Artificial Intelligence Review (2024) 57:130
https://doi.org/10.1007/s10462-024-10774-7

Survey: federated learning data security and privacy-preserving in edge-Internet of Things

Haiao Li[1,2] · **Lina Ge**[1,2,3] · **Lei Tian**[1,2]

Accepted: 24 April 2024 / Published online: 29 April 2024

Abstract
The amount of data generated owing to the rapid development of the Smart Internet of Things is increasing exponentially. Traditional machine learning can no longer meet the requirements for training complex models with large amounts of data. Federated learning, as a new paradigm for training statistical models in distributed edge networks, alleviates integration and training problems in the context of massive and heterogeneous data and security protection for private data. Edge computing processes data at the edge layers of data sources to ensure low-data-delay processing; it provides high-bandwidth communication and a stable network environment, and relieves the pressure of processing massive data using a single node in the cloud center. A combination of edge computing and federated learning can further optimize computing, communication, and data security for the edge-Internet of Things. This review investigated the development status of federated learning and expounded on its basic principles. Then, in view of the security attacks and privacy leakage problems of federated learning in the edge Internet of things, relevant work was investigated from cryptographic technologies (such as secure multi-party computation, homomorphic encryption and secret sharing), perturbation schemes (such as differential privacy), adversarial training and other privacy security protection measures. Finally, challenges and future research directions for the integration of edge computing and federated learning are discussed.

Keywords Federated learning · Edge computing · Data security · Privacy-preserving · Internet of Things

✉ Lina Ge
66436539@qq.com

[1] School of Artificial Intelligence, Guangxi Minzu University, Nanning 530006, China

[2] Key Laboratory of Network Communication Engineering, Guangxi Minzu University, Nanning 530006, China

[3] Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Nanning 530006, China

1 Introduction

As a new Internet technology, the Smart Internet of Things (SIoT) has ushered in a new wave of development in correspondence with the rise in artificial intelligence technologies, which can greatly help the development of intelligence in various fields (Shen et al. 2023). IoT devices have increasingly advanced sensing technologies, storage capabilities, and computing and processing capabilities. They are widely deployed for various sensing tasks, such as in wearable medical monitoring devices to assist in diagnosis (Jan et al. 2023), smart agricultural monitors (Garg and Alam 2023), and smart industrial control (Adhikari et al. 2023). Successful practices in these fields must consider the foundations of big data, large-scale data analyses, and the training and application of machine learning (ML) models. However, traditional centralized ML processes are facing ever-growing, multi-source, heterogeneous, and complex distributed IoT data. These data cannot be managed by relying only on the limited functional services provided by cloud service centers (Ge et al. 2023; Shuvo et al. 2022). The cloud-centric computing model uploads data collected by edge-IoT devices to a central cloud server for centralized storage, processing, and computing. Nevertheless, the centralized processing model based on the cloud center cannot withstand the explosive growth in data volume. Moreover, data related to user privacy are uploaded to the cloud center to address security issues, increasing privacy leakages. Edge computing (EC) is a new distributed computing paradigm (Shi et al. 2016) for storing, processing, and applying data collected by edge-IoT devices on the edge side close to the data source. The EC approach effectively reduces the pressure of performing data processing in the cloud center. The bottleneck in the transmission communication retains the training data in edge device nodes to avoid the risk of privacy data leakages to a certain extent (Li et al. 2022a; Ranaweera et al. 2021). EC approaches combining ML (Hua et al. 2023; Ning et al. 2023) and deep learning (DL) (Ahmad et al. 2023; Zhang et al. 2023a; Ghosh and Grolinger 2020) have been applied in many fields.

Federated learning (FL) is a paradigm for training statistical learning models on distributed edge networks, and was proposed by Google in 2016 (Mcmahan et al. 2017). It is a mainstream solution for solving the problems concerning huge communication overheads, data privacy security, and heterogeneous data fusion (Fan et al. 2023; Xu et al. 2023). As a distributed ML (DML) method, FL realizes global model training without local source data. Through FL, each local device only needs to encrypt the local model parameter updates and then upload them to a central aggregation server. The central aggregation server then uses a federated average algorithm to obtain a global model parameter update. Next, each local device can download the global model parameters and decrypt them. Finally, each local device uses the global model parameters for the local ML model training. This process cannot be stopped until the maximum number of iterations or required model accuracy is reached. FL has been successfully applied in medical (Fan et al. 2022; Li et al. 2023a; Myrzashova et al. 2023; Sharma and Guleria 2023), industrial (Zheng et al. 2023; Guo et al. 2023), agricultural (Durrant et al. 2022; Friha et al. 2021) and other fields. It enables models to be trained on edge devices, which also means that EC is an appropriate environment for using FL. Therefore, problems such as high communication costs, privacy and data security needs, and heterogeneous data isolation in edge-Internet of Things (edge-IoT) networks can be alleviated by utilizing FL technology in the EC environment. Moreover, FL solves the privacy and security problem of sensitive data in ML in the EC environment. However, the processes of uploading and downloading the model update parameters, training iterations, and other processes still expose the FL environment to a

series of risks. For example, malicious hacker attacks and dishonest participants may use the model parameters to infer the original data (Phong et al. 2017). The data security and privacy-preserving schemes in FL in an EC environment face the following threats and challenges.

- Data security threats. In EC environment, communications between devices and between devices and aggregation servers may be subject to various network attacks, and edge devices are usually distributed in untrusted environments and may be subject to physical attacks or malicious tampering, etc. FL systems in an EC environment need to send data to the edge layer close to the data source for processing, which still involves the risk of data security. These systems may also be subject to various forms of attacks, such as poisoning attack. A malicious client in the FL system may send incorrect model updates to the parameter aggregation server to probe the remaining participating side datasets or destroy the model accuracy. In the edge-IoT network environment, participants can access the FL system any time, verify their identity credibility, and grant access to the system. Building a trusted FL system in an untrusted environment remains a key issue in the design of FL security systems.
- Data privacy-preserving issues. Edge computing drives services closer to end-users, and FL transfers models from edge service nodes to the user's local for training, mitigating the risk of privacy leakage caused by user data leaving the local area. However, information such as model training parameters or gradients communicated between local devices and edge service nodes may still leak information about the local raw dataset (Ge et al. 2023). Meanwhile, since service providers are usually honest but curious and have the most private information at their disposal, even partial model updates may lead to serious privacy data leakage problems. Therefore, the constant updating of attack types also puts new requirements on the research of privacy-preserving FL techniques under edge computing.
- Communication and computation overhead. The rapid popularization of SIoT applications has led to an explosive growth in the amount of data at edge nodes, and the high computational cost brought by massive data training and the high communication cost generated by the exchange of large-scale model parameters have seriously hindered the widespread application of FL techniques. Meanwhile, the low computational performance, limited communication bandwidth, network real-time and high quality of service requirements of edge devices also pose challenges to the research of novel FL techniques in EC environments.
- In EC, FL enables large-scale device collaboration for training AI models in a privacy-preserving manner. However, the scale of edge devices involved in FL is huge, and the performance and computational power of hardware devices may vary greatly. Meanwhile, the geographic distribution of edge devices varies widely, which may lead to heterogeneity problems such as network latency, instability, and bandwidth limitations. Finally, each edge device as a FL participant usually has non-independent and homogeneously distributed datasets, and heterogeneous data sources bring huge negative impact on the accuracy of FL models. Device heterogeneity in IoT EC environments poses significant challenges for FL implementations.

Several current research works are devoted to the innovation and application of FL theory in EC. Edge terminal devices with high mobility and exposed to the open edge network environment are prone to various malicious attacks, which not only affect the performance of the FL model, but also bring serious data security and privacy leakage problems to the

FL process. To address the data security problem of the FL process under EC, (Huang et al. 2023a) proposed a reliable FL mechanism for mobile EC, designing endpoint selection algorithms based on the reputation mechanism for the construction of the reputation model and the concealment of the selected endpoints, and maintaining the model performance through elite campaigning to reduce the impact of poisoning attacks on the model. (Ni et al. 2023) proposed a new Byzantine robust FL framework, which identifies and discards malicious gradients through a dual filtering mechanism design, and uses an adaptive weight adjustment scheme to dynamically reduce the aggregation weights of potentially malicious gradients, to realize secure and trustworthy FL in IoT. (Li et al. 2023b) address the security attacks that FL is vulnerable to in distributed adversarial environments, and non-independent and homogeneous distribution of data further weakening the robustness of the existing FL methods. The Mini-FL scheme was proposed. This scheme performs unsupervised learning on the received gradients to define a grouping policy, and the aggregation server groups the received gradients according to the grouping policy and calculates the weighted average of the gradients in each group to update the global model. Existing FL data security technology schemes mainly involve endpoint selection, hardware device secure communication, model secure aggregation, and security detection, etc., and the related technology development and innovation are still ongoing.

To address the privacy protection of FL process under EC, (Zhu et al. 2023) proposed an enhanced FL model with reinforcement learning, and designed a partially encrypted secure multiparty broadcast computation algorithm by combining the advantages of end-to-end homomorphic encryption and secure multiparty computation, which realizes that the edge device and the roadside unit collaborate to train the learning model without exposing the original data. (Li et al. 2023c) proposed a ubiquitous intelligent FL privacy protection scheme, designing matrix masks to ensure secure transmission between embedded devices and edge servers, while using differential privacy mechanisms to train residual models on edge servers to provide privacy protection for data under EC. (Liu et al. 2023a) discussed the problem that most FL privacy protection protocols only provide single round privacy guarantees, and proposed a long-term privacy protection aggregation protocol, which uses a batch partition deletion update policy and integrates with advanced privacy-preserving aggregation protocols to satisfy single- or multi-round privacy guarantees. Privacy protection is an important issue in the combination of FL and EC, and the existing technical solutions mainly include encryption techniques, differential privacy techniques, etc. The FL process in EC environment involves information exchange and model training among multiple devices or edge nodes, and requires comprehensive consideration of various factors such as algorithms, protocols and policies.

Distinguished from other published works focusing on FL within EC, this paper distinctly delineates the landscape of security issues and privacy threats intrinsic to FL within edge networks. Additionally, it meticulously categorizes and expounds upon the adverse consequences stemming from diverse privacy security attacks, comprehensively dissecting their impacts on the FL process. Compared to the literature (Huang et al. 2023a; Ni et al. 2023; Li et al. 2023b, 2023c; Zhu et al. 2023; Liu et al. 2023a) that explores the data security and privacy protection issues involved in FL under EC. This review systematically sorts out the current research results of privacy-preserving FL in edge networks, and focuses on the problems of multi-party conspiracy to steal private data and malicious adversaries destroying the FL process in FL and the corresponding security defense schemes.

This article aims to provide a comprehensive discussion of the development of safe and reliable FL systems in an EC environment. First, we introduce concepts related to EC and

FL. We then summarize the data security risks and privacy leakage threats in the current edge-IoT environment. Next, we review the research progress on FL's existing privacy data security protection technologies in EC. Finally, we analyse future research hotspots in FL privacy security protection in edge-IoT, and provide several research suggestions for establishing a secure and trusted FL system under edge-IoT.

2 Fundamentals of federated learning

This section introduces the concepts and working principles of EC and FL. We also summarize the data security and privacy leakage attacks faced by FL systems in current edge-IoT environments.

2.1 Related technologies

2.1.1 Edge computing

At present, ML training data mainly comes from edge-IoT devices, which not only have a large amount of data, but also a high degree of data heterogeneity. With increasing attention being paid to the security protection of private data, the traditional architecture model of centralized processing in a cloud center can no longer meet the needs of the current technological developments. As a new type of distributed computing technology, the core of the EC technology involves loading raw data into edge network devices (such as edge servers) for storage, processing, and computing. In the edge-IoT, the breakthroughs in EC technology have meant that many ML model trainings can be implemented locally without having to be delivered to a cloud center. The EC processes and analyses data in real time near the data source, providing the advantages of high data processing efficiency, strong real-time performance, and low network latency. This mode is closer to the user and solves requirements at the edge node. It effectively improves the computing processing efficiency, reduces channel pressure, and protects the security of private data.

2.1.2 Deep learning

DL is currently used in a wide range of applications, including computer vision and natural language processing. Terminal devices such as smartphones and IoT sensors generate data to be analysed in real time using DL and to train DL models. However, DL inferences and training require significant computing resources for quick execution. EC's fine grid of computing nodes placed close to terminal devices is a viable way to meet the high computing and low latency requirements of edge device DL and also provides privacy technology protection, bandwidth efficiency, and scalability. However, there is a risk that the sensitive information of data owners in the EC environment will be leaked when the data leaves the edge server node during local upload. Service providers are usually honest, but may be curious. FL was initially proposed to provide a collaborative data-training solution. It provides considerable privacy enhancements by coordinating multiple client devices to train a shared ML model without directly exposing the underlying data.

2.1.3 Definition of FL

FL problems involve learning a single global statistical model from data stored in a large number of remote devices. An example of a traditional FL architecture is shown in Fig. 1. The goal of FL is to learn the model under the constraints of local storage and the processing of the data generated by the device, and to periodically update the model parameters to be communicated with the cloud parameter server. In other words, the goal is to minimize the following objective function, i.e., to minimize the average training loss for all customers, as follows:

$$\min_{w} F(w), F(w) = \sum_{k=1}^{m} p_k F_k(w) \tag{1}$$

where, m is the total number of devices participating in training, p_k specifies the relative weight of influence attributed to each individual device, and $F_k(w)$ is the local objective function of the k th device. $F_k(w)$ is usually defined as the empirical risk of the local data, as follows:

$$F_k(w) = \frac{1}{n_k} \sum_{i=1}^{n_k} f_i(w, x_i, y_i) \tag{2}$$

where, n_k is the data volume of the k th device, $f_i(w, x_i, y_i)$ is the loss function of the model with the parameter w on the instance (x_i, y_i) in the k th device-local dataset. The optimization process within FL focuses on the minimization of the value associated with the local loss function.

