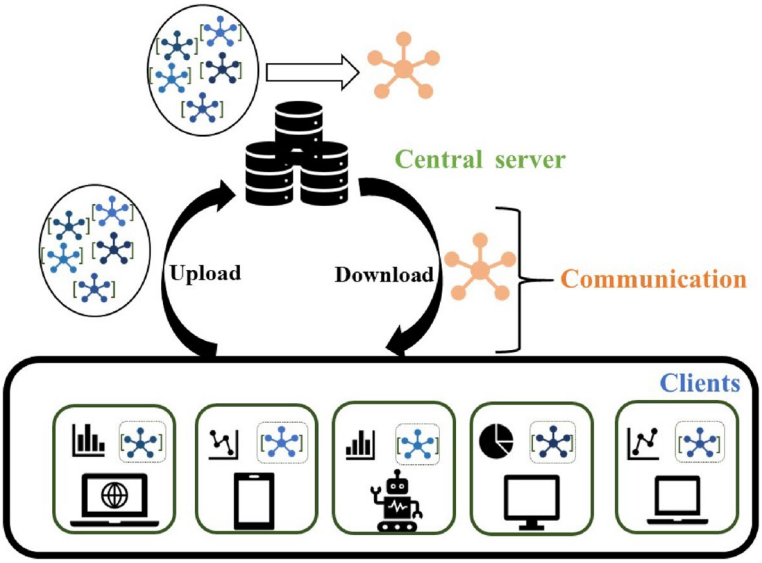

Fig. 9 Schematic of federated learning combined with homomorphic encryption approach [] represents encrpytion tachniques

protection must be balanced in FL. In Hao et al. (2019), the authors combined homomorphic encryption and differential privacy such that the experimental results ensured the accuracy of the model while protecting the clients engaged in FL. The authors in Bonawitz et al. (2017) proposed an efficient and secure aggregation method responsive to both active and passive attacks. The authors in Li et al. (2020) proposed a secure multiparty computation-based learning framework that protects the client's output. According to their complexity and security features, the authors in Goryczka et al. (2013) compared protocols with various privacy techniques and security systems. In general, Although secure multiparty computation can alleviate the privacy issues in FL, it poses some challenges. First, secure multiparty computation-based models require more time compared with traditional FL training models. The tradeoff between efficiency and security is another significant challenge. Further, designing lightweight secure multiparty computation solutions for FL customers is also a challenge.

5 Defenses in FL for non-independent identically distributed case

Owing to variations in the equipment used to gather the data and conditions of subjects, client data tends to be unevenly dispersed for each client in a real-world FL situation. As such, the training process often performs poorly, and several academics have suggested fixes Wang et al. (2020); Huang et al. (2021); Shoham et al. (2019). Similar to this, client data may be non-IID, thus rendering difficulty in differentiating between malicious and helpful customers. In this regard, anomaly identification in FL is highly challenging. For instance, most Byzantine resilience algorithms currently used consider only IID training situations when using an IID-only training dataset with benign players. However, because the quality and distribution of each training dataset differ, the training dataset is usually

non-IID. Defense strategies in FL for the non-IID case encounter several challenges due to the unique characteristics of this scenario. Firstly, the heterogeneity of data distribution poses a significant obstacle, as data distributions across clients in non-IID settings are diverse and varied, making it difficult to develop defense strategies that can effectively generalize across all clients while accommodating these distributional differences. Secondly, an imbalance in data availability among clients further complicates matters, potentially leading to biased models or vulnerabilities, as defense mechanisms may not adequately address the representation of all clients' data. Thirdly, privacy concerns are heightened in non-IID settings, as models trained on heterogeneous data may inadvertently leak sensitive information about individual clients, necessitating novel privacy-preserving approaches that balance privacy and model performance effectively. Additionally, the presence of non-IID data distributions may result in model drift or concept shift, where the underlying relationships between features and labels change over time or across clients, requiring robust defense strategies to adapt to such changes and ensure continued model effectiveness. Moreover, implementing defense strategies in FL often entails exchanging additional information or performing complex computations during model aggregation or update processes, which can increase communication overhead and resource requirements, impacting system scalability. Furthermore, non-IID data distributions may hinder the generalization performance of FL models, necessitating defense strategies to ensure models can generalize well to unseen data and perform reliably across all clients. Lastly, non-IID settings may be more susceptible to adversarial attacks, where malicious clients intentionally manipulate their data to undermine system integrity or performance, requiring resilient defense strategies to safeguard against such attacks and ensure model robustness. Addressing these difficulties requires a holistic approach that considers the unique challenges posed by non-IID data distributions in FL, involving tailored defense mechanisms, robust privacy-preserving techniques, and strategies to mitigate the impact of data heterogeneity on model performance and security. The defense strategies in FL for non-IID situations are primarily covered in this section.

