# Membership Inference Attacks and Defenses in Federated Learning: A Survey

LI BAI, HAIBO HU, QINGQING YE, and HAOYANG LI, The Hong Kong Polytechnic University, Hong Kong

 ${\sf LEIXIA~WANG, Renmin~University~of~China, China}$ 

JIANLIANG XU, Hong Kong Baptist University, Hong Kong

Federated learning is a decentralized machine learning approach where clients train models locally and share model updates to develop a global model. This enables low-resource devices to collaboratively build a high-quality model without requiring direct access to the raw training data. However, despite only sharing model updates, federated learning still faces several privacy vulnerabilities. One of the key threats is membership inference attacks, which target clients' privacy by determining whether a specific example is part of the training set. These attacks can compromise sensitive information in real-world applications, such as medical diagnoses within a healthcare system. Although there has been extensive research on membership inference attacks, a comprehensive and up-to-date survey specifically focused on it within federated learning is still absent. To fill this gap, we categorize and summarize membership inference attacks and their corresponding defense strategies based on their characteristics in this setting. We introduce a unique taxonomy of existing attack research and provide a systematic overview of various countermeasures. For these studies, we thoroughly analyze the strengths and weaknesses of different approaches. Finally, we identify and discuss key future research directions for readers interested in advancing the field.

CCS Concepts: • Security and privacy  $\rightarrow$  Privacy protections.

Additional Key Words and Phrases: Membership inference attacks, federated learning, deep leaning, privacy risk

#### **ACM Reference Format:**

Li Bai, Haibo Hu, Qingqing Ye, Haoyang Li, Leixia Wang, and Jianliang Xu. XXXX. Membership Inference Attacks and Defenses in Federated Learning: A Survey. *ACM Comput. Surv.* 37, 4, Article 111 (August XXXX), 35 pages. https://doi.org/XXXXXXXXXXXXXXXX

## 1 INTRODUCTION

With the increasing availability of extensive datasets, machine learning (ML) has emerged as a critical technology, facilitating significant advancements across various domains, including computer vision [1–5], natural language processing [6–9], and more. Notably, legal regulations such as the General Data Protection Regulation (GDPR) [10] and the California Consumer Privacy Act (CCPA) [11] establish key guidelines for data sharing between organizations and mandate the safeguarding of users' privacy when utilizing such data. Federated learning (FL), as proposed in [12], is a distributed machine learning paradigm that enables multiple clients to collaboratively

Authors' addresses: Li Bai, baili.bai@connect.polyu.hk; Haibo Hu, haibo.hu@polyu.edu.hk; Qingqing Ye, qqing.ye@polyu.edu.hk; Haoyang Li, hao-yang9905.li@connect.polyu.hk, The Hong Kong Polytechnic University, Hong Kong, Hong Kong; Leixia Wang, leixiawang@ruc.edu.cn, Renmin University of China, Beijing, China; Jianliang Xu, xujl@comp.hkbu.edu.hk, Hong Kong Baptist University, Hong Kong, Hong Kong.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

0360-0300/XXXX/8-ART111 \$15.00

https://doi.org/XXXXXXXXXXXXXXX

111:2 Bai et al.

train a machine learning model (*global model*) without directly sharing their private data. This approach offers a practical solution to overcome these privacy constraints. Unlike centralized machine learning, which relies on data aggregation, FL allows training samples to remain local across various organizations or mobile devices. This approach not only enhances the volume of available training data but also supports large-scale training processes. Meanwhile, by decoupling model training from direct access to raw data, FL enables each client to maintain their training data locally, ensuring compliance with current legal regulations and helping safeguard data privacy.

While FL is designed to be privacy-aware by preventing direct access to raw training data, it remains susceptible to significant privacy risks. Researchers have demonstrated that adversaries can exploit model updates in FL to reconstruct the original training data and labels [13-15], infer properties of other clients' training data [16, 17], and even generate representative samples [18–20]. Among these privacy risks, membership inference attack (MIA) represents a fundamental privacy violation, which seek to determine whether a specific record is part of the training dataset [21-24]. Compelling applications for MIAs include: 1) Privacy breach: MIA can reveal sensitive details about the training data of machine learning models to potential attackers. For example, if an adversary determines that a medical record was used to train a cancer prediction model, they may infer that the individual has cancer. 2) Data censorship: MIA serves as an effective tool for auditing data privacy and compliance. For instance, under the legal requirement of the right to be forgotten, MIA can be employed to verify whether a platform has successfully erased a specific data point following a deletion request. 3) Foundation of advanced attacks: MIA can act as a foundational step in strengthening more sophisticated privacy attacks. Attackers can refine their strategies to explore the target model by determining which samples were used as the training data, such as model extraction attacks [25]. A considerable body of research is dedicated to MIAs specifically crafted for the FL environment. To name a few, the study [26] first proposes an inference algorithm to deduce membership information by exploiting gradients, hidden layer output, and sample loss during the learning process. Since then, the research community has attempted to extend it to various domains, such as classification models [27-30], regression models [31] and recommender systems [32], by either exploiting exchanged model update [26, 33, 34] or the trend of model outputs [29, 30, 35, 36].

Although related work explores MIAs and defenses in both centralized and federated settings, there are significant differences between these approaches, as shown in Table 1. These differences motivate us to focus specifically on works within the federated learning framework. Considering a centralized learning (CL) setting, where the training data is gathered on a single server [37, 38], machine learning models are trained using the aggregated dataset and then released to the public. The differences in MIAs and defensive mechanisms between CL and FL settings are as follows.

- 1 Attacker/Defender role: The roles of adversaries and defenders differ between two settings. In CL, most MIAs are usually conducted by model consumers who primarily have access to model outputs [39]. In FL, however, potential attackers mainly stem from insiders, including the central server and other clients. In the case of defenses against CL-related MIAs, the responsibility for privacy protection lies with the model owners [40]. However, both the central server and clients can implement defensive strategies to prevent membership information leakage in FL [41, 42].
- 2 Attack/Defense phase: The phase of when an adversary launches an attack or when a defense algorithm is applied varies. In CL, most MIAs take place during the inference phase, after the target model has been released. In contrast, MIAs in FL focus on the training phase, attempting to compromise membership privacy during the entire convergence process. In terms of defense, both centralized and federated settings aim to protect data privacy during

- the training stage. However, the key distinction is that centralized defenses also focus on safeguarding the model's output during the inference stage.
- 3 Adversary knowledge: Attackers in FL can gain more detailed adversarial knowledge compared to those in CL. MIAs in CL can occur in two settings: in the white-box setting, where attackers have access to the target model and intermediate layer computations, or in the black-box setting, where only the model's outputs, such as prediction scores [22, 43, 44] or labels [45, 46], are available for analysis. In contrast, adversaries in FL have extensive access to the target model's gradients, intermediate computations, and final outputs throughout the learning process [26]. Furthermore, FL attackers can closely observe the model's convergence, allowing them to access multiple historical versions of the target model.
- 4 **Active strategy**: The proactive methods used by attackers to infer membership privacy differ across settings. In FL, adversaries with legitimate access to the training process can maliciously manipulate models, enabling them to conduct powerful MIAs through model poisoning [26, 47]. In contrast, CL attackers may poison the data to amplify membership information leakage [48, 49].
- 5 **Protection core**: Since the types of private information at risk of being leaked vary, the targets of defense mechanisms across settings also differ. In CL, defenses aim to make the model outputs indistinguishable between member and non-member samples. In contrast, FL countermeasures focus on protecting model updates from privacy violations by the central server or eavesdroppers, while also safeguarding the model from potential breaches by curious clients.

Aspect Centralized Learning Federated Learning FL server / client / eavesdropper Attacker role Model consumer Attack phase Inference phase Training phase Gradient Membership Model output Adversary knowledge (data) Model output Inference Intermediate computation Intermediate computation Attack Adversary knowledge (model) Target model Target model and historical versions Active strategy Data poisoning Model poisoning Membership Defender role Data / Model owner FL server / client Inference Training / Inference phase Training phase Defense phase Defense Protection core Model output Model update / Historical models

Table 1. A comparison of membership inference attacks and defenses in centralized and federated learning.

However, regarding FL-related MIAs, existing surveys provide only incomplete introductions and preliminary discussions [39, 50–58], lacking a comprehensive and systematic review. As an illustration in Table 2, earlier research [50, 52, 55, 56, 58] briefly outline privacy and security issues in the setting of FL, and [53, 54] emphasize both inference and poisoning attacks within the FL process. These surveys only mention a few early studies [17, 26] and omit more recently published research [33, 59]. Additionally, the most relevant article [39] provides a comprehensive summary of MIAs in the CL setting, while offering limited works related to FL [17, 26, 27, 29, 60, 61].

In this work, we provide a comprehensive survey of MIAs together with defense strategies on the whole FL process, as shown in Fig. 1. Starting with a unique taxonomy, we extensively review existing MIAs on FL through model updates and convergence trends in the whole FL training process. As for defenses against MIAs in the context of FL, we review four mitigation strategies used to protect exchanged updates and models, including partial sharing [62, 63], secure aggregation [64], noise perturbation [65, 66] and anomaly detection [67]. In the end, we also point out future research

111:4 Bai et al.

C	Year	Key Topic		Attacks <sup>1</sup>		Defenses <sup>2</sup>			
Survey	Теаг			A2	D1	D2	D3	D4	
[54]	2020	Privacy and robustness issues of FL	<b>/</b>	X	X	X	X	X	
[57]	2020	Privacy and security attacks in FL	/	X	X	<b>'</b>	/	X	
[56]	2021	FL vulnerabilities	nerabilities 🗸 🗡		<b>V</b>	/	/	/	
[55]	2021	Privacy and security threats in FL	Privacy and security threats in FL		X	/	/	1	
[50]	2021	FL concept and mechanism	echanism 🗸 🗶		X	<b>V</b>	<b>V</b>	X	
[53]	2022	Privacy and robustness in FL	FL 🗸 🗡		X	/	/	X	
[39]	2022	MIAs and defenses of CL	✓ X		X	X	/	X	
[58]	2023	Privacy and fairness in FL	in FL 🗸 🗡		X	<b>V</b>	1	X	
[40]	2023	Defenses against MIAs in CL	VX		X	X	/	X	
This work	-	MIAs and defenses in FL	<b>'</b>	<b>'</b>	<b>'</b>	~	<b>'</b>	<b>/</b>	

Table 2. Existing surveys about membership inference in federated learning.

<sup>&</sup>lt;sup>2</sup> D1: Partial sharing D2: Secure aggregation D3: Noise perturbation D4: Anomaly detection

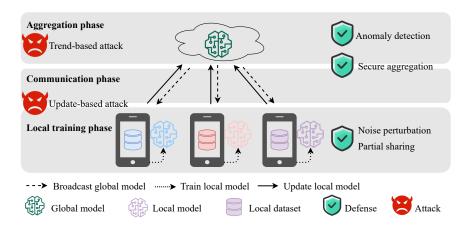


Fig. 1. Membership inference attacks and defenses in federated learning.

directions about MIAs and defenses in FL. The contributions of this research can be summarized as follows:

- We present a comprehensive review of MIAs in the FL setting by summarizing most studies in the literature. To the best of our knowledge, this paper is the first survey of MIAs and defenses in the FL domain.
- We identify MIA approaches in the context of FL, and offer both update-based and trend-based perspectives to guide this review. Based on our analysis, we present a unique taxonomy to summarize existing works on MIAs and discuss the differences from CL-related MIAs.
- We categorize existing countermeasures in FL on four approaches and analyze their advantages and limitations. Additionally, we offer a comparison of defenses in the CL setting from key viewpoints.
- We envision promising research directions for MIAs and defenses in this field.

The remaining sections of this survey are organized as follows: Section 2 first provides the preliminary background and briefly reviews CL-related MIAs and mitigation strategies. Following

<sup>&</sup>lt;sup>1</sup> A1: Update-based attack A2: Trend-based attack

this, Section 3 explores various threat models in FL. Next, Section 4 introduces and summarizes existing research on MIAs within the FL context. In Section 5, we describe the countermeasures available to defend against these attacks. Section 6 then highlights potential research directions, while Section 7concludes with final remarks.