The FL architecture can also be designed using peer-to-peer networking, as shown in Fig. 2. This architecture eliminates the hidden dangers caused by a single failure point, further ensuring system security. It is easy to scale but may consume a greater amount of computing resources in the encryption and decryption of message communication.

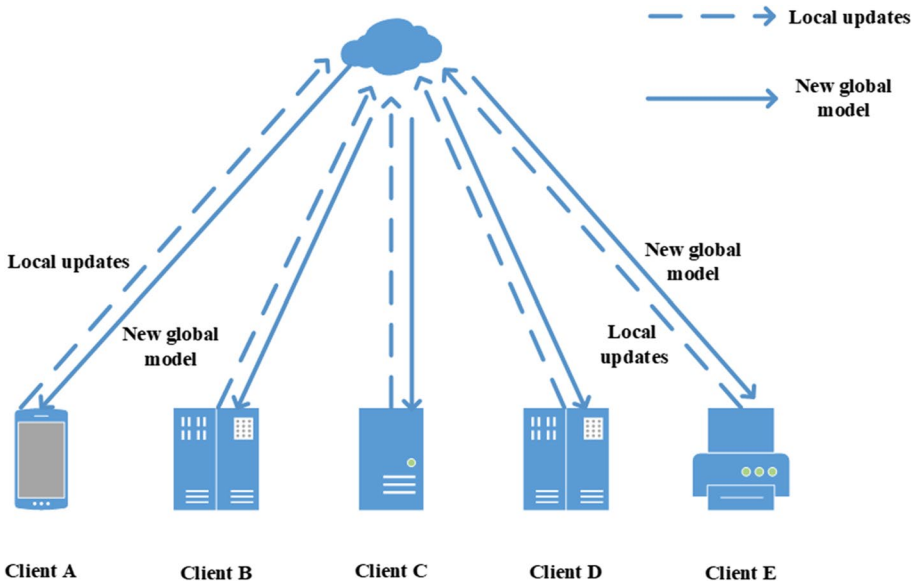

Fig. 1 Federated learning system architecture: client–server

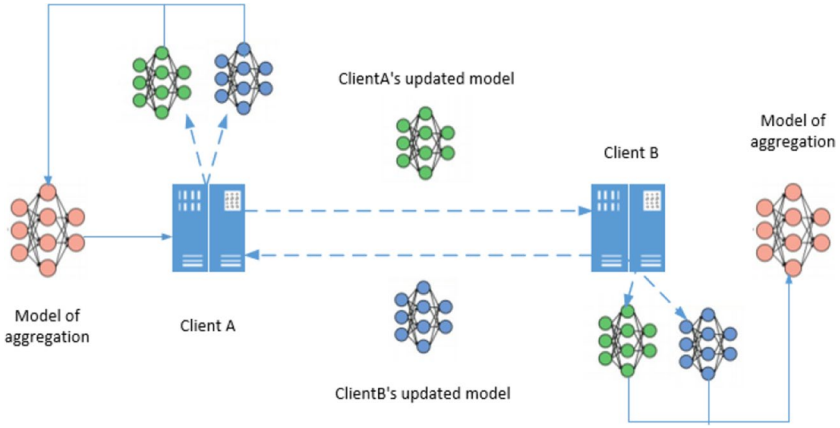

Fig. 2 Federated learning system architecture: Peer-to-peer network architecture

2.2 FL data security and privacy leakage

In edge-IoT, a large number of edge-IoT devices are often connected to the Internet, which undoubtedly significantly increases the security risks. Currently, most of the various security attack methods are conducted through the Internet, and the success rate is generally high. In this environment, there are a variety of data security and privacy leakage threats in FL systems. This paper mainly discusses the data security attacks and privacy leakage attacks involved in FL in edge-IoT.

2.2.1 FL data security attacks

FL under EC utilizes the computational power of edge devices to collaboratively train models. However, insecure network communication environments and design flaws of traditional FL structures lead to many data security issues facing the FL process. At the same time, current FL security techniques are difficult to defend against constantly innovative and sophisticated security attacks. Malicious adversaries often use vulnerable edge-IoT devices to intrude into internal targets, destroying the global model convergence as well as the model performance, thus posing security threats to the entire FL process. In the following, the more common data security attacks are described and categorized in detail.

Poisoning attack: In the FL system, a poisoning attack (Bhagoji et al. 2019; Chen et al. 2023) aims to use poisoned data to disrupt the model training and reduce the federated model accuracy. Poisoning attacks can be divided into model and data poisoning. Model poisoning attacks are injected into the model trained by the edge nodes. Data poisoning can disrupt the model training by injecting poisoned data into a local dataset of an edge device. Both methods will produce malicious updates to the model training of FL systems, and it is difficult to detect such attackers (Fang et al. 2020; Guo et al. 2022). In the edge-IoT environment, the large and complex number of networked devices participating in the FL system and degrees of device trustworthiness are unknown, posing a potential threat to the security of the FL system. Therefore, it is necessary to develop defensive measures to protect FL systems from poisoning (Rodríguez-Barroso et al. 2022a).

Sybil attacks: In FL-distributed networks, attackers can use a single node to forge multiple identities. Attackers use these forged identities to control or affect other normal nodes in a distributed network, resulting in network robustness losses (Douceur 2002; Singh and (2006) Eclipse attacks on overlay networks: Threats and defenses. In: Proceedings IEEE INFOCOM 2006;). In the context of edge intelligent networks, distributed collaborative intelligent applications need to exchange a large amount of data and information to ensure the efficient operation of different programs, and the distributed network environment provides conditions for Sybil attacks, where attackers can create multiple identities to interfere with the normal operation of the system or steal services for profit (Hammi et al. 2022). Traditional FL allows distributed clients to join and exit the system, and attackers can use multiple colluding aliases to join the system to execute attacks, for this reason, some researchers have set up a trusted third party for centralized FL systems to verify false node identities or use blockchain technology to implement decentralized and reliable FL in untrustworthy networks, and utilize the structural properties of the blockchain and cryptography to defend against Sybil attacks (Xiao et al. 2022; Fang et al. 2022).

Backdoor attack: Backdoor attack is when an attacker intentionally inserts a malicious message or "backdoor" into a system in order to trigger malicious behavior under certain conditions. Backdoor attack may cause a system to perform well under normal conditions, but perform illegal and improper tasks when certain conditions are met. Since FL involves multiple participants, an attacker can attempt to insert a backdoor into the local model of some trusted participants and then disrupt the FL process by tampering with gradient updates, sending malicious update parameters. Meanwhile, after multiple iterations of FL, the model parameter aggregation may change and the embedded backdoor may gradually fail. Therefore, malicious actors often increase the persistence of backdoor attacks in FL by slowing down the learning rate during training (Nguyen et al. 2024; Gong et al. 2022; Yang et al. 2023a).

Byzantine attack: As the most typical attack method in FL, the attacker tries to tamper with the model update parameters submitted by trusted nodes so that the actual model aggregation deviates far from the model convergence direction, resulting in a decrease in the FL model accuracy and serious deviations in the predicted values. This type of security attack is the most common and is highly effective. Correspondingly, the FL security systems for Byzantine attacks are also being updated. The use of secure robust aggregation is currently recognized as an effective means for defending against Byzantine attacks (Li et al. 2023d; Miao et al. 2022; Wan et al. 2022).

Free-riding Attack: In FL, free-riding attack is the process of a free-rider generating false model update parameters to report to the parameter server. Then, the free-rider uses the global model parameters to update its local model but does not contribute its own local data to the global model aggregation. Free-riding attacks reduce the amount of data involved in global model training. Spurious parameter updates can also affect the global model accuracy (Lin et al. 2019; Fraboni et al. 2021).

Adversarial attack: Adversarial attack in FL deceive federated models by adding subtle perturbations to the local raw data to generate adversarial samples. Adversarial attacks can be divided into white-box, black-box, targeted, and non-directed attacks. In a white-box attack, the adversary masters the model and training data information; however, this is generally inconsistent with actual situations. Under a black-box attack, the adversary has little knowledge of the model and training set information. This is in line with the actual scenarios, and is currently the main research direction. Under a targeted attack, a multi-classification ML model classifies and outputs input samples into specified categories. Under a non-directional attack, the adversary uses generative adversarial samples to deceive the

FL model. Adversarial attacks help adversaries evade FL system security detection and can generate poisoned samples to undermine the FL system accuracy (Goodfellow et al. 2014; Nair et al. 2023).

2.2.2 FL privacy leakage attacks

In terms of privacy leakage, during the entire training cycle of the FL model, information such as weight updates and gradient updates may be leaked as sensitive private data. During this period, there may be a risk of privacy leakage, whether from the participating clients of the FL, central servers, or third-party servers. Recent studies have found that even some model gradient information can be leaked as local private data samples or features. Malicious attackers can also steal the training datasets of a local client in an FL iteration, or rebuild the training datasets through an FL global model inversion.

Model inversion attack: This type of attack method obtains the training set information from a target model that has completed the training. A model inversion attack infers the training set information through a reverse analysis. This information can be the information of members participating in the training and/or certain statistical characteristics of the training set data. For example, an attack method that uses a model inversion attack to infer actor identity information is called a member inference attack. Member inference attacks are designed to obtain information by checking for the presence of raw data in the training set (Hu et al. 2023). In an FL member inference attack, with each iteration, the data contributor uploads its own model update parameters to the parameter server. At this time, the server understands the model characteristics and global model parameters of each party. Thus, it can easily be determined whether a specific data sample is in the local training dataset. In some cases, parameter servers can become malicious adversaries and conduct member inference attacks. The remaining participants in the FL system may be potential adversaries. They can determine whether the estimated data originates from the target model training set. Subsequently, the adversaries judge whether the data are member data. During the entire attack process, the FL model parameter contributors don't know the opponent's inference attack behavior, leading to the privacy data of the model contributors being leaked unknowingly (Hatamizadeh et al. 2023).

Refactoring attack: This type of attack focuses on reconstructing all of the training set information of the model contributors or certain sensitive feature categories of the training set (e.g., attribute category labels). Several reconstruction attack methods have been proposed, such as generalized adversarial network (GAN) attacks. (Hitaj et al. 2017) first proposed a GAN-based data reconstruction attack to steal the private data of model contributors. When using the GAN to attack the FL, an adversary generates a prototype of the training data of the target by training the GAN. By injecting fake training data samples into the model server, the model contributor is tricked into contributing more local training data. Eventually, the adversary may use this information to restore the local original data of the model contributor and thereby steal sample data. Through GAN attacks, adversaries can also complete data category inferences (Liu et al. 2023b) and label inferences (Jin et al. 2023). Because the FL system server does not know much regarding the reputation and honesty of the various parties, it is difficult to distinguish a GAN-based data-inference attack. Therefore, when building and maintaining FL security systems, it is necessary to strengthen the identification of and defend to such reconstruction attack methods.

Model extraction attack: In 2016, (Tramèr et al. 2016) proposed a model extraction attack method focused on reconstructing alternative models similar to the target model. In

the FL context, the adversary obtains black-box access to the target model. It obtains the return result by sending data in a loop and uses these return values to steal the FL model update parameters or model functions. It restores the FL target model or reconstructs a similar target FL model as much as possible (Li et al. 2023e).

The above data security and privacy leakage attacks are common attack types in FL. Table 1 and 2 compare and analyse the methods and effects of the security and privacy attack methods.

2.3 FL privacy security threats in EC

The traditional cloud-centric computing approaches are gradually proving inadequate in addressing the security challenges posed by the vast amounts of privacy-sensitive data generated by edge terminal devices in the intelligent Internet of Things landscape. EC technology, by processing and analyzing data near its source in real-time, not only caters to the demand at the edge nodes but also significantly enhances processing efficiency, alleviates communication overhead, and safeguards data privacy. Concurrently, FL enables model training on edge devices, making EC an apt environment for FL deployment. Nevertheless, EC, being a nascent distributed computing paradigm, harbors distinctive privacy security risks. The FL model training process within EC confronts similar privacy security threats. The intricate service model of edge computing, coupled with real-time application requisites and the resource constraints of edge terminal devices, alongside the heterogeneous nature of edge user privacy data from multiple sources, may exacerbate privacy security concerns during the FL training process. For instance, the exigency for refined security authentication mechanisms at the edge nodes, coupled with the inadequacy of traditional encryption technologies for edge computing environments, poses risks of privacy security breaches, particularly in untrusted execution environments. The primary data security threats and privacy protection challenges encountered by FL within EC encompass:

- Secure sharing and storage of privacy data: The collection of user data by edge terminal devices encompasses sensitive information such as personal location, health data, and identity details. Storing such privacy data in third-party servers, like edge servers, within edge computing environments raises concerns regarding data leakage and unauthorized tampering. Moreover, the presence of numerous unknown trust nodes within the intricate edge network environment poses additional security risks. The connection of FL model learning and training tasks to the network exposes vulnerabilities that malicious adversaries can exploit through various security attack methods. Common network security threats in the EC environment include denial of service attacks, information injection, malicious code attacks, gateway forgery, and man-in-the-middle attacks. The fundamental challenge persists in securely uploading local data to the network or entrusting it to third-party servers for storage and processing.
- Fine-grained authentication access control: EC, being a distributed computing system, operates across multiple trust domains. However, even authorized nodes within untrusted network environments face trust issues. Establishing authentication identities for network nodes across diverse trust domains and ensuring secure access verification between them imposes elevated requirements and challenges on the design of fine-grained access control mechanisms within complex edge networks.
- Design requirements of lightweight privacy security algorithms: The majority of edge terminal devices in edge computing environments suffer from limited resource perfor-