Studies on the attack and defense means for non-IID scenarios in FL are limited. Because Byzantine attacks are common in FL, most studies focus on Byzantine attacks and defenses. Most state-of- the-art approaches exploit similar updates from benign clients to mitigate Byzantine behaviors in FL. However, in numerous FL scenarios, the data between clients is non-IID; for non-IID data, current approaches are ineffective in defending against Byzantine attacks. Some studies on Byzantine attacks in the case of non-IID client data are as follows. Zhai et al. (2021) came up with an FL Byzantine robustness architecture based on a non-IID data confidence assessment. First, an adaptive anomaly detection model was combined with data validation to design a trustworthiness assessment method for Byzantine attacks. In particular, an adaptive mechanism was added to the anomaly detection model to train and predict the model. Finally, the global model was given a consistent orientation using a unified updating mechanism. Elkord et al. Elkordy et al. (2022) presented a proposal for Basil, a quick and computationally effective Byzantine resilient method for dispersed training situations. Their proposed approach uses a new memory-aided, and performance-based criterion for training on logical rings while filtering Byzantine users. The anonymous circular data sharing strategy, which allows each node to exchange a random percentage of its local non-sensitive dataset anonymously (for example, landmark photos) with all other nodes, has also been used to expand Basil to non-IID dataset distribution situations. Prakash and Avestimehr (2020) proposed DiverseFL for overcoming this challenge in a heterogeneous data distribution setting. In particular, the FL server in DiverseFL computes a bootstrap gradient for each client in each iteration; this is for a small sample of

the client's local data and is received only once before the training begins. Subsequently, the server flags Byzantine updates using a new per-client criterion, updates the model by comparing the corresponding bootstrap gradients with the client updates, and updates the model using the gradients received from the non-flagged clients. Guo et al. (2021) proposed Siren, a Byzantine-robust FL system using an active alerting mechanism. Siren can resist attacks from a higher percentage of malicious clients in the system while maintaining a global model. It allows the production of digital or hard copies for personal or classroom use of all or part of the work without charging any fees for IID and non-IID data under different settings for different attack methods; it is extensively experimented with. The experimental results demonstrated the effectiveness of Siren with several advanced defense methods.

In addition to the abovementioned methods, distillation learning is another method for solving the heterogeneity problem in the FL process. Researchers have combined FL defense methods and distillation learning to address a problem in anomaly detection and defense of the FL process when the client data is non-IID. For example, the federated robust adaptive distillation (FedRAD) technique proposed by Sturluso et al. (2021) executes an enhanced integrated knowledge distillation after detecting adversaries and robustly aggregating local models based on the characteristics of median statistics. Pertinent experiments demonstrated that FedRAD performed superior to all other aggregators in the presence of adversaries and various data distributions. In their examination of the combined issue of non-IID and long-tail data in FL, Shang et al. (2022) presented a solution in the form of federated distillation with imbalance calibration (FEDIC). Utilizing a variety of models developed on non-IID data, FEDIC employs model integration to manage non-IID data. Based on this, a novel distillation technique for successfully addressing the long-tail problem was suggested. It included calibration gating networks and logit correction. Wen et al. (2020) suggested robust joint augmentation and distillation as a two-step FL paradigm for preserving privacy, providing efficient communication, and facilitating Byzantine-intolerant on-device machine learning in wireless communications.

Other strategies for FL non-IID attack and defense are as follows. In one approach, each participant in FL uses its data to train a client's local model, and a global model is created on a reliable server by combining model updates from all clients. However, because the server has no control over or access to the training processes of participants, the global model is vulnerable to assaults such as data poisoning and model poisoning. A protection method named BARFED was suggested by Isik-Polat et al. (2021); BARFED makes no assumptions regarding the distribution of data, the similarity of participants' updates, or the percentage of malevolent participants. At each model design layer, BARFED considers the state of the outliers in participant updates, depending predominantly on the distance from the global model. FL with incremental clustering was proposed by Espinoza Castellon et al.(2022); it allows the server to benefit from client changes during federation training instead of requiring them to concurrently communicate the parameters. Thus, no further communication between the server and the client is necessary beyond that required for traditional FL. The method successfully separates customers into numerous groups based on the same data distribution for various non-IID scenarios. A client-based defense called FL white blood cell (FL-WBC) was suggested by Sun et al. Sun et al. (2021) as a way to counteract model poisoning assaults that have broken a the model from central server. The fundamental principle behind FL-WBC is to locate the parameter space where a long-term assault impact on the parameters exists and disturb that space during the local training phase. Within five communication rounds and with very little accuracy loss,

this technique can successfully mitigate the impacts of model poisoning attacks on the global model in both IID and non-IID setups. To identify hostile clients and guarantee aggregation quality, an FL poisoning attack detection approach was suggested in You et al. (2022). The technique uses a reputation system to identify attacker clients step-by-step and filters abnormal models based on similar client history changes. Experiments revealed that the strategy considerably boosted the global model's performance, even if the percentage of malicious clients wass more than 30 percent.

6 Open problems

How can privacy and efficiency be balanced? We discovered that with the majority of defenses, balancing preventing privacy attacks while maintaining the performance of the original model is challenging. For instance, in those based on differential privacy, a significant amount of noise must be included to preserve the data privacy, thus materially reducing the model performance Girgis et al. (2021). A homomorphic encryption-based privacy-preserving model increases the original FL model transmission burden and computational pressure on the central server Jia et al. (2021). Therefore, creating more effective privacy-preserving techniques and expanding privacy-preserving techniques to include defenses against all adversarial attacks remain challenging.