## 2 PRELIMINARIES

This section first presents background knowledge regarding CL and FL. Then we offer a brief overview of MIA and defense studies in CL.

## 2.1 Centralized Learning

We consider a centralized setting in which training data are pooled in the server to train an ML model [37, 38], which leverages input-output pairs to train a non-linear function that predicts outcomes for new queries. Let  $D_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=0}^N$  and  $D_{te} = \{(\mathbf{x}_i, y_i)\}_{i=0}^M$  denote the training and test datasets, respectively, where  $\mathbf{x}_i \in \mathbb{R}^d$  denotes the d-dimension feature vector of the i-th sample labeled by  $y_i$ . The learning goal is to develop an ML model F that maps an input  $\mathbf{x}$  into  $F(\mathbf{x};\theta)$ , where  $\theta$  is the model parameters. To obtain the parameters  $\theta$ , a loss function  $\ell(\cdot,\cdot)$  is typically introduced to measure the discrepancy between the predicted output  $F(\mathbf{x};\theta)$  and the ground truth label y.

To obtain a high-quality ML model, we consider the expected value of the loss on the training dataset. A common approach for model optimization is empirical risk minimization [68], which minimizes the following objective function on  $D_{tr}$ :

$$L\left(D_{tr},\theta\right) = \frac{1}{N} \sum_{i=0}^{N} \ell\left(F\left(\mathbf{x}_{i};\theta\right), y_{i}\right) \tag{1}$$

Many training algorithms have been proposed to minimize this objective function [69–71], including stochastic gradient descent (SGD) [69], which updates the ML model in the t-th iteration as follows:

$$\theta^{t} = \theta^{t-1} - \eta \sum_{(\mathbf{x}_{i}, y_{i}) \in B} \nabla_{\theta} \ell \left( F\left(\mathbf{x}_{i}; \theta^{t-1}\right), y_{i} \right), \tag{2}$$

where B is a mini-batch of random training examples from  $D_{tr}$ ,  $\theta^t$  represents the model parameters in the t-th round, and  $\eta$  denotes the learning rate. The optimization process is terminated when the model converges to a local minimum, or the iteration reaches a preset number. The trained ML model often uses the accuracy on test dataset  $D_{te}$  to validate its performance.

## 2.2 Federated Learning

2.2.1 Brief Introduction of Federated Learning. FL is a collaborative ML paradigm that is widely used in various applications, such as smart healthcare [72, 73] and the Internet of Things [74, 75]. In contrast to CL, in which training data are collected and processed in a centralized location, FL involves training a shared global model using distributed data from different organizations or devices. Instead of transmitting the raw data to a central server, FL allows local data to remain stored locally and only the model parameters or gradients are exchanged, which addresses the data island problem and alleviates data privacy concerns.

The FL system involves two entities, clients and the central server. The clients (a.k.a., participants) possess training data and collaborate to train a shared model. According to the number of clients involved, there are two main FL types: cross-silo and cross-device FL [51]. Cross-silo FL involves a relatively small number of clients, typically organizations or data centers with a significant amount of data [76, 77], while cross-device setting involves a large number of clients with small amounts of data, such as mobile devices [74, 75]. The central server is used to aggregate model updates from

111:6 Bai et al.

clients without accessing their local data. In the cross-silo setting, a client can be selected as the central server to perform aggregation and update the global model. In contrast, a powerful server is often introduced to manage model aggregation effectively in cross-device FL.

FL typically involves training a joint global model based on distributed local data through three phases: training, communication, and aggregation phase. There are K clients  $\{c_1, c_2, ..., c_K\}$  that involve in FL and each client c owns a local dataset  $D_c$ . We use  $\theta_s$  and  $\theta_c$  to denote the global and local model, respectively. To train a global model collaboratively, the central server first initializes the global model  $\theta_s$  (e.g., model weights and hyperparameters) and broadcasts it to clients, and then the FL process carries out as follows:

- 1 Local training phase. Each selected client c receives the global model  $\theta_s$  and training settings. Then, the client fine-tunes it on the local dataset  $D_c$ .
- 2 Communication phase. For each selected client, it uploads the latest version of local model  $\theta_c$  to the central server, while the central server broadcasts the global model to selected clients after the aggregation phase.
- 3 Aggregation phase. The central server aggregates all the model updates from the selected clients and renews the global model  $\theta_s$ .

Three phases are repeated until the loss value of the global model converges on a validation dataset, resulting in a well-trained model.

2.2.2 Categorization of Federated Learning. Based on the distribution of data over the sample and feature spaces, we can broadly categorize FL into horizontal FL, vertical FL, and federated transfer learning [50–52].

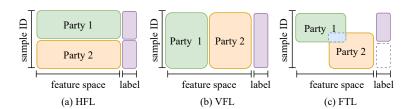


Fig. 2. Three categories of federated learning.

Horizontal Federated Learning (HFL) refers to scenarios where multiple parties possess data samples with the same features but unique sample IDs, as shown in Fig.2(a). For instance, different hospitals may possess patient records that share the same feature space but have different patient IDs. HFL is the most popular FL category in the literature [12, 78], and local models commonly share the same architecture as the global model. As such, the global model can be simply aggregated from local models.

Vertical Federated Learning (VFL) refers to scenarios where multiple parties in an FL system hold data samples with identical sample IDs but different feature spaces, as shown in Fig.2(b). A typical case for VFL is when an E-commerce company and a local bank collaborate to train a personalized loan model based on online shopping records and credit situation [52]. Typically, only one participant can access the labels, and sample alignment is performed before training a joint model across multiple clients [55].

**Federated Transfer Learning (FTL)** pertains to the situation where clients' datasets contain varying ID spaces and feature spaces [79], as shown in Fig.2(c). Inspired by transfer learning, FTL allows knowledge to be shared without compromising data privacy, enabling the transferability of

complementary knowledge by exploiting source domain knowledge to train an FL model for the target domain [80, 81].

## 2.3 Membership Inference Attacks in Centralized Learning

MIAs aim to infer whether an example was used to train an ML model [22]. Given a query example  $\mathbf{x}$  and a target model  $F(\theta)$ , an MIA algorithm  $\mathcal{A}$  infers the membership status m of  $\mathbf{x}$ . We formulate an MIA algorithm as a binary classification task:

$$m = \mathcal{A}(\mathbf{x}, F(\theta)) = \begin{cases} 0, & \mathbf{x} \notin D_{tr} \\ 1, & \mathbf{x} \in D_{tr}, \end{cases}$$
(3)

where the membership status is 1 if the adversary infers that the input  $\mathbf{x}$  was used to train the target model and 0 otherwise. MIAs in CL occur in either *white-box* scenarios, where attackers access model details (e.g., architecture and parameters), or *black-box* scenarios, where only model outputs are observed. In white-box settings, model parameters  $\theta$  are observable, but not in black-box settings [29]. Based on the construction of the attack model, MIAs can be categorized into classifier-based attacks and metric-based attacks.

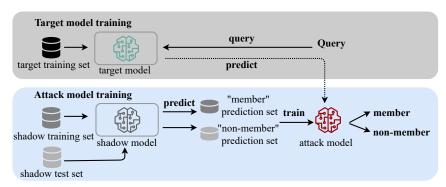


Fig. 3. Overview of the shadow training scheme in CL.

In a classifier-based MIA, a binary classifier attack model is constructed to determine membership information. Shokri et al. [22] conduct the pioneering study on this approach, utilizing the *shadow training* technique to build an attack model that outperforms random guessing (i.e., 0.5 probability of inferring membership). Fig.3 visualizes how shadow training can be used to construct a classifier-based MIA. Assuming the attacker has auxiliary knowledge, they can collect a shadow dataset whose distribution is similar to the target dataset. Next, the attacker trains one or multiple shadow models with the same architecture as the target model and collects prediction vectors from the shadow dataset. Finally, the attacker uses these prediction data to train an attack model capable of inferring membership status based on the output of a query example. Later studies extended this technique to relaxed assumptions [43] and various model architectures [31, 82–84].

A metric-based MIA is more straightforward and has less computational cost than classifier-based attacks. It deduces membership information for data records by calculating metrics on their prediction vectors. The calculated metrics are then compared with a preset threshold to determine the membership status of a data record [39]. Several metric options are available, such as prediction correctness [85], prediction loss [85], confidence score [43], and confidence entropy [43, 86]. Based on these intuitive approaches, advanced approaches have been proposed to calibrate calculated metrics or estimate the prediction distribution by introducing more shadow models [87, 88] to decrease the false positive rate of MIAs.

111:8 Bai et al.

## 2.4 Membership Inference Defenses in Centralized Learning

Numerous studies have explored membership inference defense techniques in the context of CL, primarily including regularization, knowledge distillation, differential privacy, and output perturbation, designed to thwart membership privacy violations. In this subsection, we touch upon these strategies. For a more comprehensive understanding, readers are encouraged to delve into the detailed introduction provided in [40].

Regularization and knowledge distillation effectively defend against MIAs by diminishing the overfitting level of ML models. Existing empirical and theoretical research has underscored the link between the leakage of membership information and the extent of model overfitting [22, 85, 89]. Motivated by this observation, regularization techniques strive to reduce the generalization gap, thus diminishing the vulnerability to MIAs. This category encompasses a range of approaches, including L2-norm regularization, dropout techniques [43], and model stacking methods [15]. Likewise, knowledge distillation techniques [90] encourage a smaller student model to learn from the outputs of a teacher model, rather than relying solely on the original training labels. This approach enhances overall model generalization while simultaneously thwarting attackers from inferring sensitive member data.

Differential privacy serves as a potent defense mechanism in the CL setting. This technique operates by introducing carefully calibrated noise into model gradients, thereby guaranteeing that the presence or absence of individual data points remains indiscernible. While this method inherently guards against MIAs as its theoretical definition, it invariably leads to substantial utility loss when implemented [22, 91].

In contrast to prior defense techniques that alter input data or the training process, output perturbation represents a post-processing method employed to align the model's output across training and test samples. Typically integrated during the inference phase, this approach is efficient against black-box MIAs in CL. It entails the partial disclosure of confidence scores, including the top-K confidence score [22], the associated predicted label [45, 46], and noisy confidence scores [92].

## 3 THREAT MODEL AND ATTACK TAXONOMY

#### 3.1 Threat Model

In this subsection, we overview the threat models of MIAs in the FL setting from three perspectives: the adversary's goal, role, and strategy, as shown in Fig.4. In addition, we discuss and compare the adversarial knowledge that an MIA attacker can access in each threat model.



Fig. 4. Categorization of threat models.

# 3.2 Adversary's Goal.

In the existing literature on MIAs in FL, the granularity of the target can be either *record-level* or *source-level*. As defined in Eq.(3), record-level attacks represent a privacy breach of individual data items. For instance, in a disease-prediction model trained through FL, record-level MIAs infer a patient who likely suffers from disease by identifying the presence of the patient's clinical record in the *entire training dataset*.

On the other hand, source-level attacks refine the goal of MIAs from the entire training dataset  $D_{tr}$  to a *specific client's training dataset*  $D_c$ . When a data point is a member in the context of FL, source-level MIAs can further identify the specific client to which it belongs. Formally, the training dataset consists of multiple local datasets  $D_{tr} = \{D_c | c = 1, 2, ..., K\}$ , where  $D_c$  is the local dataset of client c. Source-level MIAs aim to trace the source of a training member [29, 61, 93], formally defined as follows:

$$m = \mathcal{A}(\mathbf{x}, F(\theta)) = \begin{cases} 0, & \mathbf{x} \notin D_{tr} \\ c, & \mathbf{x} \in D_{c} \end{cases}$$
(4)

Source-level attacks are a natural extension of record-level attacks and cause greater privacy concerns [29]. For example, suppose several hospitals collaborate to train a shared global model to predict COVID-19 diagnosis. Record-level MIAs can reveal a patient's identity when her record is used as the training data. By contrast, source-level attacks can identify the hospitals where these patients were treated, which can further leak more information, e.g., patients' addresses.