Table 1 Federated learning (FL) data security attacks

Type of Attacks	Attack Methods	Attack Content	Attack Effect	Papers
Data Security Attacks	Poisoning Attack	Poisoning attacks are divided into data poisoning attacks and model-poisoning attacks, in which poisoned data is injected into the local training data of the participants or directly destroys local model parameter information	Corrupt model training with poisoning data to reduce federated model accuracy	Liu et al. 2023a; Bhagoji et al. 2019; Chen et al. 2023; Fang et al. 2020; Guo et al. 2022; Rodríguez-Barroso et al. 2022a)
	Sibyl Attack	Attackers use a single node to forge multiple identities to control or influence other normal nodes in the distributed network	Sibyl attack leads to the loss of network robustness and reduces the accuracy of the FL global model	Douceur 2002; Singh and (2006) Eclipse attacks on overlay networks: Threats and defenses. In: Proceedings IEEE INFOCOM 2006; Hammi et al. 2022; Xiao et al. 2022; Fang et al. 2022)
	Backdoor Attack	Add samples injected with special triggers into the training data to train a model embedded in the backdoor	Backdoor attack affects the prediction accuracy of FL model	Nguyen et al. 2024; Gong et al. 2022; Yang et al. 2023a)
	Byzantine Attack	The attacker attempts to tamper with the model update parameters submitted by trusted nodes, so that the actual model aggregation is far away from the model convergence direction	The accuracy of the FL model decreases and the predicted values are seriously biased	Li et al. 2023d; Miao et al. 2022; Wan et al. 2022)
	Free-riding Attack	The free-rider generates fake model update parameters and reports them to the central server to get the global model without actually contributing	Free-riding attack reduces the amount of data involved in model training, and spurious parameter updates also affect model accuracy	Lin et al. 2019; Fraboni et al. 2021)
	Adversarial Attack	Deceive the federated model by adding subtle perturbations to the local raw data to generate adversarial samples	Adversarial attack helps adversaries evade FL system security detection, and generate poisoning samples to destroy FL system accuracy	Goodfellow et al. 2014; Nair et al. 2023)

Table 2 Federated learning (FL) privacy leakage attacks

Type of Attacks	Attack Methods	Attack Content	Attack Effect	Papers
Privacy Leakage Attacks	Model Inversion Attack	Obtain the training set information from the target model that completes the training, and infer the training set information through reverse analysis	It is possible to infer whether a particular sample or a particular attribute is in its corresponding training set	Hu et al. 2023; Hatamizadeh et al. 2023)
	Refactoring Attack	Inferred labels for training samples and reconstructed training samples	Reconstruct all training set information of model contributors or certain sensitive feature categories of training set	Hitaj et al. 2017; Liu et al. 2023b; Jin et al. 2023)
	Model Extraction Attack	Attackers send data in a loop to obtain the return result, and use these return values to steal the FL model update parameters or model functions, and restore the FL target model as much as possible	Rebuild an alternative model that is similar to the target model	Tramèr et al. 2016; Li et al. 2023e)

mance, including constrained computing power and battery capacity, especially prevalent in mobile terminal devices. This limitation inhibits the effective implementation of traditional security encryption algorithms, access control mechanisms, and security defense measures on resource-constrained terminal devices. Consequently, the development of lightweight privacy and security algorithms tailored for the EC environment becomes imperative for the secure and efficient execution of FL processes.

- Data consistency and quality issues: Data in EC environments is typically distributed across edge devices, leading to synchronization challenges due to performance disparities among terminal devices. Asynchronous execution of learning algorithms and computing tasks further complicates ensuring data consistency and the accuracy of FL local model training parameters. Addressing these issues remains a formidable challenge within FL in EC.

3 Related works

This section discusses FL studies on the development of data security and privacy protection technologies in the edge-IoT environment, and analyses, compares, and summarizes the mainstream solutions in the industry.

3.1 FL data security

FL is a variant of distributed learning that enables the training of shared models without the need to access private data from different sources. Despite its benefits in terms of privacy protection, the distributed nature of FL and its privacy constraints make it vulnerable to data security attacks. These include poisoning, sybil, backdoor, and adversarial attacks. In the edge-IoT scenario, cyber-physical systems integrate sensing, computing, control, and network processes into physical objects and infrastructure elements connected via the Internet to perform common tasks. Once the learning and training of the FL model are connected to the network, hackers can use a series of security attack methods and the security mechanism vulnerabilities of the host to launch security attacks. Therefore, the tight coupling of the network and physical systems poses challenges to the stability, security, efficiency, and reliability of FL. In this section, we summarize the means and mechanisms for providing robust performance protection of FL models, including data security attack intrusion detection mechanisms and improvements in FL model security robustness.

3.1.1 Federated robust aggregation algorithms

With the steady developments in edge device computing power, storage capacity, and other performances, information transmission, local storage, and network computing tasks have gradually shifted to edge devices. In the face of access to complex edge devices, this undoubtedly has brought significant challenges to the security of FL. FL aggregation algorithms play an important role in updating the global model, such as privacy data protection, efficient model convergence, and security attack defense. Different FL aggregation algorithms are designed with different advantages and disadvantages. Secure aggregation, as an important criterion in the design of FL aggregation algorithms, aims to protect the security and reliability of the participants' local models and the FL training process. However, in the untrustworthy EC environment, the current federated aggregation algorithms cannot

well defend against the Byzantine attacks that are common in distributed computing systems. Meanwhile, the model communication of massive edge participants also puts higher requirements on efficient FL aggregation algorithms. For this reason, the development of secure and efficient FL aggregation algorithms is an important means to implement robust FL processes in untrustworthy EC environments. This section focuses on a systematic and in-depth analysis of the more advanced FL aggregation algorithms, and discusses and summarizes the model aggregation algorithms under different FL and EC application scenarios by comparing different advanced schemes in terms of data security, model performance, and system efficiency (Qi et al. 2023).

In terms of defending data security attacks, Nuria et al. (Rodríguez-Barroso et al. 2022b) deeply analyzed the model poisoning backdoor attack in FL, discussed the pattern key backdoor attack and distributed backdoor attack that may be triggered by adversarial participants with outlier behaviors in FL, and developed a new robust and resilient FL aggregation operator, i.e., robust filtering of one-dimensional outliers, in response to the above problems, which filters out the univariate outliers by performing the standard deviation method on model participant updates for each dimension to identify univariate outliers and thus filter out adversarial participants in FL. (Zhang et al. 2023b) supported backdoor detection for FL secure aggregation through two new primitives, inadvertent random grouping and partial parameter disclosure. Inadvertent random grouping divides FL participants into one-time random subgroups, which prevents collusive attackers from knowing each other's group membership assignments and detect backdoor attacks using statistical distributions of subgroup aggregation parameters based on learning iterations. Compared to the robust filtering operator for one-dimensional outliers proposed by Nuria. Zhang's design scheme has better performance in terms of communication and computational cost, but Nuria focuses more on the detection of outlier behaviors of adversarial participants, with a focus on defending against model-poisoning attacks based on data poisoning and model updating enhancements of adversarial participants.

Nevertheless, current FL systems are unable to monitor the local training process of edge devices in real-time when dealing with distributed collaborative learning among a large number of IoT devices, leading malicious attackers to exploit the vulnerability for byzantine attacks. (Ni et al. 2023) proposed a dual filtering mechanism for byzantine attacks during FL under edge-IoT to identify and discard malicious gradients and increase the security of the FL training process, while considering the impact of potential malicious gradients, designing an adaptive weight adjustment scheme to correct the size of local and normal gradients to keep the same size using a dynamic trimming method to ensure effective model aggregation. However, Ni's scheme does not consider that the real datasets usually used by FL participants in real EC environments are non-independently and identically distributed, and the non-independently and identically distributed original datasets further weaken the robust performance of existing FL aggregation algorithms and increase the possibility of the FL global model to be attacked and corrupted in non-independently and identically distributed scenarios.

For this reason, (Li et al. 2023b) evaluated the effectiveness of existing byzantine robust FL methods in non-independently and identically distributed scenarios, and proposed a mini-FL scheme, which proposes a grouping aggregation method based on participant geographic, temporal, and user characteristics as a grouping principle. The scheme introduces a clustering method, considering that the uploaded gradients naturally tend to cluster due to location, time and user clustering. The parameter server divides the received gradients into different subgroups and performs byzantine robust aggregation separately, the similar behavior of each subgroup results in a smaller range of gradients leading to a smaller attack

space. The introduction of clustering method reduces the attack surface and effectively enhances the FL robustness in practical non-independent same-distribution scenarios (He et al. 2023) similarly analyzed the challenges posed by non-independent same-distribution data to the byzantine robustness of FL, and proposed the byzantine robust stochastic model aggregation method, which utilizes robust stochastic model aggregation to obtain the byzantine robustness to non-independent same-distribution data, and analyzed and proved that the byzantine convergence of the robust stochastic model aggregation scheme in distributed nonconvex learning. Compared with the stronger generality of the scheme proposed by Li, He proved the convergence of the scheme in distributed nonconvex learning even more from theoretical analysis. (Zhang et al. 2023c) studied the existing FL attacks and detection schemes, and found that most detection schemes have high false positive rates in the setting of non-independent and identical distribution. Therefore, they proposed a Kalman filter-based cross-round detection. This detection scheme identifies adversaries by looking for behavioral changes before and after attacks, so as to adapt to data heterogeneity and improve detection accuracy.

In terms of FL model performance enhancement, existing FL secure aggregation algorithms cannot satisfy the security and reliability of the FL process for resource-constrained IoT devices without significantly affecting the model accuracy and performance. Since the performance of edge-IoT devices is usually limited to support high performance consuming FL robust aggregation algorithms, (Cao et al. 2024) proposed a secure robust FL framework with a trusted execution environment, which adopts a shared representation learning approach to classify the model into a sensitive model and a representational model used for client training, where the sensitive model is always retained in the secure environment, and the representational model is in the real environment for normal training and aggregation. The representation learning approach allows each FL participant to train its own personalized model, which improves the model convergence rate and accuracy. Meanwhile, the framework, in order to enhance the robustness of FL models in non-independent and homogeneously distributed scenarios, designs a robust affiliation-based multi-model aggregation method, which uses affiliations generated by soft clustering to classify clients and uses multi-model methods to perform aggregation separately to enhance robustness.

(Du et al. 2023a) considered that the device operating environment restricts high-quality annotated data extraction among the participants, and propose a forgotten optimized aggregation strategy combining Kalman filter and cubic exponential smoothing, which improves the global aggregation of models to enhance model performance. Meanwhile, a deep learning network combining multi-scale convolution, attention mechanism, and multilevel residual connectivity is also used to extract multi-client data features simultaneously to improve the accuracy and generalization of the aggregation algorithm in local model training with multiple participants.

(Wang et al. 2023) focused on the fact that the performance of the over-the-air FL is usually limited by the devices with the worst channel conditions of the edge servers, and considered that the use of reconfigurable smart surfaces can alleviate the communication problem of over-the-air FL and develop a learning algorithm based on graph neural network to map the channel coefficients directly to the optimized network parameters, which reduces the computational complexity of the algorithm by exploiting the alignment equivalence and invariance of the graph to achieve the aggregation algorithm dimensionality independently of the number of edge devices.

The schemes proposed by Cao et al., Du et al. and Wang et al. focus on the performance of FL robust aggregation algorithms under the constraints of edge-IoT device resources and operating environment. And the former scheme focuses on evaluating the model accuracy

of the proposed aggregation algorithms when resource-limited IoT devices are subject to byzantine and backdoor attacks, while the latter scheme focuses on improving the accuracy of the model aggregation as well as the communication and computational efficiency when a large number of local models are aggregated, focusing on improving the accuracy and efficiency of the FL system. However, allowing all devices to participate in the FL process is not a long-term feasible solution, and the heterogeneity of edge devices under IoT in terms of data quality power, arithmetic, storage, etc., and the poor communication links of some of the participants can affect the FL performance. Therefore, optimal client selection also becomes an important stage in the FL process.

3.1.2 FL client selection algorithms

In the FL process, the client devices participating in each round of training can be accurately and efficiently selected based on the device performance, connection quality and other indicators, which helps to improve the efficiency and performance of FL. Meanwhile, a fair client selection algorithm design can also prevent malicious clients from participating in training using techniques such as authentication and reputation assessment, reducing the harm of malicious attacks on the model (Mayhoub and M. Shami T, 2023). To this end, this section reviews recent client selection algorithms for FL, analyzes and evaluates client selection algorithms in terms of both features and limitations.

In designing client selection algorithms to enhance FL robustness, reduce communication and computational overhead and improve model convergence speed and accuracy. (Jiang et al. 2023) analyzed a number of security strategies to mitigate label-flipping attacks in FL, and proposed a malicious client detection approach for defending against label-flipping attacks against the huge computational overhead and lack of robustness required by these strategies, which is achieved by training a parameter server with a lightweight generator on the parameter server to detect the quality of training data for each client. The generator performs data quality detection without retraining and does not require any prior knowledge, which satisfies the lightweight design and privacy security requirements. However, Jiang et al.'s scheme does not focus on the frequent exchanges of model parameters carried out between the massive edge client devices and the server side, and due to the limited communication resources, the frequent exchanges of massive parameters cause large communication delays and affect the efficiency of the FL system.

Aiming at how to design the FL client selection algorithm to improve the FL communication efficiency, (Yang et al. 2023b) first proposed a client selection scheme based on the kernelized stan difference between the global posterior and the local skewed distributions, the updates provided by clients with large kernelized stan differences can minimize the local free energy of each iteration, and the probability of such clients to be selected is the largest, and the resulting communication overhead is also smaller.