How can additional Byzantine non-IID scenarios be defended against? The majority of Byzantine robust algorithms currently in use consider only IID training scenarios when training datasets with benign users. However, because each training dataset's quality and distribution may vary, the training datasets are non-IID in most real-world scenarios. Therefore, defenders must work harder to differentiate between good and bad updates. Present approaches only consider a small number of non-IID scenarios, although various publications have sought to offer Byzantine robust aggregation methods in non-IID scenarios Zhai et al. (2021); Guo et al. (2021).

How should multiple cross-attacks be addressed? Security and privacy attacks are two primary forms of attack on FL. However, methods used in the current study are focused only on one type of attack at once; they have not yet been examined simultaneously. In the future, techniques to guard against concurrent security and privacy assaults can be researched. One strategy is to implement robust anomaly detection techniques to identify and mitigate different types of attacks. Additionally, enhancing the security protocols within the FL framework can help prevent unauthorized access and manipulation of the system. Regular monitoring and evaluation of the FL system's security posture are also essential to adapt defenses to evolving attack strategies. Overall, a combination of proactive measures, collaborative efforts, and continuous monitoring is crucial to effectively address multiple cross-attacks in FL.

How can clients' contributions be calculated? Properly distributing client contributions in FL is crucial for fairness, accuracy, and efficiency in model training. One effective method is weighted aggregation, where each client's contribution is weighted based on factors like data quality, model performance, and trustworthiness. This weighting can be determined by considering factors such as the amount and quality of data contributed, the performance of the client's model updates, and the trustworthiness of the client based on past behavior and adherence to protocol.

7 Conclusion

Although FL contains significant security and privacy hazards, it remains a viable method for ensuring that several parties work together to train models while minimizing client data privacy breaches. In this study, we described common security and privacy attacks on FL from the perspectives of the client, central server, communication, and the corresponding defenses for locating attacks encountered during FL and taking appropriate defenses in time. We also provided an overview of attacks and defenses against FL in the non-IID context. Additionally, we addressed a few unresolved concerns with current defense techniques with hope that doing so may speed up research on strong and privacy-preserving FL. FL is a relatively new concept in machine learning, and further research and development are needed before it can be reliably applied in delicate applications.

Author contributions Chang Zhang: Data curation, Writing - original draft. Shunkun Yang: Software, Data curation. Lingfeng Mao: Conceptualization,Methodology, Software. Huansheng Ning: Writing -review & editing All authors reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B (2022) Federated learning for healthcare: systematic review and architecture proposal. ACM Trans Intel Syste Technol (TIST) 13(4):1–23

Aono Y, Hayashi T, Wang L, Moriai S (2017) Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans Inf Forensics Secur 13(5):1333–1345

Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How to backdoor federated learning. In: International Conference on Artificial Intelligence and Statistics, 2938–2948. PMLR

Barreno M, Nelson B, Joseph AD, Tygar JD (2010) The security of machine learning. Mach Learn 81(2):121–148

Bhagoji AN, Chakraborty S, Mittal P, Calo S (2018) Model poisoning attacks in federated learning. In: Proc. Workshop Secur. Mach. Learn.(SecML) 32nd Conf. Neural Inf. Process. Syst.(NeurIPS), 1–23

Bhagoji AN, Chakraborty S, Mittal P, Calo S (2019) Analyzing federated learning through an adversarial lens. In: International Conference on Machine Learning, 634–643. PMLR

Biggio B, Nelson B, Laskov P (2012) Poisoning attacks against support vector machines. arXiv preprint arXiv:1206.6389