## 3.3 Adversary's Role.

ML models in the FL setting are vulnerable to MIAs launched by either *insiders* or *outsiders*. Insiders, such as curious clients or the central server, may perform MIAs to uncover sensitive information about other participants. Outsiders, like eavesdroppers, can exploit intercepted model updates to infer membership privacy related to either local or global models in the FL system.

As insiders, the FL server and clients can legally access local and global models during training. In particular, a server is more powerful than the clients because it can observe each local model from participants and manipulate the aggregated results. Conversely, clients can only observe the aggregated model, making it challenging to perform source-level attacks without auxiliary knowledge. While as outsiders, eavesdropping attackers can intercept the communication messages between the FL server and a client, including the global model updates of the server to a client, the local model updates of a client to the server, or both.

In general, insider adversaries pose a more significant threat than outsider adversaries as they can access more information and manipulate the training process [53]. For instance, an honest-but-curious FL server may access the local models of each participant during the aggregation process and use this information to identify whether a query sample belongs to a particular participant. A malicious FL participant may even upload a poisoned model to the FL server to enhance sensitive information to leak [26]. By comparison, eavesdropping attackers can only passively access communication messages but cannot modify them.

# 3.4 Adversary's Strategy.

An adversary can infer the membership information in the context of FL through *passive* or *active* strategies.

Regarding a passive MIA, an adversary only observes and exploits the learning model without directly interfering, relying solely on the data collected during the normal operation of the FL system. For example, any client in an FL system can exploit the information they collect during training to deduce private data about other participants. Such an attack is challenging to detect

111:10 Bai et al.

because the adversary does not affect the target model during the training phase. In contrast, an insider attacker can launch an active attack to boost attack performance by altering the learning model or data. For instance, a malicious client can actively push a target model far away from the optimum to inspect the response of a training member [26]. Active inference attacks are easier to mount in FL than CL, as any participant or the server can modify the training data or manipulate the model parameters.

## 3.5 Attack Taxonomy

In this survey, we classify existing research on MIAs in FL into two categories: (1) update-based attacks, which leverage one or more historical versions of the target model to infer membership information; and (2) trend-based attacks, which analyze the trajectory of specific indicators to determine membership status. Given their different threat models, Table 3 provides a summary of the surveyed attacks.

Table 3. Taxonomy of membership inference attacks on federated learning.

Attack	Approach Reference		G	sary's		ersary's Role		ersary's rategy
			Record-level	Source-level	$Inside_r$	$O_{utsider}$	$P_{assive}$	$A_{ctiv_{\mathbf{e}}}$
	Update-based MIAs							
Model gradient-based	Original gradient	Nasr et al. [26]	<b>√</b>		✓		<b>√</b>	$\checkmark$
		Gupta et al. [31]	✓		✓		✓	
		Lu et al. [94]	✓		✓		✓	
	Gradient difference	Melis et al. [17]	✓		✓		✓	
		Li et al. [33]	✓			$\checkmark$	✓	
		Zhu et al. [59]	✓		✓		✓	
Single model-based	Shadow training	Liu et al. [32]	✓		✓		✓	
		Pustozerova et al. [28]	✓		✓		✓	
		Zhang et al. [27]		$\checkmark$	✓		✓	
		Chen et al. [61]		$\checkmark$	✓		✓	
		Zhao et al. [93]	✓		✓		✓	
		Luqman et al. [95]	✓		✓			$\checkmark$
		Banerjee et al. [96]	✓			$\checkmark$	✓	
		Truex et al. [97]	✓		✓		✓	
		Yuan et al. [98]	✓		✓		✓	
	Structure modifying	Pichler et al. [34]	✓		✓	$\checkmark$		$\checkmark$
		Nguyen et al. [99]	✓		✓			✓
	Trend-based MIAs							
Model output-based	Prediction trajectory	Zari et al. [35]	✓			$\checkmark$	✓	
		Gu et al. [30]	✓		✓		✓	$\checkmark$
		Zhang et al. [100]	✓		✓			$\checkmark$
		Liu et al. [101]	✓		✓		✓	
	Loss trajectory	Hu et al. [29]		$\checkmark$	✓		✓	
		Suri et al. [36]	✓	$\checkmark$	✓		✓	
		Zhu et al. [59]	✓		✓		✓	
Model parameter-based	Bias trajectory	Zhang et al. [102]	✓		✓		✓	$\checkmark$

#### 4 MEMBERSHIP INFERENCE ATTACKS IN FEDERATED LEARNING

In this section, we provide a detailed examination of specific attacks within the FL context from both update-based and trend-based perspectives.

# 4.1 Update-based Membership Inference Attacks

Although FL prevents access to the raw training data, it is still vulnerable to MIAs owning to the gradient/weight exchanged. An update-based MIA directly leverages exchanged information to develop an attack model that distinguishes between members and non-members. According to the exchanged information, these attacks can be further divided into those based on *model gradient* and *model parameter*. The former directly exploits original gradients or gradient differences as the input of attack models, whereas the latter extends the shadow training method to FL scenarios or uses a specialized structure of the target model to infer membership information.

4.1.1 Model Gradient-based MIAs. MIAs of this type in FL treat model gradients as part of attack feature vectors [26, 31, 94], or compare the gradients between rounds [17, 33] to infer the membership status of a query example.

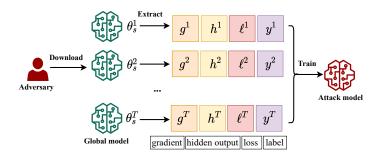


Fig. 5. Overview of update-based MIAs by using original gradients and other intermediate features.

Infer Membership via Original Gradient. This approach commonly regards the original gradients exchanged as one of the attack features and trains an attack model to discern the membership status. Nasr et al. [26] develop a pioneering inference attack against the FedAvg algorithm [12], which can infer leaked private information in the ML model by exploiting gradients and intermediate outputs in either an active or a passive manner. A participant or server can gather original gradients, hidden layer outputs, loss values, and ground truth labels over multiple iterations and exploit them to construct an attack model in a passive strategy, as depicted in Fig.5. Moreover, this work also develops an active attack strategy that causes victim participants to divulge more information through the *gradient ascent algorithm*. Specifically, for a query example  $\mathbf{x}$ , an active attack performs gradient ascent as shown in Eq.(5), where  $\eta$  represents the local learning rate, and  $L(\mathbf{x}, \theta)$  denotes the loss on  $\mathbf{x}$ .

$$\theta \leftarrow \theta + \eta \frac{\partial L\left(\mathbf{x}, \theta\right)}{\partial \theta}.$$
 (5)

Contrary to conventional SGD algorithms that reduce the loss of a training sample, an active attacker intentionally increases the loss on  $\mathbf{x}$  to widen the gap between the member and non-member data points. If a query example is a member of the training data, the SGD algorithm decreases its loss after an honest participant updates the model. As such, a non-member record maintains a high loss due to the absence of additional modifications. Furthermore, when the server is a malicious attacker, the attacker can isolate the victim participant to obtain a local view of the

111:12 Bai et al.

target model and expose more private information. This study showcases how an adversary can improve attack performance in a white-box scenario. We conjecture the rationale behind it is that gradients confer an advantage for MIAs since they reveal more detailed information about the query example. The gradients in a fully connected layer represent the inner products of the error from the next layer and the features at that layer [17].

The work of Nasr et al. [26] has been extended to other applications [31, 94]. A similar attack approach [94] is used to compare FL and coreset-based learning [94, 103] in terms of data privacy. Specifically, the passive MIA method is used in FL, and a new attack strategy for coreset is established by K-means algorithm. The results show that FL is preferable over coreset for privacy protection with high accuracy. However, if a significant loss of model accuracy is tolerable, coreset can achieve privacy protection with less computation cost. Gupta et al. [31] develop MIAs against deep regression that predicts brain age in the FL setting. This approach leverages gradients, activations, and predictions to conduct MIAs from the viewpoint of a participant. The authors show that privacy breaches are more eminent when local data is skewed or non-IID among participants.

Although MIAs discussed above [26, 31, 94] can be applied in various networks and scenarios, they have limitations in terms of assumptions and efficiency. For instance, they require access to partial member data from victim clients, which may be challenging in scenarios with strict privacy constraints (e.g., medical area). Additionally, it is time-consuming to construct an attack model as it requires a significant number of feature vectors from various iterations.

Infer Membership via Gradient Difference. To address the limitations of using original gradient, it is possible to directly infer the membership information by evaluating the gradient changes between consecutive iterations, given the distinction between members and non-members in the gradient distribution. Melis et al. [17] introduce a novel MIA based on gradient difference in which the embedding layer leaks the membership status of training samples. The core idea is that non-zero gradients in the embedding layer disclose which samples are trained in a batch. Specifically, an honest-but-curious participant collects consecutive snapshots of the global model in the (t-1)-th and t-th rounds, and calculates the difference between them as  $\Delta \theta^t = \theta_s^t - \theta_s^{t-1} = \sum_c \Delta \theta_c^t$ . The model update contributed by other participants can be represented as  $\Delta \theta^t - \Delta \theta_{adv}^t$ , where  $\Delta \theta_{adv}^t$  denotes the update of the adversary in the t-th iteration. The proposed technique effectively detects the presence of location or text samples with a small batch size in a two-party FL setting. However, as the batch size increases, the false positive ratio significantly increases as it becomes more challenging to identify the exact training representation for a large set of word candidates in a bag-of-words format. Moreover, this method is only applicable to neural networks with embedding layers and cannot be applied to deep learning models that use numeric data (e.g., tabular or image).

Li et al. [33] propose two passive MIAs called the gradient-diff attack and cosine attack. In these approaches, the adversary computes the gradient difference in consecutive rounds to deduce the membership status. The gradient-diff attack is based on gradient orthogonality, through Eq.(6) to indicate the membership status of  $\mathbf{x}$ , where  $\Delta \theta_c^{t+1}$  denotes the local model update of client c in the (t+1)-th iteration. If Eq.(6) holds,  $\mathbf{x}$  is a training record in the private set  $D_c$ .

$$\left\|\Delta\theta_c^{t+1}\right\|_2^2 - \left\|\Delta\theta_c^{t+1} - \sum_{y \in Y} \nabla_{\theta} \ell\left(F\left(\mathbf{x}; \theta_s^t\right), y\right)_2^2\right\| > 0.$$
 (6)

Moreover, the authors find a clear disparity between the distributions of the cosine similarity for member instances and non-member instances. Although both distributions resemble Gaussian distributions, their varying averages reveal distinct membership characteristics. As such, they develop a cosine attack to deduce membership information by measuring the angle between these two vectors:

$$\sum_{y \in Y} \operatorname{sgn}\left(\operatorname{cosim}\left(\nabla_{\theta} \ell\left(F\left(\mathbf{x}; \theta_{s}^{t}\right), y\right), \Delta \theta_{c}^{t+1}\right) \geq \gamma\right),\tag{7}$$

where  $sgn(\cdot)$  is an indicator function,  $cosim(\cdot, \cdot)$  denotes the cosine similarity, and  $\gamma$  is a preset threshold based on non-member data. Their MIAs are still effective and robust under privacy protection mechanisms [104, 105]. Zhu et al. [59] explore this idea further by extending it across multiple communication rounds and integrating models from non-target clients, boosting attack effectiveness significantly.

4.1.2 Single Model-based MIAs. MIAs in this category extract information about membership by utilizing either a historical global model or a local model in the FL training. Most of these attacks adapt the shadow training method from CL for conducting MIAs [28, 98], often enhanced through data augmentation [27, 32]. Additionally, manipulating a specific structure within the target model can also facilitate membership inference in FL [34, 99].