Meanwhile, (Wehbi et al. 2023) proposed an intelligent client selection method by considering the problems of data quality, computational and communication resource heterogeneity that exists among IoT devices, and analyzed the substantial problems of schemes that select FL clients based on a random selection strategy. The method overcomes the limitation that most client selection schemes follow a unilateral selection strategy, and uses matching game theory to propose a bilateral client selection method for FL, taking into account the preferences of FL servers and client devices in the selection process.

(Huang et al. 2023b) investigated the use of a clustered FL approach to solve the problem of heterogeneous data, and found that the current clustered FL process is relatively

slow. Considering the lack of an effective client selection strategy, proposed the use of active learning to select participating clients for each cluster. The scheme filters out some of the group clients that partially provide the most information in each round of FL based on active learning metrics and only aggregates their model updates to update cluster-specific models. Where the active learning metrics are set to uncertainty sampling, committee query, and loss. The active client selection FL scheme proposed by Huang requires fewer participating clients, which can significantly speed up the learning process and significantly improve the model accuracy with lower communication overhead.

Enhancing the robustness and data security of FL systems in EC environments involves multiple factors and considerations, and more current technical solutions are mostly researched from the perspectives of communication security, device security, model robustness, communication and computation overhead, and client selection. However, the network attacks in untrustworthy EC environments are gradually diversified and complicated, and how to develop FL frameworks with higher security performance while taking into account the system overhead and efficiency is still a key direction for future research on secure robust FL.

3.2 FL information privacy-preserving

FL solves the privacy security problem of sensitive data in ML environments. However, the uploads and downloads of the model update parameters, training iterations, and other processes still expose the FL environment to a series of risks. Examples include malicious speculation by semi-honest adversaries and theft of data by curious attackers (Narayanan and Shmatikov 2008). The privacy in FL can be divided into global and local privacy. Global privacy requirements and local device-generated model updates protect the privacy of all unreliable third parties except for the trusted central aggregation server in each iteration. Local privacy requires model updates to protect the server privacy. At present, typical technologies for improving FL privacy security include cryptographic technologies, disturbance technologies, adversarial training (AT), blockchain, and KD.

3.2.1 Cryptography technologies

Common cryptography technologies in FL encrypt the model parameter information that must be uploaded in plaintext. This process enhances the privacy and security protection performance of the FL systems. At present, the commonly used encryption methods include secure multi-party computation (SMPC) and homomorphic encryption (HE). Each encryption technology has unique technical characteristics. For example, SMPC can keep user input data confidential and allows multiple parties to perform joint computation on private data, but the computing overhead is expensive. Compared with SMPC, HE schemes have a similar performance in security protection. However, they consume fewer computing resources than SMPC.

SMPC, also known as MPC, was originally proposed to protect the inputs of multi-party participants. In the FL framework, the SMPC is used to protect model updates for clients. SMPC ensures that each participant in the FL system only recognizes its own inputs and outputs, and ensures that it has a complete lack of knowledge regarding other clients. Using SMPC to build an FL security model can increase efficiency by reducing the security requirements. Kilbertus et al. (Kilbertus et al. 2018) used SMPC methods for validation and model training to prevent local private data from being known to other users.

Sharemind (Bogdanov et al. 2008) designed an SMPC framework as a secure and efficient computer system. (Kalapaaking et al. 2022) proposed a CNN-based FL rack by combining SMPC-based aggregation and cryptographic inference methods. Encrypted on-premise models were sent to the cloud for SMPC-based cryptographic aggregation, resulting in an encrypted global model. Ultimately, the encrypted global model was returned to each edge server for more localized training, thereby further improving the accuracy of the model. This solution solved the FL privacy security problem in the IoT environment under 6th-generation networks, and ensured the accuracy of the model and confidentiality of the model parameters. However, this solution did not consider the costs of multi-source heterogeneous data encryption processing and communication resource consumption in the IoT environment.

(Li et al. 2023f) designed a vertical FL-ring (VFL-R) for FL models with limited communication sources and low computing power in the coordinators. VFL-R was a novel vertical FL framework combined with a ring architecture for multi-party collaborative modeling. The VFL-R framework simplified the intricate communication architecture of all parties and provided protection against semi-honest attacks. In addition, it reduced the influence of coordinators in the modeling process. (Berry and Komninos 2022) faced the problems of the SMPC's large number of computational rounds and high costs of transmitting data between parties. They proposed an efficient optimization framework for a CNN with SMPC. This framework combined various optimization methods from a broader privacy-preserving DL field. It included batch normalization for privacy-preserving and polynomial approximations of the activation functions.

The SMPC is a lossless solution that allows multiple parties to perform joint computations on private data. SMPC keeps data content confidential and provides strong privacy protection. As a research-oriented solution, the SMPC-based FL privacy protection scheme still faces many challenges. The main problem is the trade-off between the FL system efficiency and privacy. The SMPC encryption and decryption process takes a long time, which may negatively affect the model training. The design of a lightweight SMPC solution remains a significant challenge. Therefore, many researchers choose HE technologies. Under the premise of having the same security performance scheme, the lower computing consumption of HE technologies has made them the first choice for many researchers in designing FL security protocol frameworks.

HE refers to plaintext and encryption operations. The obtained result is equivalent to a result obtained by first encrypting the plaintext to obtain ciphertext and then performing the same operation on the ciphertext. Owing to this advantageous feature, ML can entrust a third party to process data without revealing the information.

HE has been introduced into the FL framework to encrypt local gradient updates, gradient parameters, and other model parameter information. This prevents private data from being leaked by adversaries. The early HE algorithms used single-key arithmetic. Homomorphic operations could only be performed between values with the same public key. This required the clients to share the same private key. However, if the same private key was shared by multiple clients, there was a risk of private key leakage. There was also an increased risk of malicious clients accessing the data of other clients. This was undoubtedly a significant test for FL privacy protection using HE schemes. Therefore, (López-Alt et al. 2012) proposed a multi-key HE method. Later generations have continued to innovate and improve on this basis to solve the limitations of single-key HE approaches. It has recently been discovered that multi-key HE can be used in FL scenarios to protect the privacy of model updates. However, they may not be protected against attacks by malicious actors that disrupt the course of learning, such as Byzantine attacks.

(Ma et al. 2022a) proposed a multi-key fully HE multi-key-Cheon-Kim-Kim-Song (MK-CKKS) scheme for supporting approximate fixed-point algorithms. It required an aggregated public key for encryption and decryption and required each device to calculate its decryption share. The ciphertext was successfully decrypted only when the number of private keys participating in decryption reached a certain threshold. MK-CKKS required all data contributors to collaborate to decrypt the aggregated results, thereby guaranteeing the confidentiality of the model updates. It was robust to attacks by malicious actors and collusive attacks by actors on servers. (Hou et al. 2021) proposed a verifiable privacy protection scheme (VPRF) based on a vertical joint random forest. The VPRF utilized homomorphic comparisons and voting statistics with a multi-key HE to protect privacy. (Zhang et al. 2023d) combined distributed Paillier cryptography and zero-knowledge proofs based on existing Byzantine robust FL algorithms. They proposed an FL scheme that balanced robustness and privacy protection.

In solving the problems of traditional privacy protection, FL solutions cannot simultaneously provide efficient data confidentiality and lightweight integrity verification. (Ma et al. 2022b) proposed a verifiable privacy-preserving FL scheme (VPFL) for EC systems. The VPFL combined the distributed selection stochastic gradient descent method with the Paillier HE system. They also proposed an online and offline signature method for lightweight gradient integrity verification. (Zhao et al. 2022) designed a decentralized privacy-preserving and verifiable federated learning framework based on efficient and verifiable cryptographic matrix multiplication. The framework effectively defends against various inference attacks by ensuring the confidentiality of global and local model updates and the verifiability of all training steps. It realizes the integrity of the federated learning model training and improves the training efficiency of the federated learning system.

In recent years, secret sharing schemes have been widely concerned and used to design federated security protocols as a special encryption method that is both secure and efficient. Secret sharing technology can be applied to the training process of FL to ensure the privacy of model parameter sharing. Secret sharing splits the parameters of the federated learning model into multiple parts and sends them to different participants, and ensures that only when the number of participants exceeds the recovery threshold specified by the system, can they cooperate with each other to recover the model information. In theory, the federated learning based on secret sharing can protect the local data set of the client from the semi-honest central aggregation server and other participants. In the case of collusion between some clients and the central aggregation server, secret sharing can still provide privacy security guarantee.

(Zhou et al. 2021) proposed a privacy-preserving FL framework that combines Shamir with HE to ensure that aggregate values can be correctly decrypted only when the number of participants is greater than t. Tasiu et al. (Muazu et al. 2024) proposed a secure FL system based on data fusion, which employs a convolutional neural network with an effective weight sharing method for prediction, and uses multi-party computation and additive secret sharing to encrypt the weights of the model to protect the privacy of the gradient parameters of the federated learning training model. (Wang et al. 2022a) proposed a privacy-preserving scheme for FL under EC, and designed a lightweight privacy-preserving protocol based on shared secret and weight mask, which achieves higher accuracy and training efficiency than HE, and can resist device dropout and collusion attacks between devices. However, the above schemes solve the trade-off between prediction accuracy and model privacy by protecting the gradient parameters, and do not consider the high communication cost caused by the transmission of massive secret segments. At the same time, in the EC

environment, the processing and transmission delay of edge devices with different performance are also quite different.

Therefore, (Liu et al. 2022) designed a secure aggregation protocol based on an effective additional secret sharing in fog computing setting to solve the problem that the training process of FL needs to perform secure aggregation frequently. Firstly, the protocol used fog nodes as intermediate processing units to provide local services to help the cloud server aggregate the sum during the training process. Then a lightweight Request-then-Broadcast method is designed to ensure that the protocol is robust to lost clients. The protocol achieves low communication and computation overhead. (Xie et al. 2022) designed a lossless multi-party federated XGBoost learning model based on secret sharing. The model framework reshaped the segmentation criterion calculation process of XGBoost in the secret sharing setting, and solved the quadratic optimization problem in a distributed way to perform leaf weight calculation, so that the model could run quickly in a secure manner. However, when facing the complex and large FL network, the secret sharing technology needs to allocate more secret segments. At the same time, due to the different equipment performance of each participant in the FL, and the different types, structures, and quality of the data sets, the local model training time and quality of each participant are different, which leads to the fact that multiple participants cannot receive or submit their own secret clips at the same time, and the FL system has higher bandwidth requirements for network transmission.

Nevertheless, it remains necessary to solve the problems of member inference attacks and reverse attacks leading to training data privacy leakages, e.g., the address security problems of most existing cryptography approaches in FL and required additional computing. HE still has great room for technological innovation and improvements in terms of the computing efficiency, complexity of the interaction logic, and secret sharing schemes in terms of communication delay and communication bandwidth cost.

3.2.2 Differential privacy technology

In DML, fuzzy processing technology is often used to protect the privacy security of the training datasets. This includes performing randomization, noise disturbance, generalization, and compression to obfuscate the training data and improve the privacy performance to a certain extent. In FL, DP is often used to add noise disturbance to the original training set data, model parameters, or gradient information, so as to hide key features of the data. DP can achieve privacy data protection.

(Dwork 2008) proposed the concept of DP in 2006 and provided a rigorous mathematical proof of its security. DP mainly protects data privacy and security by adding noise to sensitive data. The introduction of DP in FL to add noise disturbances to the model parameters uploaded by FL participants or to use generalization methods to hide key data features prevents a reverse retrieval of data, so that the ML models can resist adversarial samples (Ibitoye et al. 2021).

DP has a lower overhead than SMPC's high communication overhead. Many advanced DP algorithms have been proposed in existing studies. For example, (Wang et al. 2019a) designed a deep neural network (DNN) learning framework that supported DP by considering the risk of privacy leakages of sensitive crowdsourced data. The framework evaluated important features related to the target class labels. It used adaptive noise figures to accommodate heterogeneous input features. Finally, a noise disturbance was added to the affine transformation of the input features according to the importance and heterogeneity of

the input features. (Geyer et al. 2017) proposed a client-side DP-protection FL optimization algorithm. (McMahan et al. 2017) added user-level privacy protection to the FL averaging algorithm to design a user-level DP training algorithm for large neural networks. The purpose of both was to protect private data by hiding the local model parameters uploaded by users during training, thereby balancing the model performance and privacy loss. Both algorithms were validated using actual datasets. This proved that, with sufficient devices participating in federated training, privacy protection could be achieved with a small additional overhead. Simultaneously, both approaches guaranteed high model accuracy.

However, this method did not consider that the introduction of DP in FL with fewer participants may lead to impaired overall model accuracy. To this end, (Huang et al. 2020) substituted DP noise into a neural network by pruning a given layer of the neural network, aiming to protect private data from leakage without reducing the model accuracy. Lin et al. (Lin et al. 2022) designed a novel privacy-preserving learning framework based on graph neural networks (GNNs). The framework had a formal privacy guarantee based on edge-local DP to protect both node features and edge privacy. It was highly integrated with a GNN with a privacy utility guarantee to protect users' data privacy under a given privacy budget.

In general, the factors with greater influence on the accuracy of the model are the noise disturbances and clipping degrees. Bu et al. (Bu et al. 2021) utilized the advantages of the linear algebraic properties of neural tangent kernel matrices. A convergence analysis framework for DP DL suitable for general neural network structures and loss functions was established. In a continuous-time analysis, the authors verified that the main influence on the model convergence was not noise, but the degree of sample clipping. Thus, a global cropping method was designed. Compared to traditional local cropping methods, the global clipping loss was small, the calibration was better, and the model prediction accuracy was less-impacted. However, the discrete-time convergence at large learning rates required further study (in such cases, the addition of noise affects the model convergence to some extent.