Blanchard P, El Mhamdi EM, Guerraoui R, Stainer J (2017) Machine learning with adversaries: byzantine tolerant gradient descent. Adv Neural Inf Proc Syst. https://doi.org/10.48550/arXiv.1703.02757
Blanco-Justicia A, Domingo-Ferrer J, Martínez S, Sánchez D, Flanagan A, Tan KE (2021) Achieving security and privacy in federated learning systems: Survey, research challenges and future directions. Eng Appl Artif Intel 106:104468
Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, Kiddon C, Konečnỳ J, Mazzocchi S, McMahan B (2019) Towards federated learning at scale: system design. Proc Mach Learn Syst 1:374–388
Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K (2017) Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191
Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W (2018) Federated learning of predictive models from federated electronic health records. Int J Med Inf 112:59–67
Canetti R, Feige U, Goldreich O, Naor M (1996) Adaptively secure multi-party computation. In: Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing, 639–648
Cao D, Chang S, Lin Z, Liu G, Sun D (2019) Understanding distributed poisoning attack in federated learning. In: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), 233–239. IEEE
Cao X, Fang M, Liu J, Gong NZ (2020) Fltrust: Byzantine-robust federated learning via trust bootstrapping. arXiv preprint arXiv:2012.13995
Chen Y, Su L, Xu J (2017) Distributed statistical machine learning in adversarial settings: byzantine gradient descent. Proc ACM Measur Anal Comput Syst 1(2):1–25
Chen Y, Zhang Y, Wang S, Wang F, Li Y, Jiang Y, Chen L, Guo B (2022) Dim-ds: dynamic incentive model for data sharing in federated learning based on smart contracts and evolutionary game theory. IEEE Internet Things J 9:24572–24584
Chen S, Kahla M, Jia R, Qi G-J (2021) Knowledge-enriched distributional model inversion attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 16178–16187
Choudhury O, Gkoulalas-Divanis A, Salonidis T, Sylla I, Park Y, Hsu G, Das A (2019) Differential privacy-enabled federated learning for sensitive health data. arXiv preprint arXiv:1910.02578
Cretu GF, Stavrou A, Locasto ME, Stolfo SJ, Keromytis AD (2008) Casting out demons: Sanitizing training data for anomaly sensors. In: 2008 IEEE Symposium on Security and Privacy (sp 2008), 81–95. IEEE
Dang T, Thakkar O, Ramaswamy S, Mathews R, Chin P, Beaufays F (2021) Revealing and protecting labels in distributed training. Adv Neural Inf Proc Syst 34:1727–1738
Douceur JR (2002) The sybil attack. International Workshop on Peer-to-peer Systems. Springer, Berlin, pp 251–260
Dwork C (2008) Differential privacy: a survey of results. In: International Conference on Theory and Applications of Models of Computation. Springer, Berlin, pp 1–19
Elkordy AR, Prakash S, Avestimehr S (2022) Basil: a fast and byzantine-resilient approach for decentralized training. IEEE J Selected Areas Commun 40(9):2694–2716
Enthoven D, Al-Ars Z (2021) An overview of federated deep learning privacy attacks and defensive strategies. Fed Learn Syst. https://doi.org/10.1007/978-3-030-70604-3_8
Espinoza Castellon F, Mayoue A, Sublemontier J-H, Gouy-Pailler C (2022) Federated learning with incremental clustering for heterogeneous data. arXiv e-prints, 2206
Fang H, Qian Q (2021) Privacy preserving machine learning with homomorphic encryption and federated learning. Future Internet 13(4):94
Fang M, Cao X, Jia J, Gong N (2020) Local model poisoning attacks to {Byzantine-Robust} federated learning. In: 29th USENIX Security Symposium (USENIX Security 20), 1605–1622
Fraboni Y, Vidal R, Lorenzi M (2021) Free-rider attacks on model aggregation in federated learning. In: International Conference on Artificial Intelligence and Statistics, 1846–1854. PMLR
Fung C, Yoon CJ, Beschastnikh I (2018) Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866
Fung C, Yoon CJ, Beschastnikh I (2020) The limitations of federated learning in sybil settings. In: 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020), 301–316
Fu S, Xie C, Li B, Chen Q (2019) Attack-resistant federated learning with residual-based reweighting. arXiv preprint arXiv:1912.11464
Gao W, Guo S, Zhang T, Qiu H, Wen Y, Liu Y (2021) Privacy-preserving collaborative learning with automatic transformation search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 114–123
Geiping J, Bauermeister H, Dröge H, Moeller M (2020) Inverting gradients-how easy is it to break privacy in federated learning? Adv Neural Inf Proc Syst 33:16937–16947