Infer Membership via Shadow Training. Inspired by the shadow training technique in CL, this approach considers the global or local model as the target model. However, in contrast to the CL scenario, it can be implemented more easily as any participant can naturally access the target model and auxiliary data. Pustozerova et al. [28] develop an MIA for a sequential FL framework that exposes an individual local model to the subsequent participant immediately after training on private datasets, allowing others to infer information from the received model. Assuming the presence of an auxiliary dataset, an attacker (e.g., one of the participants) builds shadow models, collects attack features, and develops an attack model to launch an inference attack on the victim model. Luqman et al. [95] investigate CL-related MIAs based on shadow training in the peer-to-peer FL setting and reveal that membership leakage intensifies when colluding adversaries are involved. FD-Leaks, an MIA proposed by [106], is tailored for federated distillation learning that involves clients exchanging model outputs on a public dataset. Unlike the shadow training approach in CL, this method treats the attacker's model as the shadow model and avoids retraining.

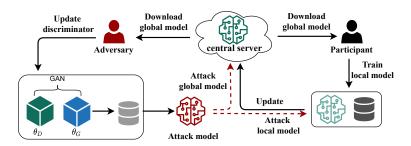


Fig. 6. Overview of updated-based MIAs enhanced by data augmentation.

Shadow training approaches can be enhanced by focusing on the construction of both shadow datasets and attack datasets [27, 32, 61, 93, 96]. Data augmentation is implemented to boost the shadow dataset in the FL setting. As illustrated in Fig.6, an attacker uses a shared global model as the discriminator of a GAN to generate diverse data and update it during the learning process. After generating sufficient attack data, the attacker trains a binary attack model by shadow training approach. Zhang et al. [27] first employ GANs to enhance the MIAs launched by insiders. Subsequently, Liu et al. [32] develop an MIA in which an eavesdropper behaves as a regular participant but has no knowledge of the private training dataset and attempts to attack the global model in FL. Assuming that clients' labels are non-overlapping among different participants, source-level MIAs can be launched by comparing the query example's prediction result and label distribution among participants [61, 93]. In terms of attack data construction, MIA-BAD [96] improves shadow training methods by introducing a batch-wise attack dataset, inspired by the ensembling phenomenon [107].

111:14 Bai et al.

As for attacks beyond classification tasks, Yuan et al. [98] investigate an MIA that utilizes the shadow training approach against federated recommender systems (FedRecs), in which an honest-but-curious server aims to determine items interacted by a user based on uploaded parameters. Attacks on classification tasks cannot be applied to FedRecs because most attacks aim to infer the existence of a query example, which is ineffective for FedRecs as only positive items (i.e., items interacted with by a user) disclose private information. Specifically, the attack process involves training a shadow recommender model by randomly assigning ratings to the relevant items, then iteratively selecting the closest items, and finally retraining the shadow model until it reaches the preset number of guessed items. This attack efficiently infers user interaction information for Fed-NCF and Fed-LightGCN frameworks [108, 109]. Additionally, the authors also observe that membership information leakage could increase with more auxiliary knowledge, such as popular items.

Infer Membership via Structure Modifying. In such attacks, a malicious server may actively manipulate model structures to infer membership information. Pichler et al. [34] study an active attack by modifying the network of the client model. This approach uses the rectified linear unit (ReLU) property that the derivative of an output with respect to the parameters is zero when the output is negative. Specifically, a malicious server embeds a network module equipped with ReLU activations into the target model, followed by configuring its parameters to enable activation by training members. When encountering an unseen query example, the parameters within the architecture remain unaltered. Consequently, the attacker determines the membership status by comparing the parameter changes with a predetermined threshold. A similar idea is also explored in [99], where a malicious server carefully crafts and embeds malicious parameters into a specific neuron, which a member sample can only activate. The dishonest server can infer membership details by examining the neuron's gradient during the learning process. These attacks are simple yet effective, with the malicious strategy applicable to networks utilizing ReLU activations.

Table 4. Summary of update-based membership inference attacks on federated learning.

Year Reference	Task	Technique	Comparison	Summary
2019 Nasr et al. [26]	Classification	Intermediate output	[22]	Use original gradients
2021 Gupta et al. [31]	Regression	Intermediate output	-	$\ominus$ Require large feature vectors
2020 Lu et al. [94]	Classification	Intermediate output	-	$\ominus$ Require auxiliary member data
2019 Melis et al. [17]	Embedding	Non-zero gradient	-	Use gradient differences
2023 Li et al. [33]	Classification	Gradient orthogonality	[110]	⊕ Require fewer feature vectors
2024 Zhu et al. [59]	Classification	Gradient similarity	[26, 33, 85]	⊕ Low computation
2022 Liu et al. [32]	Classification	Data reconstruction	-	
2022 Pustozerova et al. [28	] Classification	Sequential update	Random guessing	5
2020 Zhang et al. [27]	Classification	Data augmentation	Random guessing	5
2020 Chen et al. [61]	Classification	Non-overlapping label	[26]	Extend shadow training to FL
2021 Zhao et al. [93]	Classification	Non-overlapping label	-	⊕ Launched by any role
2023 Luqman et al. [95]	Classification	Peer-to-peer update	-	$\oplus$ Enhanced by data augmentation
2024 Banerjee et al. [96]	Classification	Batch-wise feature	[22]	$\ominus$ Underexplored information
2018 Truex et al. [97]	Classification	Decision boundary	Random guessing	5
2022 V 1 [00]	D	. Earl adding a slares as	K-means	
2023 Yuan et al. [98]	Recommendation	Embedding relevance	Randow guessing	
2022 Pichler et al. [34]	Classification	ReLU activation	-	Modify the target model
2022 Nguyen et al. [99]	Classification	Poisoned neuron	-	⊕ Low computation
				⊖ Require specialized networks

<sup>⊕:</sup> Advantages; ⊖: Disadvantages

To summary, it is intuitive to exploit original gradients to build an inference model. However, high computational and memory resources are required when collecting all intermediate outputs of target models. In addition, these original gradient-based MIAs require an adversary to access partial member data to build a high-quality attack model [26, 31, 94]. In contrast, gradient difference techniques are proposed to tackle these challenges efficiently and practically. Additionally, attacks based on model parameters mainly treat one of the snapshots as the target model and conduct MIAs to infer the private information of FL clients. Given the white-box scenario in FL, the shadow training approach is practically implemented and boosted with the help of data augmentation via generative models. On the other hand, the structure modifying approach [34, 99] is limited to a malicious central server and linear layers, unsuitable for heterogeneous FL settings and complex networks. We present additional details of trend-based attacks in Table 4, including the year of application, the application domain, the specific methods employed, attacks used for comparison with the proposed approach, and a summary of each kind for readers' reference.

# 4.2 Trend-based Membership Inference Attacks

Trend-based MIAs examine the evolution of an indicator associated with membership status to determine whether a record is a member during the learning process. Such MIAs typically collect the historical models, extract the indicator information, and make decisions by comparing the indicator distributions between members and non-members, as shown in Fig.7. According to the indicator knowledge used, trend-based attacks can be categorized into those based on *prediction score* and *prediction loss*.

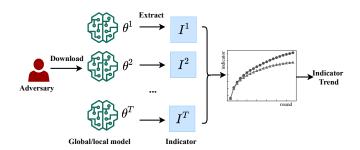


Fig. 7. Overview of trend-based MIAs.

4.2.1 Model Output-based MIAs. For an ML model, the prediction score (a.k.a, confidence) and predication loss in the private data of participants often increases faster than that observed from testing data as the model converges, partially due to overfitting and memorization. Inspired by this phenomenon, the difference in model output changes between training and test data across iterations is used when inferring membership information [30, 35, 100].

Infer Membership via Prediction Trajectory. This approach uses the trend in confidence predicted by multiple models to differentiate between member and non-member samples. A passive MIA proposed by [35] exploits a sequence of prediction scores generated by a local model to deduce the membership status. Specifically, to infer the privacy of a specific client c, a curious eavesdropper gathers many versions of local models exchanged between c and the server, and then calculates the prediction probability of the ground truth label of a query example  $\mathbf{x}$ , namely,  $\left\{F_y\left(\mathbf{x};\theta_c^t\right)\right\}_{t=1}^T$ , where T is the number of rounds. Then, the adversary constructs a fully convolutional attack model to classify these series and learn the difference between the training and test data. This approach

111:16 Bai et al.

simplifies the input requirements of the method proposed by [26], as the input vector of attack models only consists of a single number in a round, and thus, a comparative attack performance can be achieved with lower computational resources and memory.

Similarly, Gu et al. [30] propose confidence-series-based MIAs (CS-MIAs) that use advanced confidence metrics and active attack strategies. Different from [35], CS-MIAs use a novel confidence metric called *modified prediction entropy* [86] to calculate the confidence series. In addition, CS-MIAs allow a global adversary (i.e., the server) to extract more private information by active participation and selection. The global adversary fine-tunes the global model using auxiliary data and submits updates similar to a regular participant in each iteration to address the problem of lack of training data when building the attack model. This process makes the confidence scores on the shadow training data similar to those on the target training data. Furthermore, the adversary deliberately selects the target client in each round instead of random selection to force the victim participant to leak more private information. Experimental results demonstrate that CS-MIAs outperform existing state-of-the-art black-box MIAs [26, 43, 86].

Zhang et al. [100] explore the poisoning MIA (PMIA) in FL, where poisoning attacks are employed to enhance the effectiveness of MIAs. This approach injects malicious gradients to improve ML models by triggering victim clients to repair manipulations, leading to disclosing membership information to the attacker. During the training phase, this approach formulates an optimization problem to maximize the loss of victim training data while evading detection from Byzantine-robust aggregation mechanisms by introducing a "cover" dataset to conceal the adversary's update. Subsequently, membership information is inferred using prediction correctness or trend analysis. It is noteworthy that the prediction trends behind PMIA differ from the aforementioned works. Specifically, if a client observes an incorrect prediction on a training sample after an attack round, the subsequent training rounds are expected to reach a correct prediction.

Rather than employing prediction series directly, Liu et al. [101] presented a novel attack technique based on the temporal evolution of adversarial robustness when an adversary's access is restricted solely to prediction labels. Their approach stems from the conspicuous disparities observed in the convergence patterns of adversarial robustness between training and test data.

Infer Membership via Loss Trajectory. This approach typically examines the change in the loss of samples throughout the training process to infer the membership status. Suri et al. [36] propose a subject-level MIA that aims to infer the privacy of a particular individual's (a.k.a, subject's) data in the cross-silo FL setting. Subject-level privacy is equivalent to source-level privacy in the cross-device FL setting, in which FL has a one-to-one mapping between data subjects (each individual owning a participating subject). This equivalence no longer works in the cross-silo setting when an individual's data is spread across several federation users or organizations [36, 111]. By assuming an auxiliary dataset about inferred subject distribution, the authors develop a loss-across-round attack to infer the subject membership. The loss-across-round attack from a participant or server exploits the change in the loss values as the training rounds progress to deduce the subject membership status. To this end, the attacker records loss values of the subject's dataset  $D_s$  across each training round t, counts the number of training rounds where the loss decreases, and finally compares the final value  $\mathcal L$  with a threshold by Eq.(8). In addition to utilizing multi-round model updates from the target client, Zhu et al. [59] enhance the effectiveness of MIAs by incorporating models from

non-target clients, particularly in scenarios with homogeneous data distributions.

$$\mathcal{L}_{t} = \sum_{(\mathbf{x}, y) \in D_{s}} \ell_{t}(\mathbf{x}, y)$$

$$\mathcal{L} = \sum_{t=1}^{r} \mathbb{I} \left[ \mathcal{L}_{t} < \mathcal{L}_{t-1} \right]$$
(8)

Hue et al. [29] targets source-level privacy and leverages the prediction loss to determine the source of a training sample from a server perspective. This is based on the fact that the probability of an example being a member depends on its loss [89]. SIA allows an honest-but-curious server to launch a source-level attack in the FL setting by comparing the loss across local models. The source of that example is the participant whose local model yields the smallest loss. The success of SIA relies on the generalization of local models, which highlights the privacy risks arising from the non-IID phenomenon in FL. Furthermore, SIA provides an alternative approach for an FL server that conducts inference attacks against a specific participant, that is, it can leverage out-of-the-box MIAs on CL, such as metric-based attacks [85, 86, 112], to easily breach the privacy of participants' local data.