The above improvements strove to strike a balance between privacy protection and model accuracy. However, they did not deeply consider the privacy computing costs of model iterations or the added model complexity with the introduction of DP. To this end, (Zhao et al. 2021) designed a multi-level and multi-participation dynamic allocation method for a privacy budget. A new adaptive differential private FL algorithm was designed to balance privacy and utility. (Andrew et al. 2021) proposed an adaptive gradient-clipping strategy. This strategy added noise to a specified layer while applying adaptive fractional clipping to an iterative DP mechanism. This strategy alleviated the problem of excessive hyperparameters in DP algorithms.

3.2.3 Adversarial training

In recent years, information privacy-preserving methods based on cryptography and disturbance technologies have been widely applied in FL. These technologies primarily concern raw data, parameter encryption, and secure local computing. They pass the results of the computation to a third party to aggregate the computation results, which can significantly reduce the risk of privacy leakages in the distributed learning process. However, malicious attackers can steal data from other honest actors by deploying GANs. (Wang et al. 2019b) designed a GAN on a central aggregation server to steal the private data of users. Using calculated gradient information, the adversary could reverse some or all of the private

data. In recent years, many approaches have been proposed for stealing private data in FL systems through GANs. Aiming to resist such adversarial attacks, a significant amount of research has been conducted aiming to protect the privacy in FL. The main objectives are detection and defense. Defenses against adversarial attacks have the following three main directions.

- Modify the training process or modify the input sample during the testing stage.
- Modify the neural network, such as by improving the activation function or loss function and adding or deleting the number of sublayers in the neural network.
- Identify adversarial samples or completely classify adversarial samples.

AT is the first line of defense against adversarial attacks, and has been introduced into FL to strengthen data privacy security. AT participates in federated model training, using the real and adversarial samples as a training set. AT enhances local real-world data privacy security through adversarial sample perturbation. AT is an active defense technique. It attempts to arrange all attacks from the training phase of the client, making the FL global model robust to known adversarial attacks (Tramèr et al. 2017). AT requires the use of large amounts of training data and high-intensity adversarial samples; therefore, it can regularize neural networks to reduce overfitting. In turn, the resistance of the neural network is enhanced and the best empirical robustness is obtained (Papernot et al. 2018; Croce and Hein 2020; Tramer et al. 2020).

Early AT defenses were focused on detection and prevention. For example, (Baracaldo et al. 2017) used background information such as sources and transformations to detect toxic sample points in a training set, and (Arjovsky et al. 2017) prevented inference attacks by generating fake training data. However, recent studies have found that the security threats to concentrated AT are gradually increasing. (Song et al. 2019) found that AT makes ML models more vulnerable to member inference attacks than those trained using the original training set. (Mejia et al. 2019) found that with a model inversion attack, an attacker can use an AT model to generate images that are similar to actual training samples. (Zhang et al. 2022) developed a new privacy attack method that destroyed the privacy of the DL systems in AT models. First, the feature information was recovered from the gradient. The recovered features were then used as supervised reconstruction inputs.

Facing insecure AT, (Ryu and Choi 2022) proposed a hybrid AT method. This scheme used clean images denoised by denoising networks, clean images without denoising, and adversarial samples to train DNN models. This scheme improved the robustness of DNNs against a wide range of adversarial attacks. (Wang et al. 2022b) introduced a semi-supervised learning mechanism with virtual AT to avoid overfitting during DL model training. (Rashid et al. 2022) used IoT datasets to explore the impacts of adversarial attacks on DL, and proposed a method using AT. This method significantly improved the performance of IDSs in adversarial attacks.

At present, most AT methods use AT examples to improve the model robustness. However, most AT approaches require additional computational time and overhead for computational gradients. To this end, (Jia et al. 2022) designed an adversarial initialization method for the dependent samples of a fast AT. This method realized sample dependence by generating benign samples and their gradient information in the training target network. However, this law did not consider the high computational costs and time required to deploy large-scale AT on resource-constrained edge devices in FL networks. (Tang et al. 2022) proposed a federated adversarial decoupling learning framework. The framework applied decoupled greedy learning (DGL) to federated AT to reduce computations and

memory usage. In addition, the framework added an auxiliary weight decay to improve the vanilla DGL and mitigate target inconsistencies. The experimental results showed that the federated adversarial decoupling learning framework significantly reduced the computing resources consumed by AT while maintaining almost the same accuracy and robustness as concentrated joint training.

Most researchers have focused on the accuracy of AT models. AT injects adversarial sample samples into the model training to improve the robustness of DNN models against adversarial GANs. A slight perturbation of the adversarial sample to the original sample may affect the accuracy of the model. To this end, (Yu et al. 2022a) designed a meta-learning-based AT algorithm framework to avoid the performance degradations caused by the generated adversarial samples. (Zhou et al. 2022) proposed a latent-boundary-guided AT framework. The framework trained DNN models on adversarial samples, as guided by potential boundaries. High-quality adversarial sample samples were generated by adding perturbations to potential features. This approach achieved a better trade-off between the standard accuracy and adversarial robustness. In general, AT improves the privacy of the user data. Adding AT samples minimizes the threat of inference to the actual training data. In the latest research on improving the robustness in AT, the trade-off between standardization and robustness has received widespread attention. This also provides a new direction for the future development of federal confrontation training.

3.2.4 Blockchain

The traditional centralized FL framework relies on a central aggregation server and therefore has a single point of failure. When communication is busy, the central node incurs higher communication costs and low efficiency. Participants lack incentive mechanisms and are not highly motivated to participate in joint learning. There is also a lack of security mechanisms to identify malicious users who compromise the model. To address these shortcomings, many researchers have combined blockchain with FL. First, the participating nodes of the blockchain are used to replace the central server to reduce the single-point-failure problem. Next, miner nodes are used to calculate the local device model update parameters without uploading raw data. Subsequently, the consensus mechanism of the blockchain is used to verify and record the local device model updates. The aggregate model parameters are uploaded by local devices, and the global model updates are added to new blocks. Finally, each local device downloads the global model from the blockchain blocks.

(Miao et al. 2022) designed a blockchain-based privacy-preserving Byzantine robust FL (PBFL) scheme. The PBFL used cosine similarity to determine the gradients uploaded by malicious clients. Fully HE was used to provide secure aggregation. In general, PBFL uses a blockchain system to facilitate the implementation of transparent PBFL processes and regulations, thereby mitigating the impacts of central servers and malicious clients. (Durga and Poovammal 2022) proposed a novel framework based on blockchain and FL models. The FL model was responsible for reducing the complexity, whereas the blockchain helped protect the privacy of distributed data. This framework used a hybrid capsule learning network to develop models that protected privacy while performing accurate predictions. (Wang et al. 2022c) aimed at the problem of untrusted third parties in FL by adopting a distributed blockchain to distribute tasks and collection models. A reputation calculation method was proposed to calculate the real-time reputations of task participants. (Yu et al. 2022b) designed an overall framework for a blockchain-based FL system. The framework

utilized distributed ledger technology to mitigate the problems of single points of failure and low-quality or poisoned data interference models in FL systems. It was designed to enhance the security and scalability of FL systems.

The blockchain-based FL technologies introduced above have focused on using the characteristics of blockchain technology to alleviate privacy and security issues in FL systems. However, blockchain technology can increase the complexity of FL systems. It is also possible that the FL system may become inefficient because of an inefficient consensus mechanism. To this end, many researchers regard the research and development of efficient and lightweight blockchain consensus mechanisms as a major research hotspot for the future development of blockchain-based FL technologies. (Yang et al. 2022) designed a credit data and model-sharing architecture based on FL and blockchain. The framework ensured the secure storage and sharing of credit information in a distributed environment. The framework proposed a permission control contract and credit verification contract for the security authentication of the results of the credit sharing model under FL. The efficient credit data storage mechanism, combined with a removable bloom filter, ensured a unified consensus of the training and calculation processes. (Li et al. 2022b) discussed existing quantum blockchain schemes and analyzed the reasons for the inefficiency of the current blockchain consensus mechanisms. A consensus mechanism called a quantum-delegated proof was constructed using quantum voting and provided rapid decentralization for quantum blockchain schemes. (Du et al. 2023b) investigated a blockchain-assisted EC scenario. A matching mechanism based on a smart contract was proposed to establish a lease association between EC nodes and data service operators. A trust-driven proof-of-benefit consensus mechanism was designed to realize verification of the transactions and fair remuneration distributions.

The integration of blockchain and FL technologies has largely alleviated the issues faced in the traditional FL field. However, after the integration of these two technologies, there remain problems caused by the blockchain itself. For example, the traditional blockchain consensus mechanism and network structure cause problems such as long transaction confirmation times, limited throughput, and complex communication structures. This also leads to an increase in the model update parameter aggregation delays in the blockchain network for each round of the FL process. Each FL participant uses a different local device. When the uploaded model is updated in the blockchain network, the time delays of each device may not be uniform. This can also lead to a decrease in the prediction accuracy of the trained global model (Zhu et al. 2022). Considering the existing problems of the blockchain-based FL frameworks, the current decentralized FL architecture approach is rapidly gaining popularity (Hu et al. 2019). Decentralized training has also been shown to be more efficient than centralized training when running federated systems in low-bandwidth or high-latency networks (Xiao et al. 2020; Liu et al. 2020; Jiang et al. 2020). Combined with the blockchain-based asynchronous FL framework proposed by (Feng et al. 2021), blockchain ensures that the model parameters in the chain are not tampered with. Simultaneously, the asynchronous FL accelerates the global aggregation. We find that an asynchronous FL framework based on blockchain can solve the problems in balancing privacy, security, and efficiency faced in current FL technology development to a certain extent.

3.2.5 Knowledge distillation

KD technology originated from the concept of transferring knowledge from large models to small models, and was formally proposed by (Hinton et al. 2015) in 2015. The core idea

of KD concerns the transfer of knowledge. A student model obtains an accuracy comparable to that of a teacher model by imitating it. A complete KD system consists of three parts: knowledge, distillation algorithm, and teacher-student architecture. The distillation algorithm is the core step for determining how the knowledge of the teacher model is transferred to the student model.

In the area of privacy protection, traditional ML and DL methods are vulnerable to privacy attacks. For example, an attacker can obtain individual information or relevant training data from the model parameters, gradient information, and target model. Therefore, the relevant organizations and individuals have high requirements for data privacy and security. FL has been widely adopted as a way to access private raw data training sets without disclosing them to other participants. KD can isolate the access to the original training datasets of each participant in the FL system through the teacher-student architecture of KD, e.g., by letting the teacher learn the private data training model and then transferring this knowledge to the outside model.

(Wang et al. 2022d) proposed an adversarial KD. This scheme combined KD with backdoor attacks. The KD reduced the anomalous features in the model results caused by label flipping, allowing the model to bypass defenses. Meanwhile, there also exists a problem of privacy leakage when using co-distillation to solve the problem of unbalanced data distribution. (Gong et al. 2023) proposed a FL framework for an integrated attention distillation to protect privacy. The framework utilized unlabeled public data for one-way offline KD and learning from local knowledge with an integrated attention distillation. The framework isolated dispersed and heterogeneous local data through the KD, thereby significantly reducing the risk of privacy leakage. Nevertheless, while focusing on privacy protection performance, the framework ignored the huge communication costs of FL and the huge computing overhead caused by KD. (Wu et al. 2022) proposed the FedKD FL method. This method was based on adaptive mutual KD and dynamic gradient compression techniques. FedKD accelerated the efficiency of FL model training, thereby alleviating the huge communication costs in FL and improving the communication efficiency and effectiveness.

Studies have also considered the effects of data structures and distribution heterogeneity on FL performance. (Li et al. 2022c) proposed an FL framework with a decentralized KD. This framework introduced a decentralized KD module. That is, no data were stored on the server to protect local privacy. The global model was trained by extracting the knowledge of the local model based on a divergence measure defined in the loss function and by approaching the mean value of the neural network map. The impact of the heterogeneous data on the FL system performance was mitigated while protecting the local private data. Facing complex model data in the edge intelligence scenario, (Sepahvand et al. 2022) proposed an adaptive teacher-student learning algorithm with decomposition KD. The algorithm used Tucker decomposition to decompose a high-dimensional feature graph at the end of the teacher. It obtained a core tensor that students could easily understand from the teacher's feature graph. The teacher-student architecture designed by the algorithm was used in edge intelligent devices, and greatly improved the FL system efficiency and privacy security performance in edge-IoT.

KD can satisfy the FL privacy and security requirements to a certain extent. Different types of knowledge can be leveraged in complex source data scenarios. This alleviates the problem of the original training data being heterogeneous and less diverse. However, the KD technology itself has limitations. When dealing with large-scale models, large amounts of system resources must be consumed. For large-scale distributed FL scenarios, the training of more heterogeneous datasets and a large number of communication rounds

are challenges to the use of KD technology. Therefore, there is still a large space for further research on the combination of KD and FL to improve the efficiency of using KD and FL for privacy protection.

4 Challenges and future directions

We discussed the major security and privacy issues of FL under edge networks. These include the major mechanisms, attacks, and possible countermeasures. However, there are still emerging privacy security challenges and issues that have yet to be explained or require further exploration from the perspective of the EC-assisted IoT paradigm. This section explains some of these challenges, and provides insights into promising future research directions.

4.1 Challenges

The rapid development of the SIoT not only promotes the wide application of FL; it also results in higher requirements for FL. This section summarizes the challenges facing the future development of FL by combining the current development status of FL with the development needs of edge-IoT.