Gheibi O, Weyns D, Quin F (2021) Applying machine learning in self-adaptive systems: a systematic literature review. ACM Trans Auton Adaptive Syst (TAAS) 15(3):1–37
Ghourab EM, Bariah L, Muhaidat S, Sofotasios PC, Al-Qutayri M, Damiani E (2022) Blockchain-enabled moving target defense for secure cr networks. In: 2022 International Telecommunications Conference (ITC-Egypt), 1–6. IEEE
Ghourab EM, Samir E, Azab M, Eltoweissy M (2018) Diversity-based moving-target defense for secure wireless vehicular communications. In: 2018 IEEE Security and Privacy Workshops (SPW), 287–292. IEEE
Girgis A, Data D, Diggavi S, Kairouz P, Suresh AT (2021) Shuffled model of differential privacy in federated learning. In: International Conference on Artificial Intelligence and Statistics, 2521–2529. PMLR
Goryczka S, Xiong L, Sunderam V (2013) Secure multiparty aggregation with differential privacy: A comparative study. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops, 155–163
Gouissem A, Abualsaud K, Yaacoub E, Khattab T, Guizani M (2022) Federated learning stability under byzantine attacks. In: 2022 IEEE Wireless Communications and Networking Conference (WCNC), 572–577. IEEE
Guo X, Wang P, Qiu S, Song W, Zhang Q, Wei X, Zhou D (2023) Fast: Adopting federated unlearning to eliminating malicious terminals at server side. IEEE Transactions on Network Science and Engineering
Guo H, Wang H, Song T, Hua Y, Lv Z, Jin X, Xue Z, Ma R, Guan H (2021) Siren: Byzantine-robust federated learning via proactive alarming. In: Proceedings of the ACM Symposium on Cloud Computing, 47–60
Guo Q, Wu D, Qi Y, Qi S, Li Q (2022) Flmjr: Improving robustness of federated learning via model stability. In: European Symposium on Research in Computer Security, 405–424. Springer
Ha T, Dang TK (2022) Inference attacks based on gan in federated learning. Int J Web Inf Syst 18:117–136
Han Q, Lu S, Wang W, Qu H, Li J, Gao Y (2024) Privacy preserving and secure robust federated learning: a survey. Concurr Comput. https://doi.org/10.1002/cpe.8084
Hao M, Li H, Xu G, Liu S, Yang H (2019) Towards efficient and privacy-preserving federated deep learning. In: ICC 2019-2019 IEEE International Conference on Communications (ICC), 1–6. IEEE
Hayes J, Melis L, Danezis G, De Cristofaro E (2017) Logan: Membership inference attacks against generative models. arXiv preprint arXiv:1705.07663
Hijazi NM, Aloqaily M, Guizani M, Ouni B, Karray F (2023) Secure federated learning with fully homomorphic encryption for iot communications. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2023.3302065
Hitaj B, Ateniese G, Perez-Cruz F (2017) Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 603–618
Hong S, Chandrasekaran V, Kaya Y, Dumitraş T, Papernot N (2020) On the effectiveness of mitigating data poisoning attacks with gradient shaping. arXiv preprint arXiv:2002.11497
Hu R, Guo Y, Li H, Pei Q, Gong Y (2020) Personalized federated learning with differential privacy. IEEE Internet Things J 7(10):9530–9539
Huang Y, Chu L, Zhou Z, Wang L, Liu J, Pei J, Zhang Y (2021) Personalized cross-silo federated learning on non-iid data. In: AAAI, 7865–7873
Hu S, Lu J, Wan W, Zhang LY (2021) Challenges and approaches for mitigating byzantine attacks in federated learning. arXiv preprint arXiv:2112.14468
Hu H, Salcic Z, Sun L, Dobbie G, Zhang X (2021) Source inference attacks in federated learning. In: 2021 IEEE International Conference on Data Mining (ICDM), 1102–1107. IEEE
Isik-Polat E, Polat G, Kocyigit A (2021) Barfed: Byzantine attack-resistant federated averaging based on outlier elimination. arXiv preprint arXiv:2111.04550
Islam TU, Ghasemi R, Mohammed N (2022) Privacy-preserving federated learning model for healthcare data. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 0281–0287. IEEE
Jagielski M, Oprea A, Biggio B, Liu C, Nita-Rotaru C, Li B (2018) Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP), 19–35. IEEE
Jebreel N, Blanco-Justicia A, Sánchez D, Domingo-Ferrer J (2020) Efficient detection of byzantine attacks in federated learning using last layer biases. In: International Conference on Modeling Decisions for Artificial Intelligence, 154–165. Springer