4.2.2 Model Parameter-based MIAs. The optimization of an ML model aims to reduce the difference between its output and the true output on the training set, resulting in adjustments to the model's parameters. MIAs based on model parameter exploit these changes by identifying distinct patterns in how the model behaves for member versus non-member samples.

Infer Membership via Bias Trajectory. Zhang et al. [102] are the first to leverage changes in model bias to determine membership in federated learning (FL). Their approach was motivated by the observation that non-member samples induce significant changes in model parameters, resulting in a more pronounced shift in bias. This work specifically examines the bias values of the final layer and incorporates feature amplification through an exponential function. By gathering the model over multiple epochs, it calculates the bias changes and achieves membership inference. Empirical evidence suggests that these bias differences incur minimal overhead while providing effective features for MIAs.

To conclude, trend-based attacks exploit the trajectory of model output or parameter to deduce membership status and exhibit the following advantages. First, instead of collecting a large feature vector like [26, 31, 94], they gather only a single value in a round. Moreover, these approaches can make decisions about membership information by comparing the indicator with a preset threshold or observing the trend direction of the indicator without developing an attack classifier model. These advantages lower the computation and memory resources cost and the implementation complexity. We present additional details of trend-based attacks in Table 5, including the year of application, the application domain, the specific methods employed, attacks used for comparison with the proposed approach, and a summary for each kind.

## 4.3 Comparison to Attacks in Centralized Learning

In this section, we examine the existing update-based and trend-based MIAs in the context of FL. Update-based approaches focus on extracting membership information from model updates exchanged between clients and the server, including model gradients and historical models throughout the federated learning process. In contrast, trend-based approaches analyze the evolving patterns of the training data to conduct MIAs, capitalizing on the observation that member samples exhibit distinct behaviors compared to non-member samples from an indicator perspective as the learning process progresses.

111:18 Bai et al.

Table 5. Summary of trend-based membership inference attacks on federated learning.

Year Reference	Task	Technique	Comparison	Summary
2021 Zari et al. [35] 2022 Gu et al. [30] 2023 Zhang et al. [100] 2023 Liu et al. [101]	Classificatio	n Predication sequence n Modified predication sequence n Poisoning gradient n Adversarial robustness	[26] [26] [17, 26] [26, 46]	Use prediction trajectory  ⊕ Require fewer feature vectors  ⊖ Require ground truth labels
2021 Hu et al. [29] 2022 Suri et al. [36] 2024 Zhu et al. [59]	Classificatio	n Loss comparison n Multi-round loss n Multi-round-client loss		Use loss trajectory  ⊕ Require fewer feature vectors  ⊖ Require ground truth labels
2023 Zhang et al. [102]	] Classificatio	n Multi-round bias	[26, 112, 113]	Use parameter changes  ⊕ Require fewer feature vectors  ⊕ Without label requirement

 $<sup>\</sup>oplus$ : Advantages;  $\ominus$ : Disadvantages

Table 6. A comparison of attack approaches in centralized and federated learning.

Setting	Туре	Approach	Phase	Attack Feature	Unique <sup>1</sup>	
	Classified-based	Shadow training		Model output		
	Classifieu-baseu	Shadow training		Intermediate output		
CL		Prediction correctness	Inference	Model output	•	
CL	Metric-based	Prediction loss	interence	Loss of model output	0	
	Wietric-baseu	Prediction confidence		Model output	•	
		Prediction entropy		Model output	•	
				Model output		
	Update-based	Original gradient	Aggregation	Gradient	•	
				Intermediate output		
		Gradient difference	Aggregation	Gradients within		
		Gradient dinerence	Communication	consecutive iterations	•	
FL		Shadow training	Local training	Global model	0	
		Structure modifying	Local training	Global model	•	
		Prediction trajectory	Local training	Model output	0	
	Trend-based	Loss trajectory	Local training	within several iterations	U	
	11chu-baseu	Bias trajectory	Local training	Model bias		
		Dias trajectory	Local training	within several iterations		

 $<sup>^{1}</sup>$   $\bullet$ : unique attack  $\bullet$ : partially unique attack  $\circ$ : common attack

For a comprehensive understanding of MIAs in the FL context, we compare them with attacks relevant to CL, encompassing both classifier-based and metric-based methods mentioned in Section 2.3. Table 6 highlights key differences between FL-related MIAs and those in CL, focusing on two main aspects: the phase in the model's lifecycle when the attack occurs, and the adversarial knowledge leveraged to infer membership. These differences give rise to innovative attack methodologies in the FL setting. With access to the internal details of the target model, an attacker can leverage additional adversarial knowledge to enhance the effectiveness of inference attacks, employing the original gradient or gradient difference methods. However, these techniques are generally not applicable to most black-box MIAs in CL. Moreover, in FL, adversaries often have access to historical versions of the target model, enabling them to track the trajectory of query examples over time. This can greatly increase the risk of membership information leakage and supports trend-based attacks in FL, whereas such information is typically less available in CL scenarios.

Additionally, source-level MIAs are distinct in the FL setting because of its collaborative training strategy. From the perspective of the central server, this is intuitively equivalent to using record-level MIAs to sequentially infer the membership status of each local model [27, 61]. However, for a more effective and efficient attack, it is essential to consider the local models from all clients simultaneously. Hu et al. [29] compare the loss values across clients and identify the client with the lowest loss as having the source dataset. Zhang et al. [102] leverage the bias differences in the final layer and analyze multiple model epochs to identify the member source. They assign the data to the participant exhibiting the smallest change in bias. Current MIA research primarily emphasizes record-level privacy risks, as the membership leakage of an individual sample is central to private information exposure. Additionally, most record-level attacks can be extended to source-level attacks [102].

## 5 MEMBERSHIP INFERENCE DEFENSES IN FEDERATED LEARNING

Various defenses have been proposed to alleviate the growing concerns regarding private information leaked through MIAs in the FL setting. In this section, we review existing defense strategies and divide them into four categories, namely, partial sharing, secure aggregation, noise perturbation and anomaly detection. Each category is classified further based on specific approaches, as illustrated in Table 7.

Туре	Approach	Advantages	Disadvantages
Partial sharing	Gradient compression	Slight utility loss	Limited mitigation effect
rartiai sharing	Weight pruning	Weight pruning Low computation cost	
	Secure multi-party computation	Lossless model utility	High computation cost
Secure aggregation	Homomorphic encryption		Vulnerable to malicious clients
	Tromomorphic encryption	Protection from the central server	
Noise perturbation	Differential privacy	Strong privacy guarantee	High model utility loss
ivoise perturbation	Random perturbation	Protection from a server and clients	Ingli model utility loss
Anomaly detection	Misbehaving identification	Defend against poisoning MIAs	Fail for passive attacks

Table 7. Taxonomy of membership inference defenses in federated learning.

## 5.1 Partial Sharing

Partial sharing aims to reduce the effectiveness of inference attacks by suppressing certain updates exchanged during the FL process. Since the internal state of FL models is susceptible to MIA attacks through sharing parameters or gradients, one straightforward approach is to limit the information available to adversaries by reducing the amount of data shared [66]. Defenses based on this category can be further divided into *gradient compression* and *weight pruning*.

5.1.1 Gradient Compression. This defense strategy does not upload all gradient tensors but transmits only a few of them via selective strategies, such as by choosing the top-K largest values [62, 63, 114] or those exceeding a certain threshold [115]. The intuition is that although a global model depends on local updates to learn knowledge from training data, not all updates equally contribute to model parameters. In addition to privacy protection, this strategy can also decrease the transferred data volume and save communication costs during the FL process.

A partial sharing algorithm, namely, distributed selective SGD (DSSGD) [62], uses only a tiny fraction (i.e., 1%) of gradients shared per-participant to achieve better performance than CL. DSSGD allows an FL participant to download the latest global model, compute the top-K largest gradients, and upload them to the server. Melis et al. [17] explore the use of DSSGD to defend MIAs on a

111:20 Bai et al.

two-party sentiment classifier and show that the attack accuracy decreases from 0.93 AUC to 0.84 AUC when only 10% of the updates are shared during FL.

In addition to intuitive selection, researchers have also developed advanced techniques to share fewer gradients while maintaining the model performance. For example, deep gradient compression (DGC) [116] incorporates momentum correction and local gradient clipping after gradient sparsification to preserve the model performance. Moreover, warm-up training is introduced to address the staleness issue [117] in the learning process. By using only 0.1% of the gradient exchange in distributed SGD, this approach achieves a high compression magnitude of 270-600x. This work [118] proposes a novel gradient compression technique that delays the transmission of ambiguously estimated gradient elements whose amplitude is small than their variance over the data points. These methods can drastically compress the exchanged gradients while maintaining the original model's accuracy.

Instead of directly exchanging the gradient values, signSGD [105] exchanges the sign of the gradient through majority vote aggregation. This technique enables convergence in large-scale and mini-batch datasets with theoretical and empirical evidence. Recent work [33] has shown that signSGD is an effective countermeasure against MIAs, and the validation accuracy of the target model decreases by less than 1%. Moreover, signSGD provides a powerful defense against label inference attacks [119] and Byzantine attacks, even in the case of up to 50% of adversarial workers.

An extreme solution for compression is to hide all raw gradient values and directions [120]. In essence, this approach exchanges the predictions of local models and obtains a robust mean estimation on all predictions, thereby mitigating MIAs and poisoning attacks [26]. Intuitively, the release of only predictions is safer than the release of gradients, as high-dimensional gradients encode more information pertaining to the local data. Since this approach essentially transforms the attack setting from white-box to black-box, off-the-box CL defenses can be integrated to enhance the defense performance. Moreover, this approach can be applied in heterogeneous FL, as opposed to aforementioned approaches that are only suitable for homogeneous FL.

5.1.2 Weight Pruning. Since FL frameworks facilitate the exchange of model parameters among participants, weight pruning transmits only a few model parameters rather than all of them from each participant. Inspired by weight pruning techniques in ML, researchers [121] develop an approach to pruning parameters in the global model to guard against MIAs. A sparsified model containing fewer than 5% of the original model parameters can achieve comparable performance to that of the original model while more effectively mitigating MIAs. However, balancing the sparsity and accuracy of pruned neural networks may require model retraining, and the reuse of training samples can increase the risk of privacy violations through memorization [122].

In a nutshell, partial sharing defense aims to exchange incomplete model updates to mitigate the threat of MIAs and lower communication costs without significant utility loss in FL. In addition, extreme compression methods can defend against other privacy and security threats, such as model inversion attacks [116, 118] and Byzantine attacks [105]. However, such approaches only provide limited protection of membership privacy [17, 123], as they often have to exchange gradients or parameters that contain essential information about the training data to maintain the original performance.

## 5.2 Secure Aggregation

Secure aggregation typically relies on cryptographic techniques to prevent information leakage from potential adversaries. The underlying idea is to disclose the encrypted model updates instead of the clear ones throughout the learning process. This privacy-preserving approach can be divided into secure multi-party computation (SMC) and homomorphic encryption (HE).

5.2.1 Secure Multi-party Computation. SMC is a conventional technique to safeguard input when multiple participants collectively compute a specific output [55, 64]. An effective SMC protocol must fulfill two essential requirements [124, 125]: correctness, i.e., the outcome computed by the protocol must be accurate; and privacy, i.e., the information of no participant should be revealed to others. SMC restricts the central aggregator from learning only the summation or average of the updates of clients and safeguards individual local updates from potential eavesdroppers or a curious server.

Active research has been conducted on SMC for the protection of sensitive information. For example, Mohassel et al. [126] introduce a groundbreaking protocol for safeguarding privacy in ML, which enables the computation of non-linear activation functions, allows SGD optimization, and supports linear regression and neural networks. Following this study, SecAgg [127] is proposed to protect FL from an honest-but-curious server. This protocol uses secret sharing and double-masking approaches to ensure that the server gains no knowledge beyond an aggregated model and cannot observe clear individual local models. Similarly, Sayyad et al. [128] generate secret shared entities to address privacy concerns for the data involved in deep learning. However, SMC protocols suffer from high computational costs in real-world applications. To address this problem, Fereidooni et al. [129] propose an efficient aggregation protocol named SAFELearn, which prevents client-side information leakage and limits the access of the central aggregator access to only the aggregated results of client updates. This approach has garnered significant attention because it does not rely on a trusted server, requires only two communication rounds, and allows clients to drop out.