4.1.1 Secured and efficient edge FL

Currently, numerous research endeavors underscore the compromise of FL model training efficiency in favor of bolstering privacy safeguards within the FL framework. Specifically, privacy-preserving FL that relies on encryption techniques confronts a delicate balance between model training efficiency and privacy assurance. For instance, the utilization of encryption to secure model parameters introduces a noteworthy consideration: escalating the encryption level leads to an augmented computational overhead. Consequently, with the substantial influx of client interactions in the domain of Edge- IoT, the task of enhancing both model training efficiency and privacy security performance within Edge FL becomes a substantial challenge. This challenge is particularly pertinent as the goal is to establish a dependable FL system within the context of an inherently unreliable edge network environment.

4.1.2 Lightweight byzantine robust FL

A large number of client devices exist in the mobile edge-IoT scenario. FL may not be able to safely and efficiently integrate various model trainers with varying access, identities, purposes, and requirements. Secure authentication authorization is required for participants who access FL systems anytime and anywhere. The effectiveness and safety of training data with complex and varied structures must also be evaluated. The training process must be resilient to malicious enemy poisoning damage and other privacy security threats. The above issues require us to design a set of Byzantine robust security FL architectures for

future edge-IoT environments. These architectures can accommodate a large user base and provide lightweight security against hostile Byzantine adversaries.

4.1.3 Heterogeneous data

FL model training requires considerable safe and reliable data support. The unified integration, efficient utilization, safe storage, and sharing of the heterogeneous data resources in edge-IoT environments remain urgent issues that need to be addressed. They also require further exploration from an edge computation-assisted FL perspective and strategies for data confidentiality, integrity, and privacy. There is also a need to develop flexible, fine-grained, and adaptive data-analysis solutions. Such schemes can automatically identify the degree of sensitivity of edge user data and provide a corresponding appropriate security mechanism. In addition, we consider the significant differences in computing and storage capabilities between the cloud servers and edge nodes. Protecting node data may not be feasible when using the traditional security approach originally proposed for cloud servers to protect edge devices. In addition, an edge-IoT network is distributed, scalable, and heterogeneous. This is a challenge for security mechanisms that must maintain the efficiency and privacy of the data storage with auditing, backup, and recovery.

4.1.4 FL privacy security and mobile 6G

Future sixth-generation (6G) mobile networks are envisioned as heterogeneous, ultra-dense, and highly dynamic intelligent networks. The emergence of sixth-generation mobile network technology will also accelerate the realization of the edge-SIoT. The integration of 6G connects physical systems with the digital space and enables powerful and instant wireless connectivity. With growing concerns about data privacy, FL is considered as a promising solution for deploying distributed data processing and learning in wireless networks. FL can take full advantage of the distributed computing resources in mobile EC systems, allowing users to retain their private data locally. However, the unreliable communication channels, limited resources, and lack of trust between users hinder the effective application of FL in the IoT. This is because, in a mobile edge environment, the system bandwidth is limited and shared by all connected mobile devices (which may interfere with one another). In addition, owing to mobility and channel fading, a selected device may have different computing powers and dynamic wireless channel conditions. Therefore, the 6G era will lead to higher requirements for safe and reliable FL models.

4.1.5 Asynchronous FL privacy-preserving

Current IoT smart devices may not have sufficient computing resources to train and deploy an entire learning model. Simultaneously, the transmission of continuous real-time data to a central server with high computing resources incurs huge communication costs and raises data security and privacy concerns. FL is a promising solution for training ML models by using resource-limited devices and edge servers. However, most existing studies adopt an unrealistic synchronous parameter update method for homogeneous IoT nodes under stable communication connections. Therefore, asynchronous FL has been proposed as a new research approach that avoids using a single central server. Asynchronous FL improves the training efficiency of heterogeneous IoT devices in unstable communication networks by

allowing nodes to join or exit during the learning process. In business-based FL instances, asynchronous FL may be more beneficial when customers do not trust third parties. Although there is no raw data sharing, the open architecture and extensive collaboration in asynchronous FL still provide some malicious actors with a great opportunity to infer the training data of other parties. This can lead to serious privacy issues. Therefore, future research should be conducted on asynchronous FL privacy security mechanisms for EC.

4.1.6 FL universal safety mechanism design

The research and development of privacy protection mechanisms should be applicable to different FL classification scenarios. The current federal privacy and security protocols are mainly developed for HFL. When applied to VFL and federated transfer learning, they are not only inefficient, but also insecure. Therefore, the development of privacy protection mechanisms applicable to various FL classification scenarios remains a major challenge for the development of privacy-preserving FL.

4.1.7 Stringent latency demands

Contemporary intelligent IoT applications at the edge impose significant requirements on FL to satisfy the stringent latency prerequisites of client services. Examples encompass intelligent driving reliant on edge terminals and real-time medical diagnosis and analysis. Nonetheless, the issue of delay stemming from repeated communication rounds in FL curbs its convergence velocity. This delay arises due to FL clients awaiting the aggregation of all local models before initiating a new global model round and commencing the subsequent training phase. Consequently, the delay encountered in the FL communication process becomes a substantial impediment, substantially constraining the integration and application of FL within the domain of edge-based intelligent IoT networks.

4.2 Future research directions

This section combines the proposed future challenges for FL in the edge-IoT context as a direction guide. We propose meaningful research directions for the development of future FL based on new technologies with broad development prospects.

4.2.1 Secure and efficient FL based on over-the-air computing

The proliferation of edge Internet of Things devices has prompted the necessity for expedited processing of extensive sensor data, resulting in protracted data processing delays. Over-the-air computing, which merges communication and computation, presents a solution by accomplishing computing tasks during data transmission, potentially mitigating the data processing delay conundrum. Rooted in the concept of "communication and computing integration", over-the-air computing harnesses the signal waveform superposition attribute during transmission to facilitate rapid data aggregation.

The fusion of over-the-air computing with FL involves executing computational tasks while transmitting model updates. By adopting a collaborative design strategy that integrates computation and communication, the expedited aggregation of the FL global model is achieved. Concurrently, the computation for updating FL's local models takes place during the communication phase. This method adeptly alleviates the possible vulnerability

of the local model computation process to semi-honest or malicious central aggregation servers, resulting in a notable enhancement of data confidentiality. In the forthcoming landscape of edge-intelligent networks teeming with terminal devices, over-the-air computing emerges as a promising avenue to enhance the performance of distributed model training. Exploiting this technology to expedite FL model aggregation warrants comprehensive exploration. Additionally, over-the-air computing leverages the wireless channel's superposition characteristics to thwart privacy breaches during data communication. The optimization of radio resource allocation to ensure heightened confidentiality throughout the FL process in over-the-air computing also necessitates extensive investigation. Hence, the incorporation of over-the-air computing for secure and efficient FL design constitutes a compelling avenue for future research, holding considerable potential in the context of advancing both privacy-preserving and model efficiency.

4.2.2 Quantum technology designs lightweight robust FL

An FL security protocol designed using encryption technology can provide a large degree of defense against illegal attacks and prevent privacy leakages. However, the high computing overhead and heavy communication costs of encryption technology in large FL systems render FL less suitable. Therefore, we consider that the high encryption rate, high execution rate, and high security of quantum encryption (QE) technology may be suitable for an FL lightweight secure and robust protocol design. QE technology mainly uses quantum properties and principles. It comprises a series of encryption technologies such as key generation, plaintext obfuscation encryption, ciphertext restoration and decryption, key preservation and transmission, and anti-eavesdropping. Eavesdropping by an intermediate adversary, copying, or tampering may cause a quantum-state change that exposes such eavesdropping. The key to the QE is generated randomly in the process of communication. This ensures that the key cannot be eavesdropped upon or cracked. In addition, quantum key distribution technology ensures that the communication parties only need to share the secret key once, thereby ensuring the security of transmitting the encrypted information in the open channel.

In the design of a secure and robust FL aggregation protocol, QE technology and applications can be mixed with classical encryption technology and applications. For example, for key management, we can design a quantum key pool to realize fast key distribution and reduce the waiting time for the key exchange for a large number of edge devices. We can combine quantum key distribution technology to ensure the random and secure generation and sharing of keys. Compared with traditional encryption methods, QE technology can effectively improve the encryption and decryption efficiency in large FL scenarios and simplify the complex key generation and differentiation processes. As a hot research direction in the future, quantum technology can not only be used to design robust edge FL security protocols that are more secure and reliable, but can also help realize lightweight communication, encryption, and decryption.

4.2.3 Federated transfer learning based on knowledge distillation

The massive data structures in edge-IoT are heterogeneous and cannot be effectively integrated. Consequently, FL always faces heterogeneity challenges caused by the distribution of non-identically independently distributed data from different clients with different computing and communication capabilities. Severe data heterogeneity leads to client-side drift,

resulting in unstable model convergence and poor performance. Federated transfer learning comprises a combination of FL and transfer learning techniques for allowing knowledge to be shared while protecting private data. This is set up for FL not only for different sample spaces but also for different feature spaces, which can help build effective and accurate ML models for applications with only a small amount of data and weak supervision. KD is a teacher-student training structure. Usually, the teacher model provides knowledge and the student model extracts the knowledge of the teacher model through distillation training. Knowledge from the complex teacher model is then transferred to the student model at a small cost.

As an attractive option, federated transfer learning employing KD addresses the problems of data heterogeneity and small amounts of training data. By employing KD to perform federated transfer learning from large models to compact models, the FL models can be compressed. Simultaneously, KD can reduce the bandwidth occupied by the FL training and improve the communication efficiency in FL. Therefore, as an effective method to solve data heterogeneity, federated transfer learning with KD can alleviate the costs in large-scale FL training and communication to a certain extent.

4.2.4 SIoT privacy security based on FL and blockchain

A large amount of the data stored by edge-IoT devices are locally involved in the FL model training process. This process is always at risk of theft, poisoning, and tampering by malicious adversaries. Blockchain is used as a distributed ledger to store records, and data written in the blocks cannot be corrupted. This increases the transparency and immutability of data and facilitates data sharing while reducing opportunities for deception and fraud.

In the FL blockchain, each participating device acts as a client to update the parameters and aggregate the learning models in a decentralized manner. Each local client uses a local dataset to train a local model and uploads the trained model parameters to a group of miners. The miner then merges the model parameters uploaded by the local client into a block, which is validated by the miner using the mining process. Once the block is mined, these model parameters are attached to the blockchain and broadcast to the entire network. Each local client downloads the latest update block and calculates the global model parameters. This iteration is performed only until the model accuracy requirement or maximum number of iterations is reached. The immutable and de-neutral nature of the blockchain helps eliminate the FL's need for a central server. This can be used to solve privacy security issues in edge-IoT scenarios. At the same time, blockchain incentives can attract users in the federated system to actively participate in the training of statistical models and alleviate problems such as communication costs and communication delays to a certain extent. To this end, blockchain technology has been introduced into the development of the SIoT. Researchers have used the confidentiality, integrity, and availability of blockchains to improve the security of private data.

4.2.5 FL-based intrusion detection system protects 6G wireless communications

The provision of communication services via portable and mobile devices (such as air base stations) is a key concept for implementation of such services in 6G networks. Traditional FL partially solves privacy concerns by sharing models with base stations. However,

the centralized nature of FL only allows devices near the base station to share the trained model. Moreover, there are many potential threats to wireless channel communication, such as denial of service attacks, data security attacks, and illegal authentication accesses. In addition, long-distance communication forces devices to increase their transmission power, raising energy efficiency issues.

Most of the existing IDS models are built using ML and DL algorithms. This makes it difficult to train a participant's local dataset without compromising user privacy. In addition, the amount of training data held by a single organization is limited; this has a significant impact on the accuracy of the ML and DL models. Therefore, it is necessary to expand the amount of data while protecting its security. In addition, the increasing complexity of the network environment also leads to higher requirements for the applicability of the IDS. The processing of massive data also results in higher requirements for the efficiency of an IDS. The emergence of FL has enabled ML and DL models to be efficiently trained while protecting participant data privacy. A distributed IDS under the FL mechanism significantly improves the training efficiency and guarantees the privacy data security and classification accuracy. Therefore, in an unreliable 6G wireless communication channel, improving the FL-based IDS mechanism can facilitate the security filtering of such massive data to a certain extent. This reduces the threat of wireless channel security attacks while mitigating the risk of private data leakages.

4.2.6 Asynchronous FL based on trusted blockchain

Considering the integration of blockchain and FL technologies, there may still be problems caused by blockchain itself. For example, the traditional blockchain consensus mechanism and network structure cause problems such as long transaction confirmation times, limited throughput, and complex communication structures. These problems may lead to an increase in the model update parameter aggregation delay in the blockchain network for each round of the FL process. Each participant in the federation has a different device: when models are uploaded to the blockchain network separately, the time delay of each device may not be uniform. This could also result in a decrease in the prediction accuracy of the trained global model. In asynchronous FL, the central aggregation server undertakes global aggregation shortly after amassing a limited set of local models. This prompt aggregation strategy mitigates the influence of underperforming clients on the overall efficiency of FL global model training.

Asynchronous FL solves the problem of inefficient FL systems caused by the different performances of the edge devices. However, the times of uploading or downloading the model information by each device node are not synchronized. This may cause gradient delays at some nodes. Asynchronous FL introduces a trusted blockchain as an FL server. The child blockchains are used for partial model parameter updates, and the main blockchain is used for global model parameter updates. The blockchain is decentralized, transactions are conducted in real time, and the nodes in the various blocks communicate with each other in a timely manner. These features can also alleviate the problem of gradient staleness in the model training process of asynchronous FL, resulting in degradation of the accuracy of the global model. The trusted identity authentication mechanism of blockchain can filter out unreliable devices that apply to access the FL in the edge-IoT, thereby reducing FL privacy and security risks from the source.