Jere MS, Farnan T, Koushanfar F (2020) A taxonomy of attacks on federated learning. IEEE Secur Priv 19(2):20–28
Jia B, Zhang X, Liu J, Zhang Y, Huang K, Liang Y (2021) Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in iiot. IEEE Trans Ind Inf 18(6):4049–4058
Jiang Y, Wang S, Valls V, Ko BJ, Lee W-H, Leung KK, Tassiulas L (2022) Model pruning enables efficient federated learning on edge devices. IEEE Trans Neural Networks Learn Syst. https://doi.org/10.1109/TNNLS.2022.3166101
Jiang Y, Li Y, Zhou Y, Zheng X (2020) Mitigating sybil attacks on differential privacy based federated learning. arXiv preprint arXiv:2010.10572
Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, Bonawitz K, Charles Z, Cormode G, Cummings R (2021) Advances and open problems in federated learning. Found Trends® Mach Learn 14(1–2):1–210
Karapanos N, Capkun S (2014) On the effective prevention of man-in-the-middle attacks in web applications. In: 23rd USENIX Security Symposium (USENIX Security 14), 671–686
Kingma DP, Welling M (2019) An introduction to variational autoencoders. Found Trends® Mach Learn 12(4):307–392
Koh PW, Steinhardt J, Liang P (2018) Stronger data poisoning attacks break data sanitization defenses. arXiv preprint arXiv:1811.00741
Koloskova A, Stich S, Jaggi M (2019) Decentralized stochastic optimization and gossip algorithms with compressed communication. In: International Conference on Machine Learning, 3478–3487. PMLR
Lamport L, Shostak R, Pease M (2019) The byzantine generals problem. In: Concurrency: the Works of Leslie Lamport, 203–226
Lee H, Kim J, Ahn S, Hussain R, Cho S, Son J (2021) Digestive neural networks: a novel defense strategy against inference attacks in federated learning. Comput Secur 109:102378
Li Y, Zhou Y, Jolfaei A, Yu D, Xu G, Zheng X (2020) Privacy-preserving federated learning framework based on chained secure multiparty computing. IEEE Internet Things J 8(8):6178–6186
Li T, Sahu AK, Talwalkar A, Smith V (2020) Federated learning: challenges, methods, and future directions. IEEE Signal Proc Mag 37(3):50–60
Li L, Fan Y, Tse M, Lin K-Y (2020) A review of applications in federated learning. Comput Ind Eng 149:106854
Liang X, Liu Y, Chen T, Liu M, Yang Q (2023) Federated transfer reinforcement learning for autonomous driving. Federated and transfer learning. Springer, Berlin, pp 357–371
Li S, Cheng Y, Liu Y, Wang W, Chen T (2019) Abnormal client behavior detection in federated learning. arXiv preprint arXiv:1910.09933
Li S, Cheng Y, Wang W, Liu Y, Chen T (2020) Learning to detect malicious clients for robust federated learning. arXiv preprint arXiv:2002.00211
Li M, Jin Y, Zhu H (2021) Surrogate gradient field for latent space manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6529–6538
Lim WYB, Xiong Z, Miao C, Niyato D, Yang Q, Leung C, Poor HV (2020) Hierarchical incentive mechanism design for federated machine learning in mobile networks. IEEE Internet Things J 7(10):9575–9588
Lin W-T, Chen G, Huang Y (2022) Incentive edge-based federated learning for false data injection attack detection on power grid state estimation: a novel mechanism design approach. Appl Energy 314:118828
Lin J, Du M, Liu J (2019) Free-riders in federated learning: Attacks and defenses. arXiv preprint arXiv:1911.12560
Liu M, Zhao C, Zhang Z, Deng R, Cheng P, Chen J (2021) Converter-based moving target defense against deception attacks in dc microgrids. IEEE Trans Smart Grid 13:3984–3996
Liu K, Dolan-Gavitt B, Garg S (2018) Fine-pruning: Defending against backdooring attacks on deep neural networks. In: International Symposium on Research in Attacks, Intrusions, and Defenses, 273–294. Springer
Long G, Shen T, Tan Y, Gerrard L, Clarke A, Jiang J (2022) Federated learning for privacy-preserving open innovation future on digital health. Humanity Driven AI. Springer, Berlin, pp 113–133
Lu Y, Huang X, Dai Y, Maharjan S, Zhang Y (2020) Federated learning for data privacy preservation in vehicular cyber-physical systems. IEEE Network 34(3):50–56
Luo X, Wu Y, Xiao X, Ooi BC (2021) Feature inference attack on model predictions in vertical federated learning. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE), 181–192. IEEE
Lyu L, Yu H, Ma X, Sun L, Zhao J, Yang Q, Yu PS (2020) Privacy and robustness in federated learning: Attacks and defenses. arXiv preprint arXiv:2012.06337

Lyu L, Yu H, Yang Q (2020) Threats to federated learning: A survey. arXiv preprint arXiv:2003.02133
Ma X, Zhou Y, Wang L, Miao M (2022) Privacy-preserving byzantine-robust federated learning. Comput Stand Interfaces 80:103561
Ma C, Li J, Shi L, Ding M, Wang T, Han Z, Poor HV (2022) When federated learning meets blockchain: a new distributed learning paradigm. IEEE Comput Intel Mag 17(3):26–33
Madi A, Stan O, Mayoue A, Grivet-Sébert A, Gouy-Pailler C, Sirdey R (2021) A secure federated learning framework using homomorphic encryption and verifiable computing. In: 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS), 1–8. IEEE
Melis L, Song C, De Cristofaro E, Shmatikov V (2019) Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP), 691–706. IEEE
Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G (2021) A survey on security and privacy of federated learning. Future Gener Comput Syst 115:619–640
Muñoz-González L, Co KT, Lupu EC (2019) Byzantine-robust federated machine learning through adaptive model averaging. arXiv preprint arXiv:1909.05125
Nasr M, Shokri R, Houmansadr A (2019) 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), 739–753. IEEE
Nguyen TD, Marchal S, Miettinen M, Fereidooni H, Asokan N, Sadeghi A-R (2019) Dïot: A federated self-learning anomaly detection system for iot. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), 756–767. IEEE
Nuding F, Mayer R (2022) Data poisoning in sequential and parallel federated learning. In: Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics, 24–34
Park J, Lim H (2022) Privacy-preserving federated learning using homomorphic encryption. Appl Sci 12(2):734
Park J, Han D-J, Choi M, Moon J (2021) Sageflow: robust federated learning against both stragglers and adversaries. Adv Neural Inf Proc Syst 34:840–851
Park J, Han D-J, Choi M, Moon J (2021) Handling both stragglers and adversaries for robust federated learning. In: ICML 2021 Workshop on Federated Learning for User Privacy and Data Confidentiality. ICML Board
Pillutla K, Kakade SM, Harchaoui Z (2019) Robust aggregation for federated learning. arXiv preprint arXiv:1912.13445
Prakash S, Avestimehr AS (2020) Mitigating byzantine attacks in federated learning. arXiv preprint arXiv:2010.07541
Qu K, Guo F, Liu X, Lin Y, Zou Q (2019) Application of machine learning in microbiology. Front Microbiol 10:827
Reddi S, Charles Z, Zaheer M, Garrett Z, Rush K, Konečnỳ J, Kumar S, McMahan HB (2020) Adaptive federated optimization. arXiv preprint arXiv:2003.00295
Rodríguez-Barroso N, López DJ, Luzón MV, Herrera F, Martínez-Cámara E (2022) Survey on federated learning threats: concepts, taxonomy on attacks and defences, experimental study and challenges. Inf Fusion. https://doi.org/10.48550/arXiv.2201.08135
Sattler F, Müller K-R, Wiegand T, Samek W (2020) On the byzantine robustness of clustered federated learning. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8861–8865. IEEE
Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T (2018) Poison frogs! targeted clean-label poisoning attacks on neural networks. Adv Neural Inf Proc Syst. https://doi.org/10.48550/arXiv.1804.00792
Shang X, Lu Y, Cheung Y-m, Wang H (2022) Fedic: Federated learning on non-iid and long-tailed data via calibrated distillation. arXiv preprint arXiv:2205.00172
Sharma S, Bhatt M, Sharma P (2020) Face recognition system using machine learning algorithm. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES), 1162–1168. IEEE
Shen S, Tople S, Saxena P (2016) Auror: Defending against poisoning attacks in collaborative deep learning systems. In: Proceedings of the 32nd Annual Conference on Computer Security Applications, 508–519
Shoham N, Avidor T, Keren A, Israel N, Benditkis D, Mor-Yosef L, Zeitak I (2019) Overcoming forgetting in federated learning on non-iid data. arXiv preprint arXiv:1910.07796
Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), 3–18. IEEE
Sozinov K, Vlassov V, Girdzijauskas S (2018) Human activity recognition using federated learning. In: 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous

Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom), 1103–1111. IEEE
Stripelis D, Abram M, Ambite JL (2022) Performance weighting for robust federated learning against corrupted sources. arXiv preprint arXiv:2205.01184
Sturluson SP, Trew S, Muñoz-González L, Gram, M, Passerat-Palmbach J, Rueckert D, Alansary A (2021) Fedrad: Federated robust adaptive distillation. arXiv preprint arXiv:2112.01405
Sun J, Li A, DiValentin L, Hassanzadeh A, Chen Y, Li H (2021) Fl-wbc: enhancing robustness against model poisoning attacks in federated learning from a client perspective. Adv Neural Inf Proc Syst 34:12613–12624
Sun Z, Kairouz P, Suresh AT, McMahan HB (2019) Can you really backdoor federated learning? arXiv preprint arXiv:1911.07963
Su L, Xu J (2018) Securing distributed machine learning in high dimensions. arXiv preprint arXiv:1804.10140, 1536–1233
Tan J, Zhang H, Zhang H, Hu H, Lei C, Qin Z (2021) Optimal temporospatial strategy selection approach to moving target defense: a flipit differential game model. Comput Secur 108:102342
Tang J, Xu H, Wang M, Tang T, Peng C, Liao H (2024) A flexible and scalable malicious secure aggregation protocol for federated learning. IEEE Trans Inf Foren Secur. https://doi.org/10.1109/TIFS.2024.3375527
Tolpegin V, Truex S, Gursoy ME, Liu L (2020) Data poisoning attacks against federated learning systems. In: European Symposium on Research in Computer Security, 480–501. Springer, Berlin
Tran B, Li J, Madry A (2018) Spectral signatures in backdoor attacks. Adv Neural Inf Proc Syst. https://doi.org/10.48550/arXiv.1811.00636
Truex S, Liu L, Gursoy ME, Yu L, Wei W (2019) Demystifying membership inference attacks in machine learning as a service. IEEE Trans Serv Comput 14(6):2073–2089
Wahab OA, Mourad A, Otrok H, Taleb T (2021) Federated machine learning: survey, multi-level classification, desirable criteria and future directions in communication and networking systems. IEEE Commun Surv Tutor 23(2):1342–1397
Wang H, Sreenivasan K, Rajput S, Vishwakarma H, Agarwal S, Sohn J-Y, Lee K, Papailiopoulos D (2020) Attack of the tails: yes, you really can backdoor federated learning. Adv Neural Inf Proc Syst 33:16070–16084
Wang D, Li C, Wen S, Nepal S, Xiang Y (2020) Man-in-the-middle attacks against machine learning classifiers via malicious generative models. IEEE Trans Depend Secur Comput 18(5):2074–2087
Wang H, Kaplan Z, Niu D, Li B (2020) Optimizing federated learning on non-iid data with reinforcement learning. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications, 1698–1707. IEEE
Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H (2019) Beyond inferring class representatives: User-level privacy leakage from federated learning. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, 2512–2520. IEEE
Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, Jin S, Quek TQ, Poor HV (2020) Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans Inf Forensics Secur 15:3454–3469
Wen H, Wu Y, Yang C, Duan H, Yu S (2020) A unified federated learning framework for wireless communications: Towards privacy, efficiency, and security. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 653–658. IEEE
Wibawa F, Catak FO, Kuzlu M, Sarp S, Cali U (2022) Homomorphic encryption and federated learning based privacy-preserving cnn training: Covid-19 detection use-case. In: Proceedings of the 2022 European Interdisciplinary Cybersecurity Conference, 85–90
Wong H, Luo T (2020) Man-in-the-middle attacks on mqtt-based iot using bert based adversarial message generation. In: KDD 2020 AIoT Workshop
Wu C, Yang X, Zhu S, Mitra P (2020) Mitigating backdoor attacks in federated learning. arXiv preprint arXiv:2011.01767
Wu N, Zhang N, Wang W, Fan L, Yang Q (2022) Fadman: Federated anomaly detection across multiple attributed networks. arXiv preprint arXiv:2205.14196
Xie C, Huang K, Chen P-Y, Li B (2019) Dba: Distributed backdoor attacks against federated learning. In: International Conference on Learning Representations
Xie C, Koyejo S, Gupta I (2019) Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In: International Conference on Machine Learning, 6893–6901. PMLR
Xu X, Hu H, Liu Y, Tan J, Zhang H, Song H (2022) Moving target defense of routing randomization with deep reinforcement learning against eavesdropping attack. Dig Commun Netw 8:373–387

Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: concept and applications. ACM Trans Intel Syst Technol (TIST) 10(2):1–19
Yin D, Chen Y, Kannan R, Bartlett P (2018) Byzantine-robust distributed learning: Towards optimal statistical rates. In: International Conference on Machine Learning, 5650–5659. PMLR
Yin H, Mallya A, Vahdat A, Alvarez JM, Kautz J, Molchanov P (2021) See through gradients: Image batch recovery via gradinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16337–16346
You X, Liu Z, Yang X, Ding X (2022) Poisoning attack detection using client historical similarity in non-iid environments. In: 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 439–447. IEEE
Yu L, Wu L (2020) Towards byzantine-resilient federated learning via group-wise robust aggregation. Federated learning. Springer, Berlin, pp 81–92
Yuan F-G, Zargar SA, Chen Q, Wang S (2020) Machine learning for structural health monitoring: challenges and opportunities. Sens Smart Struct Technol Civil Mech Aerosp Syst 11379:1137903
Yuan X, Ma X, Zhang L, Fang Y, Wu D (2021) Beyond class-level privacy leakage: breaking record-level privacy in federated learning. IEEE Internet Things J 9(4):2555–2565
Zhai K, Ren Q, Wang J, Yan C (2021) Byzantine-robust federated learning via credibility assessment on non-iid data. arXiv preprint arXiv:2109.02396
Zhang L, Xu J, Vijayakumar P, Sharma PK, Ghosh U (2022) Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system. IEEE Trans Netw Sci Eng. https://doi.org/10.1109/TNSE.2022.3185327
Zhang C, Li S, Xia J, Wang W, Yan F, Liu Y (2020) {BatchCrypt}: Efficient homomorphic encryption for { Cross-Silo} federated learning. In: 2020 USENIX Annual Technical Conference (USENIX ATC 20), 493–506
Zhao Y, Zhao J, Yang M, Wang T, Wang N, Lyu L, Niyato D, Lam K-Y (2020) Local differential privacy-based federated learning for internet of things. IEEE Internet Things J 8(11):8836–8853
Zhao L, Li J, Li Q, Li F (2021) A federated learning framework for detecting false data injection attacks in solar farms. IEEE Trans Power Electr 37(3):2496–2501
Zhao L, Jiang J, Feng B, Wang Q, Shen C, Li Q (2021) Sear: secure and efficient aggregation for byzantine-robust federated learning. IEEE Trans Depend Secure Comput 19(5):3329–3342
Zhao B, Mopuri KR, Bilen H (2020) idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610
Zhou S, Huang H, Chen W, Zhou P, Zheng Z, Guo S (2020) Pirate: a blockchain-based secure framework of distributed machine learning in 5g networks. IEEE Network 34(6):84–91
Zhou X, Xu M, Wu Y, Zheng N (2021) Deep model poisoning attack on federated learning. Future Internet 13(3):73
Zhou Z, Xu C, Wang M, Ma T, Yu S (2021) Augmented dual-shuffle-based moving target defense to ensure cia-triad in federated learning. In: 2021 IEEE Global Communications Conference (GLOBECOM), 01–06. IEEE
Zhu MH, Ezzine LN, Liu D, Bengio Y (2022) Fedilc: Weighted geometric mean and invariant gradient covariance for federated learning on non-iid data. arXiv preprint arXiv:2205.09305
Zhu L, Liu Z, Han S (2019) Deep leakage from gradients. Adv Neural Inf Proc Syst. https://doi.org/10.48550/arXiv.1906.08935
Zhu H, Ling Q (2021) Broadcast: Reducing both stochastic and compression noise to robustify communication-efficient federated learning. arXiv preprint arXiv:2104.06685
Zizzo G, Rawat A, Sinn M, Buesser B (2020) Fat: Federated adversarial training. arXiv preprint arXiv:2012.01791