While existing studies assert that SMC can safeguard participants' private information in FL, its practical and theoretical effectiveness as a defense mechanism remains limited [17, 129, 130]. Previous research [30] shows that SecAgg [127] fails to guarantee data privacy against client-side MIAs, and [131] points out that a malicious server can collude with other clients to compromise privacy, even with the use of SMC. Additionally, this study[130] presents a formal analysis of SecAgg against MIAs from the perspective of DP, arguing that its privacy guarantee depends on the model's dimensionality and the number of clients. It theoretically concludes that SecAgg provides weak privacy protection in FL, as the model size is usually much larger than the client pool.

To address these limitations, hybrid countermeasures have been developed to prevent information leakage from a curious participant that can access the global model. Truex et al. [82] present a privacy-preserving approach that prevents intermediate messages and the final trained model from inference attacks. This approach combines SMC and differential privacy [132] to achieve privacy protection without sacrificing much accuracy. Each participant adds noise to their local model, which is then encrypted using the Paillier cryptosystem before being sent to the server. By combining both defenses, this approach ensures data privacy with little utility loss.

5.2.2 Homomorphic Encryption. HE [133] is a practical technique to protect sensitive data by encrypting the uploaded parameters in FL. An HE cryptosystem **H** provides an operation  $\star$  that satisfies  $\mathbf{H}(m_1) \star \mathbf{H}(m_2) = \mathbf{H}(m_1 \star m_2)$ , where  $m_1$  and  $m_2$  are sets of plaintext. In essence, this technique allows certain operations to be performed in the ciphertext space without restoring the plaintext, and thus, the performance of the training model will not be affected [124, 125, 133].

Several privacy-preserving approaches for FL have been developed based on HE. Phong et al. [123] exploit HE to construct a secure FL system, thereby alleviating the data leakage from semi-trusted third-party servers. Notably, although this method can prevent the attacker from directly obtaining the local data of other parties, the attacker can still infer the distribution of private data [32]. However, since HE involves high computational and communication overheads, researchers have attempted to simplify the aggregation scheme when protecting privacy in FL. Bai et al. [134] combine HE and a selective parameter scheme to defend against MIAs and poisoning attacks from

111:22 Bai et al.

participants and the server. Designed for a cross-silo FL scenario, BatchCrypt [135] encodes a batch of quantized gradients into a long integer instead of treating full-precision gradients while incurring an accuracy drop of less than 1%, helps accelerate the training process and significantly reduces the computation and communication costs associated with HE. Ma et al. [136] propose a novel aggregation protocol using individual client model initialization and model updating to prevent eavesdroppers from inferring the local and global models, resulting in privacy enhancement and lowered computational costs in FL.

To summarize, secure aggregation utilizes encryption methods to reveal essential information to designated participants while safeguarding private data from being deduced [137]. Furthermore, it can deal with the encrypted updates without affecting the training results of the model [138]. Nonetheless, cryptographic algorithms introduce computational and communication overhead. Additionally, a curious participant may compromise the privacy of a global model protected by SMC and HE [30].

## 5.3 Noise Perturbation

Noise perturbation relies on the idea that the introduction of noises can hinder adversaries from discerning sensitive information during inference. These noises are incorporated into the target model in accordance with privacy requirements or optimization objectives. This defense category contains both *differential privacy (DP)* and *random perturbation*.

5.3.1 Differential Privacy. DP [132, 139, 140] is a lightweight privacy-preserving technique in which noise is added to sensitive data to offer strict privacy protection. It can be formally defined as follows: A randomized mechanism  $\mathcal{M}: \mathbf{X}^n \to \mathbf{Y}$  satisfies  $(\epsilon, \delta)$ -DP for input  $\mathbf{X}^n$  and output  $\mathbf{Y}$ , if for two neighboring subsets  $D, D' \in \mathbf{X}^n$  that differ at most one element and for any  $Y \in \mathbf{Y}$ , the following inequality holds:

$$P\left[\mathcal{M}\left(D\right) \in Y\right] \le e^{\epsilon} P\left[\mathcal{M}\left(D'\right) \in Y\right] + \delta,\tag{9}$$

where  $\epsilon$  refers to the privacy budget, and  $\delta$  is the failure probability. In essence, DP is a natural countermeasure to MIAs, as DP constrains the discrepancy between the presence and absence of a sample. Extensive research based on theoretical and empirical perspectives has proven that DP can ensure data privacy against MIAs.

## (i) Effectiveness of DP from the Theoretical Perspective.

Several studies have demonstrated the effectiveness of DP in thwarting MIAs and derived theoretical leakage upper bounds based on different assumptions and attack evaluation metrics [85, 89, 141, 142]. Yeom et al. [85] draw inspiration from the correlation between DP and model generalization to establish a formal association between DP and MIAs. The authors introduce a new metric, *membership advantage*, which evaluates the attack performance by measuring the disparity between the true positive rate (TPR) and false positive rate (FPR), and demonstrate that the advantage of an attacker from a DP model is smaller than  $e^{\epsilon}-1$ . Later, based on the proposition in [143] that limits the TPR of any test to  $(\epsilon, \delta)$ -DP, Erlingsson et al. [141] derive a tighter bound for the membership advantage, that is,  $1-e^{-\epsilon}(1-\delta)$ .

The abovementioned bounds are derived under two restrictive assumptions: (1) the attacker derives membership information from a black-box model, and (2) the member and non-member instances adhere to the IID principle. These assumptions do not always hold in FL settings, which involve white-box target models and non-IID data distributions. As such, researchers have attempted to extend them to realistic settings associated with FL scenarios. Considering a white-box setting, Bernau et al. [142] establish the expected membership advantage for an all-powerful adversary that can access arbitrary background information, except for one identified record. The adversary

is assumed to be able to obtain the neighborhood dataset and observe the gradient during model training, similar to an insider attacker in FL. Motivated by data dependency in the training samples, other researchers [144] explore the protection strength of DP when statistical dependencies exist among instances. According to a strong-privacy membership experiment [145], the authors derive an upper bound for the membership advantage under  $(\epsilon,\delta)$ -DP as  $(e^{\epsilon}-1+2\delta)/(e^{\epsilon}+1)$ . However, this bound about membership advantage can be very large in non-IID scenarios when there are statistical dependencies.

Different from theoretical bounds derived from membership advantage [85, 141, 142], Sablayrolles et al. [89] focus on the likelihood of an instance being a member and derive an upper bound of this probability under  $(\epsilon, \delta)$ -DP. An ML model can be treated as a posterior distribution over parameters, and the probability of an instance being a member is bounded as  $\lambda + \frac{\epsilon}{4} + \delta$ , where  $\lambda$  denotes the prior probability determined through random guessing. Although these theoretical works indicate that DP can reduce the effectiveness of MIAs, they do not accurately reflect the risks of MIAs in practice, particularly for large  $\epsilon$  [85, 144].

# (ii) Effectiveness of DP from the Empirical Perspective.

A substantial body of literature has presented empirical evidence for the effectiveness of DP, mainly by local DP (LDP) [104, 146] or central DP (CDP) [41, 42], in mitigating the threat of MIAs in FL. In LDP-based approaches, a participant chooses a DP mechanism  $\mathcal{M}$  and adds noise locally to their data under privacy requirements to achieve record-level (individual data record) protection for private data. By contrast, in CDP, noise is added to the aggregation model on the server, which renders the model outputs indistinguishable from adversaries and helps ensure client-level (users who participate in FL) privacy against MIAs.

*LDP-based Defenses*. Researchers have used LDP to conceal the effect of individual training examples within the private dataset of a client [66, 91, 147]. DPSGD [104] is a typical algorithm to implement LDP. A moment accountant scheme is used to assign the privacy budget in the learning process. Rahman et al. [147] use DPSGD as a defense against MIAs for an ML model. However, the authors highlight that realizing perfect protection requires a stringent privacy budget of  $\epsilon \leq 2$ , which comes at the cost of reduced model utility. Mohammad et al. [66] find that LDP can effectively protect membership information and reduce the attack accuracy from 73% to 52%. Hu et al. [29] observe that vanilla LDP is ineffective against their attacks based on the prediction loss in FL settings. When the privacy budget is smaller than 2, a defender is close to random guessing at the expense of significant model utility loss.

Some efforts have been made to address the trade-off between LDP effectiveness and degraded utility. Relaxed versions can decrease the utility loss under the same privacy requirement, but they reveal more private information [91]. A hybrid countermeasure [32] is developed by combining LDP with trust domain division. The proposed approach exploits authentication based on certificates issued by a trusted certificate authority, restricts communication to certified clients, prevents potential eavesdroppers from inferring, and helps balance the model utility and privacy guarantee.

**CDP-based Defenses.** Studies on CDP offer client-level privacy protection by concealing the individual contributions of participants to FL [17, 36, 66, 148]. In this approach, the server clips the  $l_2$  norm of each participant's update, aggregates the clipped updates, and then adds noise to the aggregated result to ensure privacy [41, 42].

Several studies have validated the effectiveness of CDP against MIAs in practice. Mohammad et al. [66] investigate the suitability of CDP in ensuring privacy protection and robustness against MIAs and backdoor attacks and demonstrate that CDP mechanisms could successfully prevent the white-box passive and active MIAs proposed in [26]. The defense reduces attack accuracy from 75% to 52% with a utility drop of about 16% on the CIFAR100 dataset. By examining the performances of DP with various granularities [104, 111, 111], Suri et al. [36] highlight that client-level privacy

111:24 Bai et al.

protection provides the best defense against membership information leakage, and LDP provided the least protection. Despite its promising potential, the practical application of this approach faces challenges in some scenarios. For example, Melis et al. [17] conduct empirical evaluations in a federated learning (FL) setting with fewer than 30 participants. They find that the global model fails to converge because the magnitude of the noise added is inversely proportional to the number of participants.

5.3.2 Random Perturbation. This method aims to protect private information by introducing carefully crafted noise under an optimization objective that minimizes the model utility loss and provides a privacy guarantee.

Well-crafted perturbations can be introduced into model updates while conducting FL training. An accuracy-lossless noise perturbation method [65] can address the problem that DP offers solid theoretical guarantees but invariably impairs model accuracy [17, 18] by adding removable noise in the FL setting. Inspired by the random sketching technique [149] used to defend against property inference and model construction attacks, researchers [65] developed a technique to prevent malicious clients from accessing true global model parameters and local gradients through Hadamard products and linear outputs. This approach reduces an attack accuracy of approximately 50% while maintaining the learning accuracy. However, this approach applies only to networks that use ReLU as the activation function and cannot be extended to networks involving sigmoid and tanh functions. Additionally, it relies on a trustworthy server that will not attempt to infer private information from the recovered updates. Besides, with reference to the MemGuard approach [92] against black-box attacks in the context of CL, Xie et al. [150] devise a noise addition technique to deceive adversaries into random guessing the membership status in the FL setting.

Another research line within this type involves perturbing the model's input data. Yang et al. [151] introduce a client-level input perturbation called CIP to enhance the FL framework's data privacy by altering each client's local data distribution. Similarly, Lee et al. [60] design a digestive neural network for private data and provide anonymized training inputs to mitigate the risk of disclosing membership information.

In short, noise perturbation reduces the dependence between model performance and input data to maintain private information by adding noise. Theoretical and empirical studies have demonstrated that DP schemes allow participants to maintain local data privacy within the FL framework. However, the effectiveness of DP often comes at the cost of significant utility loss, and it is challenging to select a suitable privacy budget to balance privacy and utility. In contrast, random perturbation adds well-crafted noises restricted by optimization objectives, leading to a better utility-privacy trade-off than DP.