4.2.7 Low-latency FL for edge computing systems

Edge computing involves real-time data processing and analysis proximate to the data source, effectively addressing demands at the edge node. This approach enhances data processing efficiency, alleviates data communication burdens, and fortifies data privacy and security. In this context, FL framework has been tailored to align with the edge computing paradigm. FL's local model updates are outsourced to the edge server for aggregation. Edge clients' local models undergo multiple rounds of training updates on the edge computing server before being uploaded to the central server for global aggregation. This process efficiently alleviates communication pressures stemming from centralized model training, thereby enhancing FL's convergence rate. The reduction in communication process delay satisfies the prerequisites of edge-IoT applications with stringent FL processing delay requirements.

5 Conclusion

This study comprehensively investigates and analyzes the threat of data security attacks and privacy leakage security issues faced by FL under the edge-IoT. First, we introduce relevant concepts and basic working principles of FL, EC, and other technologies. We also summarize the data security attacks and privacy leakage attacks in FL. We then discuss the current mainstream security protection technologies for FL data security and privacy leakage attacks. For data security attacks, we investigate mainstream defense measures (such as the intrusion detection mechanism) and the robust performance of FL security aggregation. For privacy leakage attacks, we analyze mainstream defense measures such as cryptography technologies, disturbance technologies, AT, blockchain, and KD. The advantages and limitations of the FL security models developed using various defense technologies are discussed. Finally, in the context of EC IoT-assisted FL, we discuss the challenges of FL security protocol design and possible future research directions.

The rapid development of SIoT has increased the complexity of DML environments at the edge. EC offloads heavy data storage processing to edge server nodes, alleviating the computing pressure of the central node and further compensating for the shortcomings of edge devices. Implementing FL under the EC paradigm will further mitigate the negative impacts of the FL privacy enhancement on the FL efficiency and model accuracy. Simultaneously, FL security protocols are being developed to balance privacy performance and system efficiency.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 61862007, Guangxi Natural Science Foundation under Grant 2020GXNSFBA297103.

Authors contribution Haiao Li: Writing – original draft, Methodology, Formal analysis, Data curation, Resources, Visualization. Lina Ge: Writing – review & editing, Conceptualization, Supervision, Project administration. Lei Tian: Project administration, Supervision, Validation.

Declarations

Competing interests The authors declare no competing interests.

References

Adhikari M, Menon VG, Rawat DB, Li XW (2023) Guest Editorial Introduction to the Special Section on Computational Intelligence and Advanced Learning for Next-Generation Industrial IoT. IEEE Transac Network Sci Eng 10(5):2740–2744

Ahmad S, Shakeel I, Mehfuz S, Ahmad J (2023) Deep learning models for cloud, edge, fog, and IoT computing paradigms: Survey, recent advances, and future directions. Computer Sci Rev 49:100568

Andrew G, Thakkar O, McMahan B, Ramaswamy S (2021) Differentially private learning with adaptive clipping. Adv Neural Inf Process Syst 34:17455–17466

Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, vol 70, pp. 214–223. PMLR, Sydney, NSW, Australia

Baracaldo N, Chen B, Ludwig H, Safavi JA (2017) Mitigating poisoning attacks on machine learning models: A data provenance-based approach. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 103–110. ACM, Dallas, Texas, USA

Berry C, Komninos N (2022) Efficient optimisation framework for convolutional neural networks with secure multiparty computation. Comput Secur 117:102679

Bhagoji AN, Chakraborty S, Mittal P, Calo S (2019) Analyzing federated learning through an adversarial lens. In: Proceedings of the 36th International Conference on Machine Learning, pp. 634–643. PMLR, Long Beach, California, USA

Bogdanov D, Laur S, Willemson J (2008) Sharemind: A framework for fast privacy-preserving computations. European Symposium on Research in Computer Security. Springer, Berlin, Heidelberg, pp 192–206

Bu ZQ, Wang H, Dai ZY, Long Q (2021) On the convergence and calibration of deep learning with differential privacy. arXiv preprint arXiv:2106.07830

Cao YH, Zhang JB, Zhao YR, Su PC, Huang HX (2024) SRFL: A Secure & Robust Federated Learning framework for IoT with trusted execution environments. Expert Syst Appl 239:122410

Chen X, Yu HN, Jia XH, Yu XZ (2023) APFed: Anti-Poisoning Attacks in Privacy-Preserving Heterogeneous Federated Learning. IEEE Trans Inf Forensics Secur 18:5749–5761

Croce F, Hein M (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning, pp.2206–2216, PMLR, Virtual Event

Douceur JR (2002) The sybil attack. International workshop on peer-to-peer systems, vol 2429. Springer, Berlin, Heidelberg, pp 251–260

Du Y, Wang Z, Li J, Shi L, Jayakody DNK, Chen W, Han Z (2023b) Blockchain-Aided Edge Computing Market: Smart Contract and Consensus Mechanisms[J]. IEEE Trans Mob Comput 22(6):3193–3280

Du J, Qin N, Huang D, Jia XM, Zhang YM (2023) An Efficient Federated Learning Framework for Machinery Fault Diagnosis with Improved Model Aggregation and Local Model Training. IEEE Transactions on Neural Networks and Learning Systems 1–24. (Early Access)

Durga R, Poovammal E (2022) FLED-Block: Federated Learning Ensembled Deep Learning Blockchain Model for COVID-19 Prediction. Front Public Health 10:892499

Durrant A, Markovic M, Matthews D, May D, Enright J, Leontidis G (2022) The role of cross-silo federated learning in facilitating data sharing in the agri-food sector. Comput Electron Agric 193:106648

Dwork C (2008) Differential privacy: A survey of results. International conference on theory and applications of models of computation. Springer, Berlin, Heidelberg, pp 1–19

Fan JQ, Wang XH, Guo YX, Hu XP, Hu B (2022) Federated learning driven secure internet of medical things. IEEE Wirel Commun 29(2):68–75
Fan MC, Ji KL, Zhang ZF, Yu HF, Sun G (2023) Lightweight Privacy and Security Computing for Blockchained Federated Learning in IoT. IEEE Internet Things J 10(18):16048–16060
Fang C, Guo YB, Ma JL, Xie HD, Wang YF (2022) A privacy-preserving and verifiable federated learning method based on blockchain. Comput Commun 186:1–11
Fang MH, Cao XY, Jia JY, Gong N (2020) Local model poisoning attacks to {Byzantine-Robust} federated learning. In: 29th USENIX Security Symposium (USENIX Security 20), pp. 1605–1622. USENIX, Boston, Massachusetts, USA
Feng L, Zhao Y, Guo S, Qiu X, Li W, Yu P (2021) BAFL: A Blockchain-Based Asynchronous Federated Learning Framework. IEEE Trans Comput 71(5):1092–1103
Fraboni Y, Vidal R, Lorenzi M (2021) Free-rider attacks on model aggregation in federated learning. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, vol 130, pp. 1846–1854. PMLR, Buenos Aires, Argentina
Friha O, Ferrag MA, Shu L, Maglaras L, Wang XC (2021) Internet of things for the future of smart agriculture: A comprehensive survey of emerging technologies. IEEE/CAA J Automatica Sinica 8(4):718–752
Garg D, Alam M (2023) Smart agriculture: a literature review. J Management Anal 10(2):359–415
Ge LN, Li HA, Wang X, Wang Z (2023) A review of secure federated learning: privacy leakage threats, protection technologies, challenges and future directions. Neurocomputing 561:126897
Geyer R C, Klein T, Nabi M (2017) Differentially private FL: A client level perspective. arXiv preprint arXiv:1712.07557
Ghosh AM, Grolinger K (2020) Edge-cloud computing for internet of things data analytics: embedding intelligence in the edge with deep learning. IEEE Trans Industr Inf 17(3):2191–2200
Gong XL, Chen YJ, Huang HY, Liao YQ, Wang S, Wang Q (2022) Coordinated Backdoor Attacks against Federated Learning with Model-Dependent Triggers. IEEE Network 36(1):84–90
Gong X, Song L, Vedula R, Sharma A, Zheng M, Planche B, Innanje A, Chen T, Yuan JS, Doermann D, Wu ZY (2023) Federated Learning with Privacy-Preserving Ensemble Attention Distillation. IEEE Trans Med Imaging 42(7):2057–2067
Goodfellow I J, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
Guo JJ, Li HY, Huang FR, Liu ZQ, Peng YG, Li XH, Ma JF, Menon VG, Lgorevich KK (2022) ADFL: A poisoning attack defense framework for horizontal federated learning. IEEE Trans Industr Inf 18(10):6526–6536
Guo W, Wang YJ, Chen X, Jiang PY (2023) Federated transfer learning for auxiliary classifier generative adversarial networks: framework and industrial application. J Intell Manuf 2023:1–16
Hammi B, Idir YM, Zeadally S, Khatoun R, Nebhen J (2022) Is it really easy to detect sybil attacks in C-ITS environments: a position paper. IEEE Trans Intell Transp Syst 23(10):18273–18287
Hatamizadeh A, Yin H, Molchanov P, Myronenko A, Li WQ, Dogra P, Feng A, Flores MG, Kautz J, Xu DG, Roth HR (2023) Do gradient inversion attacks make federated learning unsafe? IEEE Trans Med Imaging 42(7):2044–2056
He X, Zhu H, Ling Q (2023) C-RSA: Byzantine-robust and communication-efficient distributed learning in the non-convex and non-IID regime. Signal Process 213:109222
Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531
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, pp. 603–618. ACM, Dallas, Texas, USA
Hou J, Su M, Fu A, Yu Y (2021) Verifiable privacy-preserving scheme based on vertical federated random forest. IEEE Internet Things J 9(22):22158–22172
Hu L, Yan AL, Yan HY, Huang T, Zhang YY, Dong CY, Yang CS (2023) Defenses to Membership Inference Attacks: A Survey. ACM Comput Surv 56(4):1–34
Hu C, Jiang J, Wang Z (2019) Decentralized federated learning: A segmented gossip approach. arXiv preprint arXiv:1908. 07782
Hua HC, Li YT, Wang TH, Dong NQ, Li W, Cao JW (2023) Edge computing with artificial intelligence: A machine learning perspective. ACM Comput Surv 55(9):1–35
Huang XH, Han L, Li DD, Xie K, Zhang Y (2023a) A reliable and fair federated learning mechanism for mobile edge computing. Comput Netw 226:109678
Huang Y, Su Y, Ravi S, Song Z, Arora S, Li K (2020) Privacy-preserving learning via deep net pruning. arXiv preprint arXiv:2003.01876

Huang HL, Shi W, Feng YH, Niu CY, Cheng GQ, Huang JC, Liu Z (2023) Active Client Selection for Clustered Federated Learning. IEEE Transactions on Neural Networks and Learning Systems 1–15 (Early Access)
Ibitoye O, Shafiq M O, Matrawy A (2021) DiPSeN: Differentially Private Self-normalizing Neural Networks For Adversarial Robustness in FL. arXiv preprint arXiv:2101.03218
Jan MA, Zhang W, Khan F, Abbas S, Khan R (2023) Lightweight and smart data fusion approaches for wearable devices of the Internet of Medical Things. Information Fusion 103:102076
Jia X, Zhang Y, Wu B, Wang J, Cao X (2022) Boosting fast AT with learnable adversarial initialization. IEEE Trans Image Process 31:4417–4430
Jiang J, Hu L, Hu C, Liu J, Wang Z (2020) BACombo—Bandwid-th-aware decentralized federated learning. Electronics 9(3):440–455
Jiang Y, Zhang W, Chen Y (2023) Data quality detection mechanism against label flipping attacks in federated learning. IEEE Trans Inf Forensics Secur 18:1625–1637
Jin S, Li Y, Chen X, Li RX, Shen ZB (2023) Blockchain-based fairness-enhanced federated learning scheme against label flipping attack. J Inform Secur App 77:103580
Kalapaaking AP, Stephanie V, Khalil I, Atiquzzaman M, Yi X, Almashor M (2022) SMPC-Based Federated Learning for 6G-Enabled Internet of Medical Things. IEEE Network 36(4):182–189
Kilbertus N, Gascón A, Kusner M, Veale M, Gummadi K, Weller A (2018) Blind justice: Fairness with encrypted sensitive attributes. In: International Conference on Machine Learning, pp. 2630–2639. PMLR, Stockholm, Sweden
Li XW, Chen BH, Yang DQ, Wu GF (2022a) Review of Security Protocols in Edge Computing Environments. J Comp Res Develop 59(4):765–780
Li Q, Wu J, Quan J, Shi J, Zhang S (2022b) Efficient Quantum Blockchain with a Consensus Mechanism QDPoS. IEEE Trans Inf Forensics Secur 17:3264–3276
Li H, Li CC, Wang J, Yang AM, Ma ZZ, Zhang ZQ, Hua DB (2023a) Review on security of federated learning and its application in healthcare. Futur Gener Comput Syst 144:271–290
Li YL, Yuan D, Sani AS, Bao W (2023b) Enhancing Federated Learning robustness in adversarial environment through clustering Non-IID features. Comput Secur 132:103319
Li DF, Lai JH, Wang RJ, Li X, Vijayakumar P, Gupta BB, Alhalabi W (2023c) Ubiquitous intelligent federated learning privacy-preserving scheme under edge computing. Futur Gener Comput Syst 144:205–218
Li BB, Wang PR, Shao ZR, Liu A, Jiang YK (2023d) Defending Byzantine attacks in ensemble federated learning: A reputation-based phishing approach. Futur Gener Comput Syst 147:136–148
Li J, Yan T, Ren P (2023f) VFL-R: a novel framework for multi-party in vertical federated learning. Appl Intell 53:12399–12415
Li X, Chen B, Lu W (2022) FedDKD: Federated Learning with Decentralized Knowledge Distillation. arXiv preprint arXiv:2205.00706
Li J, Rakin A S, Chen X, Yang L, He ZZ, Fan DL, Chakrabarti C (2023) Model Extraction Attacks on Split Federated Learning. arXiv preprint arXiv:2303.08581
Lin J, Du M, Liu J (2019) Free-riders in federated learning: Attacks and defenses. arXiv preprint arXiv:1911.12560
Lin W, Li B and Wang C, Towards Private Learning on Decentralized Graphs with Local Differential Privacy, IEEE Transactions on Information Forensics and Security 17: 2936–2946
Liu W, Chen L, Chen Y, Zhang W (2020) Accelerating federated learning via momentum gradient descent. IEEE Trans Parallel Distrib Syst 31(8):1754–1766
Liu Y, Dong Y, Wang H, Jiang H, Xu Q (2022) Distributed fog computing and federated-learning-enabled secure aggregation for IoT devices. IEEE Internet Things J 9(21):21025–21037
Liu Z, Lin HY, Liu Y (2023a) Long-Term Privacy-Preserving Aggregation with User-Dynamics for Federated Learning. IEEE Trans Inf Forensics Secur 18:2398–2412
Liu HF, Li B, Gao CL, Xie P, Zhao CL (2023b) Privacy-Encoded Federated Learning Against Gradient-Based Data Reconstruction Attacks. IEEE Trans Inf Forensics Secur 18:5860–5875
López-Alt A, Tromer E, Vaikuntanathan V (2012) On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In: Proceedings of the forty-fourth annual ACM symposium on Theory of computing, pp.1219–1234. ACM, New York, NY, USA
Ma J, Naas SA, Sigg S, Lyu X (2022a) Privacy-preserving FL based on multi-key homomorphic encryption. Int J Intell Syst 37(9):5880–5901
Ma X, Zhou Y, Wang L, Miao M (2022b) Privacy-preserving byzantine-robust FL. Computer Standards & Interfaces 80:103561
Mayhoub S, M. Shami T (2023) A Review of Client Selection Methods in Federated Learning. Archives of Computational Methods in Engineering 1–24