## 5.4 Anomaly Detection

Anomaly detection is crucial in safeguarding FL processes by identifying irregular updates and thwarting malicious influences. Collaborative training renders FL models susceptible to malevolent manipulations like model and data poisoning. These attacks have given rise to innovative FL techniques known as poisoning MIAs [26, 48, 49], which exploit vulnerabilities to enhance the leakage of membership information. Notably, the poisoning MIAs focus on sample-level privacy leakage, diverging from conventional model poisoning attacks that seek to compromise overall model performance. Robust aggregation algorithms such as Multi-Krum [152], Trimmed-Mean [153] and FLTrust [154] are inadequate in countering this active threat [67]. Consequently, the anomaly detection approach are tailored to combat active MIAs in FL.

Ma et al. [67] introduce an innovative client-side countermeasure called LoDEN to defend against active attacks in [26] by identifying and removing suspicious training samples. The active

attack, executed through the gradient ascent algorithm, deliberately alters the model update of specific training examples, allowing the attacker to infer the membership status by observing gradient changes. LoDEN employs a localized approach to counteract its impact on the FL model. It identifies malicious updates by monitoring abrupt changes in the model outputs of training samples. Specifically, if the predicted label for a training sample shifts from correct to incorrect over several iterations, LoDEN identifies it as a malicious example. This example and its neighbors are removed from subsequent training, thus safeguarding the membership privacy of target models.

Table 8. Summary of studies on membership inference defenses in federated learning.

Туре	Year	Reference	Defender	Corres. attack	Defense approach	Comparison
	2019	Melis et al. [17]	Server	[17]	Selective gradient	-
Partial sharing	2018	Bernstein et al. [105]	Client Server	[33]	Gradient sign	Differential privacy Mix-up+MMD [155]
raitiai siiaiiiig	2019	Chang et al. [120]	Server	[26]	Prediction aggregation	-
	2022	Stripelis et al. [121]	Server	[31]	Weight pruning	FedAvg
Secure aggregation	2017	Bonawitz et al.[127]	Client	[30]	Secret sharing & Double-masking	-
	2017	Bai et al.[134]	Client	[26]	HE	-
	2018	Rahman et al. [147]	Client Server	[22]	LDP	Random guessing
	2021	Hu et al. [29]	Client	[29]	LDP	-
	2022	Suri et al. [36]	Client	[36]	LDP Subject DP	Random guessing
Noise perturbation	2022	Liu et al. [32]	Server	[22]	LDP & Trust domain	-
	2022	Naseri et al. [66]	Client Server	[26]	LDP CDP	Norm bounding
	2022	Yang et al. [65]	Server	[26]	Random perturbation	FedAvg, PPDL [62] DBCL [156], SPN [149]
	2021	Xie et al. [150]	Server	[26]	Adversarial example	Random guessing
	2021	Lee et al. [60]	Client	[26]	Digestive network	DP
Anomaly detection	2023	Ma et al. [67]	Client	[26]	Sample selection	Robust aggregation

Table 9. A comparison of defense approaches in centralized and federated learning.

Setting	Туре	Phase	Protection	Unique <sup>1</sup>
	Output perturbation	Inference	Model output	•
CL	Regularization	Training	Target model	0
	Knowledge distillation	Training	Target model	•
	Differential privacy	Training	Target model	0
	Partial sharing	Communication	Model update	•
FL	Secure aggregation	Communication and aggregation	Model update and itself	•
rL .	Noise perturbation	Training and aggregation	Target model	•
	Anomaly detection	Aggregation	Target model	•

 $<sup>^{1}</sup>$   $\bullet$ : unique defense  $\bullet$ : partially unique defense  $\bigcirc$ : common defense

111:26 Bai et al.

## 5.5 Comparison to Defenses in Centralized Learning

This section discusses four strategies to mitigate MIAs in FL, but their capabilities to protect against attacks vary. Partial sharing hides certain updates to diminish the effectiveness of attacks, which can be applied to arbitrary threat models. Regarding secure aggregation utilizing cryptographic techniques, SMC and HE generally provide defense against server-side attacks but become ineffective to the inference performed by a client with legal access to the global model. One of the noise perturbation methods, DP offers a strict theoretical guarantee in safeguarding against attacks from all sides in an arbitrary strategy but with an inevitable utility loss, which motivates researchers to address this limitation by incorporating removable or well-crafted noises. The anomaly detection method is tailored to counter active attacks but proves ineffective against passive attacks. Table 8 summarizes representative defense studies that have reported empirical performance in terms of release year, defender role, defense approach, corresponding MIAs (i.e., corres.attack), and comparison methods.

Moreover, our research explores a comparative analysis of mitigation strategies in CL and FL, with the goal of deepening the understanding of countermeasures specifically within the FL context. We examine defensive strategies from two angles: the phase when a defender implements countermeasure algorithms and the objects they aim to safeguard against information leaks. These comparisons are outlined in Table 9. Notably, the process of model updates, a distinctive trait within the FL framework, stands out as an additional avenue for potential breaches in membership privacy. Consequently, defense strategies employed in FL, such as partial sharing, secure aggregation, and anomaly detection, significantly differ from those related to CL contexts.

## **6 FUTURE DIRECTION**

The field of MIA privacy is rapidly evolving, presenting numerous challenges and opportunities. In the following, we discuss potential research directions on MIAs and defense strategies in the context of FL.

# 6.1 Research Direction about Membership Inference Attacks

Holistic Evaluation Metrics. The existing evaluation metrics for MIAs provide an incomplete picture of attack performance [157]. Although previous attacks perform well in the FL setting, their evaluation is focused only on member classes, such as the attack accuracy [17, 26, 29], precision, and recall [27, 93]. Such evaluations are often misleading owing to high false positive rate (FPR) or false alarm rate (FAR), which inspires researchers to consider the performance of negative samples by adding FAR or TPR at low FPR [88, 157]. Furthermore, current metrics provide an average evaluation of record-level membership, which is unsuitable for source-level MIAs due to the non-IID phenomenon in FL. Hence, it is necessary to develop a comprehensive and fair evaluation metric for various MIAs.

Extension to Realistic Assumptions. Current strategies for MIAs in FL typically involve strict assumptions that may not hold in real-world scenarios, including the presence of IID training members [33], a limited number of participants [17, 26, 33], and all federation round joining [27, 93]. These unrealistic assumptions lead to a misunderstanding of the effectiveness of previous attacks. For example, heterogeneous data distributions are associated with a higher inference risk than IID data [36]. Moreover, different epochs in which participants join FL influence the performance of MIAs for a local adversary [26]. Future development should consider relaxing these assumptions and implementing MIAs in real-world and dynamic scenarios.

**Attacks toward Emerging Frameworks**. The membership privacy risks associated with emerging FL frameworks remain underexplored. Although several researchers have focused on centralized FL

[12, 158], in which a central server performs aggregations and broadcasts the aggregated results to participants, research on decentralized FL, such as peer-to-peer [78, 159] and blockchain [160, 161], remain limited. In the case of decentralized frameworks, a curious attacker can observe the latest model updated by a known client, thereby facilitating the launch of source-level MIAs. Furthermore, when it comes to widely-used VFL framework [162], previous MIAs prove to be generally ineffective due to the fact that each participant possesses only a portion of the feature space and has access to an incomplete target model.

Emsembled Attack Strategies. Considering that FL is susceptible to various security attacks, such as poisoning attacks [152, 163–166] and backdoor attacks [167–170], there is an opportunity for future work to incorporate these attacks to enhance the effectiveness of existing MIA techniques or develop potent attack approaches. In particular, data poisoning techniques have emerged as a significant concern, as they can significantly increase the privacy risks associated with benign training samples and amplify the membership exposure of the targeted class effectively [48, 49, 100]. Reasons for Information Leakage. The comprehensive analysis of membership information leakage in the context of FL remains inadequate. In the CL scenario, there is considerable evidence that overfitting facilitates the success of MIAs in the black setting [22, 85, 86, 89, 112]. In contrast, regarding white-box FL models, internal exposure, non-IID training samples, and cooperative learning offer adversaries more opportunities [26]. They present challenges in exploring the underlying reasons behind these factors theoretically and empirically. Such investigations can uncover vulnerabilities within the FL framework and advance the development of effective attack approaches.

## 6.2 Research Direction about Membership Inference Defenses

**Unified Evaluation Benchmark.** There is a lack of a standardized evaluation benchmark for assessing the effectiveness of defense mechanisms. Currently, countermeasures are typically evaluated based on distinct MIAs [29, 36, 62] or discussed generally without experimental evidence [128, 129]. Although several researchers [65, 66, 120, 150] have evaluated the effectiveness of defense strategies for the same MIAs [26], their results cannot be compared owing to the use of different datasets and networks. Consequently, it is challenging to identify suitable mitigation mechanisms for a defender. A unified evaluation benchmark can address this problem and provide valuable guidance in making practical choices.

**Defense Algorithm Assessment.** Previous research has often overstated the effectiveness of defense mechanisms in mitigating information leakage. Non-IID data and homogeneous model architecture have a significant impact on privacy leakage risk. Unfortunately, they are often overlooked when designing new defense mechanisms [31, 36, 98]. Secure aggregation is often considered an effective technique for an honest-but-curious server. However, recent research [131] argued that a server could infer categorical information about a specific participant. Indeed, it indicates that when the server acts as an attacker, relying on secure aggregation in FL is insufficient to protect membership information adequately. As a result, there is a growing need for researchers to evaluate these mechanisms under realistic assumptions and attack scenarios.

**Privacy-Utility-Efficiency Defense Mechanisms.** There is a scarcity of effective countermeasures that guarantee privacy preservation with small utility losses and low computational costs. DP mechanisms provide strong privacy guarantees at the cost of large utility loss [66, 91, 147] and are thus unsuitable for mission-critical applications. Partial sharing mechanisms only slightly decrease the effectiveness of attack performance [17, 33], and secure aggregation suffers from high computational costs. Existing defense mechanisms fall short of providing comprehensive protection for membership information. This issue could be addressed by integrating multiple privacy preservation methods, leading to enhanced privacy safeguards [62, 137].

111:28 Bai et al.

Robust Defense Mechanisms. Exploration of robust protection strategies shows promise in preventing membership leakage while avoiding introducing additional risks. An instance of this is when FL protected by DP inadvertently creates an opportunity for model poisoning attacks, enabling attackers to evade anomaly detection [171]. Additionally, given the strong negative correlation between MIAs and model extraction attacks [172, 173], the mitigation of MIAs may improve the effectiveness of model extraction attacks. Such occurrences should be avoided as the objective is to build a robust and safe FL model in most scenarios. Hence, researchers should consider the latent correlation among attacks when developing defense approaches against MIAs.

#### 7 CONCLUSION

MIAs represent a critical and rapidly evolving research area within FL. This survey comprehensively summarizes the landscape of MIAs and corresponding defenses within the FL framework, providing structured taxonomies to categorize existing literature. Based on the utilization of attack knowledge in these studies, we categorize MIA research into two primary approaches: update-based and trend-based. In addition to attacks in FL, we highlight prevalent defense mechanisms currently deployed to mitigate the vulnerabilities MIAs pose. These defenses encompass a range of strategies, from partial sharing techniques to noise perturbation. Furthermore, this survey identifies promising avenues for future research in MIA and defense strategies. Key opportunities include exploring novel attack vectors in diverse FL settings, refining existing defense mechanisms, and integrating robust privacy-preserving techniques into FL frameworks. These future studies can help enhance security and privacy guarantees in FL.

#### **ACKNOWLEDGMENTS**

This work was supported by the National Natural Science Foundation of China (Grant No: 62072390, 92270123, 62102334), and the Research Grants Council, Hong Kong SAR, China (Grant No: 15222118, 15218919, 15203120, 15226221, 15225921, 15209922, and C2004-21GF).

#### REFERENCES

- [1] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35, 4 (2003), 399–458.
- [2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition. 248–255.
- [4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.
- [6] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.
- [7] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research 304 (2021), 114135.
- [8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations* (2013).
- [9] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).
- [10] GDPR. 2019. General Data Protection Regulation. https://gdpr-info.eu/.
- [11] CCPA. 2018. California Consumer Privacy Act. https://leginfo.legislature.ca.gov/.
- [12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics. PMLR, 1273–1282.