Mcmahan B, Moore E, Ramage D, Hampson S, Arcas BAY (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, pp. 1273–1282. PMLR, Ft Lauderdale, USA

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

Mejia F A, Gamble P, Hampel-Arias Z, Lomnitz M, Tindall L, Barrios MA (2019) Robust or Private? AT Makes Models More Vulnerable to Privacy Attacks. arXiv preprint arXiv: 1906.06449

Miao YB, Liu ZT, Li HW, Choo KKR, Deng RH (2022) Privacy-Preserving Byzantine-Robust Federated Learning via Blockchain Systems. IEEE Trans Inf Forensics Secur 17:2848–2861

Muazu T, Mao Y, Muhammad AU, Ibrahim M, Kumshe UMM, Samuel O (2024) A federated learning system with data fusion for healthcare using multi-party computation and additive secret sharing. Comput Commun 216:168–182

Myrzashova R, Alsamhi SH, Shvetsov AV, Hawbani A, Wei X (2023) Blockchain meets federated learning in healthcare: A systematic review with challenges and opportunities. IEEE Internet Things J 10(16):14418–14437

Nair AK, Raj ED, Sahoo J (2023) A robust analysis of adversarial attacks on federated learning environments. Computer Standards & Interfaces 86:103723

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy, pp. 111–125. IEEE, Oakland, California, USA

Nguyen TD, Nguyen T, Nguyen PL, Pham HH, Doan KD, Wong K (2024) Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions. Eng Appl Artif Intell 127:107166

Ni L, Gong X, Li JF, Tang YC, Luan Z, Zhang JQ (2023) rFedFW: Secure and Trustable Aggregation Scheme for Byzantine-Robust Federated Learning in Internet of Things. Inf Sci 653:119784

Ning ZL, Hu H, Wang XJ, Guo L, Guo S, Wang GY, Gao XB (2023) Mobile Edge Computing and Machine Learning in The Internet of Unmanned Aerial Vehicles: A Survey. ACM Comput Surv 56(1):1–31

Papernot N, McDaniel P, Sinha A, Wellman MP (2018) Sok: Security and privacy in machine learning. In: 2018 IEEE European Symposium on Security and Privacy, pp. 399–414. IEEE, London, United Kingdom

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

Qi P, Chiaro D, Guzzo A, Ianni M, Fortino G, Piccialli F (2023) Model aggregation techniques in federated learning: A comprehensive survey. Futur Gener Comput Syst 150:272–293

Ranaweera P, Jurcut AD, Liyanage M (2021) Survey on multi-access edge computing security and privacy. IEEE Commun Surv Tutorials 23(2):1078–1124

Rashid MM, Kamruzzaman J, Hassan MM, Lmam T, Wibowo S, Gordon S, Fortino G (2022) AT for Deep Learning-based Cyberattack Detection in IoT-based Smart City Applications. Comput Secur 120:102783

Rodríguez-Barroso N, Martínez-Cámara E, Luzón MV, Herrera F (2022a) Dynamic defense against byzantine poisoning attacks in federated learning. Futur Gener Comput Syst 133:1–9

Rodríguez-Barroso N, Martínez-Cámara E, Luzón MV, Herrera F (2022b) Backdoor attacks-resilient aggregation based on Robust Filtering of Outliers in federated learning for image classification. Knowl-Based Syst 245:108588

Ryu G, Choi D (2022) A hybrid AT for deep learning model and denoising network resistant to adversarial examples. Appl Intell 52(15):1–14

Sepahvand M, Abdali-Mohammadi F, Taherkordi A (2022) An adaptive teacher–student learning algorithm with decomposed knowledge distillation for on-edge intelligence. Eng Appl Artif Intell 117:105560

Sharma S, Guleria K (2023) A comprehensive review on federated learning based models for healthcare applications. Artif Intell Med 146:102691

Shen M, Gu A, Kang J, Tang XY, Lin XD, Zhu LH, Niyato D (2023) Blockchains for Artificial Intelligence of Things: A Comprehensive Survey. IEEE Internet Things J 10(16):14483–14506

Shi W, Cao J, Zhang Q, Li YHZ, Xu LY (2016) Edge computing: Vision and challenges. IEEE Internet Things J 3(5):637–646

Shuvo MMH, Islam SK, Cheng JL, Morshed BI (2022) Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. Proc IEEE 111(1):42–91

Singh A (2006) Eclipse attacks on overlay networks: Threats and defenses. In: Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications, IEEE, Barcelona, Catalunya, SPAIN

Song L, Shokri R, Mittal P (2019) Privacy risks of securing machine learning models against adversarial examples. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 241–257. ACM, London, United Kingdom

Tang M, Zhang J, Ma M, Divalentin L, Ding A, Hassanzadeh A, Li H, Chen Y (2022) FADE: Enabling Large-Scale Federated AT on R-esource-Constrained Edge Devices. arXiv preprint arXiv:2209.03839

Tramer F, Carlini N, Brendel W, Madry A (2020) On adaptive attacks to adversarial example defenses. Adv Neural Inf Process Syst 33:1633–1645

Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T (2016) Stealing machine learning models via prediction {APIs}. In: 25th USENIX security symposium (USENIX Security 16), pp. 601–618. USENIX, Austin, TX, USA

Tramèr F, Kurakin A, Papernot N, Goodfellow L, Boneh D, Mcdaniel P (2017) Ensemble Adversarial Training: Attacks and defenses. arXiv preprint arXiv:1705.07204

Wan FY, Ma T, Hua Y, Liao B, Qing XL (2022) Secure distributed estimation under Byzantine attack and manipulation attack. Eng Appl Artif Intell 116:105384

Wang Y, Gu M, Ma J, Jin Q (2019a) DNN-DP: Differential privacy enabled deep neural network learning framework for sensitive crowdsourcing data. IEEE Transact Comput Soc Systems 7(1):215–224

Wang R, Lai J, Zhang Z, Li X, Vijayakumar P, Karuppiah M (2022a) Privacy-preserving federated learning for internet of medical things under edge computing. IEEE J Biomed Health Inform 27(2):854–865

Wang F, Wu X, Wang H (2022b) Seismic horizon identification using semi-supervised learning with virtual AT. IEEE Trans Geosci Remote Sens 60:1–11

Wang WL, Wang YJ, Huang Y, Mu CC, Sun ZC, Tong XR, Cai ZP (2022c) Privacy protection federated learning system based on blockchain and edge computing in mobile crowdsourcing. Comput Netw 215:109206

Wang Z, Zhou Y, Zou Y, Bennis M (2023) A graph neural network learning approach to optimize ris-assisted federated learning. IEEE Trans Wireless Commun 22(9):6092–6106

Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H (2019) Beyond inferring class representatives: User-level privacy leakage from FL. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, pp. 2512–2520. IEEE, Paris, France

Wang Y, Fan W, Yang K, Alhusaini N, Li J (2022) A Knowledge Distillation-Based Backdoor Attack in Federated Learning. arXiv preprint arXiv:2208.06176

Wehbi O, Arisdakessian S, Wahab OA, Otrok H, Otoum S, Mourad A, Guizani M (2023) FedMint: Intelligent Bilateral Client Selection in Federated Learning with Newcomer IoT Devices. IEEE Internet Things J 10(23):20884–20898

Wu C, Wu F, Lyu L, Huang Y, Xie X (2022) Communication-efficient federated learning via knowledge distillation. Nat Commun 13(1):3–8

Xiao P, Cheng S, Stankovic V, Vukobratovic D (2020) Averaging is probably not the optimum way of aggregating parameters in federated learning. Entropy 22(3):314–325

Xiao X, Tang Z, Li CY, Xiao B, Li KL (2022) SCA: sybil-based collusion attacks of IIoT data poisoning in federated learning. IEEE Trans Industr Inf 19(3):2608–2618

Xie L, Liu J, Lu S, Chang TH, Shi Q (2022) An efficient learning framework for federated XGBoost using secret sharing and distributed optimization. ACM Transact Intell Systems Technol (TIST) 13(5):1–28

Xu CH, Qu YY, Xiang Y, Gao LX (2023) Asynchronous federated learning on heterogeneous devices: A survey. Computer Sci Rev 50:100595

Yang DS, Luo SL, Zhou JJ, Pan LM, Yang XN, Xing JY (2023a) Efficient and persistent backdoor attack by boundary trigger set constructing against federated learning. Inf Sci 651:119743

Yang J, Liu Y, Kassab R (2023b) Client Selection for Federated Bayesian Learning. IEEE J Sel Areas Commun 41(4):915–928

Yang F, Qiao Y, Abedin MZ, Huang C (2022) Privacy-Preserved Credit Data Sharing Integrating Blockchain And Federated Learning For Industrial 4.0. IEEE Transactions on Industrial Informatics 18(12): 8755–8764

Yu C, Zhang Z, Li H, Sun J, Xu Z (2022a) Meta-learning-based AT for deep 3D face recognition on point clouds. Pattern Recogn 134:109065

Yu F, Lin H, Wang X, Yassine A, Hossain MS (2022b) Blockchain-empowered secure federated learning system: Architecture and applications. Comput Commun 196:55–65

Zhang F, Wu RF, Guan JW, Zheng Z, Guo XG, Zhang X, Du XY, Shen XP (2023a) Expanding the Edge: Enabling Efficient Winograd CNN Inference with Deep Reuse on Edge Device. IEEE Trans Knowl Data Eng 35(10):10181–10196

Zhang Z, Li J, Yu S, Makaya C (2023b) SAFELearning: Secure Aggregation in Federated Learning with Backdoor Detectability. IEEE Trans Inf Forensics Secur 18:3289–3304

Zhang JL, Liu Y, Wu D, Lou S, Chen B, Yu S (2023d) VPFL: A verifiable privacy-preserving FL scheme for edge computing systems. Digital Communications and Networks 9(4):981–989
Zhang J, Chen Y, Li H (2022) Privacy Leakage of AT Models in FL Systems. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.108–114. IEEE, New Orleans, LA, USA
Zhang X, Liu Q, Ba Z, Hong Y, Zheng T, Lin F, Lu L, Ren K (2023) Fltracer: Accurate poisoning attack provenance in federated learning. arXiv preprint arXiv:2310.13424.
Zhao JZ, Mao KM, Huang CX, Zeng YY (2021) Utility Optimization of FL with Differential Privacy. Discret Dyn Nat Soc 2021:3344862
Zhao J, Zhu H, Wang F, Lu R, Liu Z, Li H (2022) PVD-FL: A privacy-preserving and verifiable decentralized federated learning framework. IEEE Trans Inf Forensics Secur 17:2059–2073
Zheng W, Cao Y, Tan H (2023) Secure sharing of industrial IoT data based on distributed trust management and trusted execution environments: a federated learning approach. Neural Comput Appl 35(29):21499–21509
Zhou Z, Tian Y, Peng C (2021) Privacy-preserving federated learning framework with general aggregation and multiparty entity matching. Wirel Commun Mob Comput 2021:1–14
Zhou X, Tsang IW, Yin J (2022) LADDER: Latent boundary-guided adversarial training. Mach Learn 111(11):1–29
Zhu JC, Cao JN, Saxena D, Jiang S, Ferradi, (2022) Blockchain-empowered Federated Learning: Challenges, Solutions, and Future Directions. ACM Comput Surv 55(11):1–31
Zhu RB, Li MY, Yin JJ, Sun LB, Liu H (2023) Enhanced Federated Learning for Edge Data Security in Intelligent Transportation Systems. IEEE Trans Intell Transp Syst 24(11):13396–13408