- [13] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020).
- [14] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. Advances in neural information processing systems 32 (2019).
- [15] Akiyoshi Sannai. 2018. Reconstruction of training samples from loss functions. arXiv preprint arXiv:1805.07337 (2018).
- [16] Meng Shen, Huan Wang, Bin Zhang, Liehuang Zhu, Ke Xu, Qi Li, and Xiaojiang Du. 2020. Exploiting unintended property leakage in blockchain-assisted federated learning for intelligent edge computing. *IEEE Internet of Things Journal* 8, 4 (2020), 2265–2275.
- [17] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE symposium on security and privacy (SP). IEEE, 691–706.
- [18] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. 603–618.
- [19] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 1322–1333.
- [20] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. 2019. Beyond inferring class representatives: User-level privacy leakage from federated learning. In IEEE INFOCOM 2019-IEEE conference on computer communications. IEEE, 2512–2520.
- [21] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS genetics 4, 8 (2008).
- [22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP). IEEE, 3–18.
- [23] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2018. Knock knock, who's there? Membership inference on aggregate location data. *Network and Distributed System Security Symposium* (2018).
- [24] Hongyang Yan, Shuhao Li, Yajie Wang, Yaoyuan Zhang, Kashif Sharif, Haibo Hu, and Yuanzhang Li. 2022. Membership Inference Attacks Against Deep Learning Models Via Logits Distribution. IEEE Transactions on Dependable and Secure Computing (2022).
- [25] Yaxin Xiao, Qingqing Ye, Haibo Hu, Huadi Zheng, Chengfang Fang, and Jie Shi. 2022. MExMI: Pool-based Active Model Extraction Crossover Membership Inference. Advances in Neural Information Processing Systems 35 (2022), 10203–10216.
- [26] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In 2019 IEEE symposium on security and privacy (SP). IEEE, 739–753.
- [27] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. 2020. Gan enhanced membership inference: A passive local attack in federated learning. In ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 1–6.
- [28] Anastasia Pustozerova, Rudolf Mayer, and " ". 2020. Information leaks in federated learning. In *Proceedings of the Network and Distributed System Security Symposium*, Vol. 10.
- [29] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. 2021. Source inference attacks in federated learning. In 2021 IEEE International Conference on Data Mining (ICDM). IEEE, 1102–1107.
- [30] Yuhao Gu, Yuebin Bai, and Shubin Xu. 2022. CS-MIA: Membership inference attack based on prediction confidence series in federated learning. *Journal of Information Security and Applications* 67 (2022), 103201.
- [31] Umang Gupta, Dimitris Stripelis, Pradeep K Lam, Paul Thompson, José Luis Ambite, and Greg Ver Steeg. 2021. Membership inference attacks on deep regression models for neuroimaging. In *Medical Imaging with Deep Learning*. PMLR, 228–251.
- [32] Zhenpeng Liu, Ruilin Li, Dewei Miao, Lele Ren, and Yonggang Zhao. 2022. Membership Inference Defense in Distributed Federated Learning Based on Gradient Differential Privacy and Trust Domain Division Mechanisms. Security and Communication Networks 2022 (2022).
- [33] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2023. Effective passive membership inference attacks in federated learning against overparameterized models. In 11th International Conference on Learning Representations (ICLR).
- [34] Georg Pichler, Marco Romanelli, Leonardo Rey Vega, and Pablo Piantanida. 2022. Perfectly Accurate Membership Inference by a Dishonest Central Server in Federated Learning. arXiv preprint arXiv:2203.16463 (2022).
- [35] Oualid Zari, Chuan Xu, and Giovanni Neglia. 2021. Efficient passive membership inference attack in federated learning. In NeurIPS PriML workshop.

111:30 Bai et al.

[36] Anshuman Suri, Pallika Kanani, Virendra J Marathe, and Daniel W Peterson. 2022. Subject Membership Inference Attacks in Federated Learning. arXiv preprint arXiv:2206.03317 (2022).

- [37] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. 2022. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [38] Georgios Drainakis, Konstantinos V Katsaros, Panagiotis Pantazopoulos, Vasilis Sourlas, and Angelos Amditis. 2020. Federated vs. centralized machine learning under privacy-elastic users: A comparative analysis. In 2020 IEEE 19th International Symposium on Network Computing and Applications (NCA). IEEE, 1–8.
- [39] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR) 54, 11s (2022), 1–37.
- [40] Li Hu, Anli Yan, Hongyang Yan, Jin Li, Teng Huang, Yingying Zhang, Changyu Dong, and Chunsheng Yang. 2023. Defenses to Membership Inference Attacks: A Survey. Comput. Surveys (2023).
- [41] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. In *International Conference on Learning Representations*.
- [42] Robin C Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557 (2017).
- [43] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In 26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019. The Internet Society.
- [44] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. 2021. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14820–14829.
- [45] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 880–895.
- [46] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *International conference on machine learning*. PMLR, 1964–1974.
- [47] Klas Leino and Matt Fredrikson. 2020. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In 29th USENIX security symposium (USENIX Security 20). 1605–1622.
- [48] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. Truth serum: Poisoning machine learning models to reveal their secrets. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2779–2792.
- [49] Yufei Chen, Chao Shen, Yun Shen, Cong Wang, and Yang Zhang. 2022. Amplifying Membership Exposure via Data Poisoning. In Advances in Neural Information Processing Systems.
- [50] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2021. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge* and Data Engineering (2021).
- [51] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning 14, 1–2 (2021), 1–210.
- [52] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10, 2 (2019), 1–19.
- [53] Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and S Yu Philip. 2022. Privacy and robustness in federated learning: Attacks and defenses. IEEE transactions on neural networks and learning systems (2022).
- [54] Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to Federated Learning: A Survey. ArXiv abs/2003.02133 (2020).
- [55] Junpeng Zhang, Mengqian Li, Shuiguang Zeng, Bin Xie, and Dongmei Zhao. 2021. A survey on security and privacy threats to federated learning. In 2021 International Conference on Networking and Network Applications (NaNA). 319–326. https://doi.org/10.1109/NaNA53684.2021.00062
- [56] Nader Bouacida and Prasant Mohapatra. 2021. Vulnerabilities in federated learning. IEEE Access 9 (2021), 63229-63249.
- [57] Malhar S Jere, Tyler Farnan, and Farinaz Koushanfar. 2020. A taxonomy of attacks on federated learning. IEEE Security & Privacy 19, 2 (2020), 20–28.
- [58] Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S Yu. 2023. Privacy and Fairness in Federated Learning: on the Perspective of Trade-off. *Comput. Surveys* (2023).
- [59] Gongxi Zhu, Donghao Li, Hanlin Gu, Yuxing Han, Yuan Yao, Lixin Fan, and Qiang Yang. 2024. Evaluating Membership Inference Attacks and Defenses in Federated Learning. arXiv preprint arXiv:2402.06289 (2024).
- [60] Hongkyu Lee, Jeehyeong Kim, Seyoung Ahn, Rasheed Hussain, Sunghyun Cho, and Junggab Son. 2021. Digestive neural networks: A novel defense strategy against inference attacks in federated learning. computers & security 109 (2021), 102378.

- [61] Jiale Chen, Jiale Zhang, Yanchao Zhao, Hao Han, Kun Zhu, and Bing Chen. 2020. Beyond model-level membership privacy leakage: an adversarial approach in federated learning. In 2020 29th International Conference on Computer Communications and Networks (ICCCN). IEEE, 1–9.
- [62] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 1310–1321.
- [63] Alham Aji and Kenneth Heafield. 2017. Sparse Communication for Distributed Gradient Descent. In EMNLP 2017: Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (ACL), 440–445.
- [64] Andrew C Yao. 1982. Protocols for secure computations. In 23rd annual symposium on foundations of computer science (sfcs 1982). IEEE, 160–164.
- [65] Xue Yang, Yan Feng, Weijun Fang, Jun Shao, Xiaohu Tang, Shu-Tao Xia, and Rongxing Lu. 2022. An accuracy-lossless perturbation method for defending privacy attacks in federated learning. In *Proceedings of the ACM Web Conference* 2022, 732–742.
- [66] Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. 2022. Local and Central Differential Privacy for Robustness and Privacy in Federated Learning. In 29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022. The Internet Society.
- [67] Mengyao Ma, Yanjun Zhang, Pathum Chamikara Mahawaga Arachchige, Leo Yu Zhang, Mohan Baruwal Chhetri, and Guangdong Bai. 2023. LoDen: Making Every Client in Federated Learning a Defender Against the Poisoning Membership Inference Attacks. In Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security. 122–135.
- [68] Vladimir Vapnik. 1991. Principles of risk minimization for learning theory. Advances in neural information processing systems 4 (1991).
- [69] David Saad. 1998. Online algorithms and stochastic approximations. Online Learning 5, 3 (1998), 6.
- [70] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research 12, 7 (2011).
- [71] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
- [72] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. 2021. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research 5 (2021), 1–19.
- [73] Dinh C Nguyen, Quoc-Viet Pham, Pubudu N Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia Dobre, and Won-Joo Hwang. 2022. Federated learning for smart healthcare: A survey. ACM Computing Surveys (CSUR) 55, 3 (2022), 1–37.
- [74] Latif U Khan, Walid Saad, Zhu Han, Ekram Hossain, and Choong Seon Hong. 2021. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials* 23, 3 (2021), 1759–1799.
- [75] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. 2021. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 23, 3 (2021), 1622–1658.
- [76] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. 2021. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7865–7873.
- [77] Othmane Marfoq, Chuan Xu, Giovanni Neglia, and Richard Vidal. 2020. Throughput-optimal topology design for cross-silo federated learning. Advances in Neural Information Processing Systems 33 (2020), 19478–19487.
- [78] Qinbin Li, Zeyi Wen, and Bingsheng He. 2020. Practical federated gradient boosting decision trees. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 4642–4649.
- [79] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. 2020. A secure federated transfer learning framework. *IEEE Intelligent Systems* 35, 4 (2020), 70–82.
- [80] Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, and Qiang Yang. 2019. Privacy-preserving Heterogeneous Federated Transfer Learning. In 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2552–2559.
- [81] Shreya Sharma, Chaoping Xing, Yang Liu, and Yan Kang. 2019. Secure and efficient federated transfer learning. In 2019 IEEE international conference on big data (Big Data). IEEE, 2569–2576.
- [82] Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. 2019. A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM workshop on artificial intelligence and security. 1–11.
- [83] Kin Sum Liu, Chaowei Xiao, Bo Li, and Jie Gao. 2019. Performing co-membership attacks against deep generative models. In 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 459–467.

111:32 Bai et al.

[84] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. 2021. EncoderMI: Membership inference against pre-trained encoders in contrastive learning. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2081–2095.

- [85] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st computer security foundations symposium (CSF). IEEE, 268–282.
- [86] Liwei Song and Prateek Mittal. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models. In USENIX Security Symposium.
- [87] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2022. On the Importance of Difficulty Calibration in Membership Inference Attacks. In *International Conference on Learning Representations*.
- [88] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 1897–1914.
- [89] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*. PMLR, 5558–5567.
- [90] Junxiang Zheng, Yongzhi Cao, and Hanpin Wang. 2021. Resisting membership inference attacks through knowledge distillation. Neurocomputing 452 (2021), 114–126.
- [91] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In 28th USENIX Security Symposium (USENIX Security 19). 1895–1912.
- [92] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 259–274.
- [93] Yanchao Zhao, Jiale Chen, Jiale Zhang, Zilu Yang, Huawei Tu, Hao Han, Kun Zhu, and Bing Chen. 2021. User-Level Membership Inference for Federated Learning in Wireless Network Environment. Wireless Communications and Mobile Computing 2021 (2021).
- [94] Hanlin Lu, Ming-Ju Li, Ting He, Shiqiang Wang, Vijaykrishnan Narayanan, and Kevin S Chan. 2020. Robust coreset construction for distributed machine learning. *IEEE Journal on Selected Areas in Communications* 38, 10 (2020), 2400–2417.
- [95] Alka Luqman, Anupam Chattopadhyay, and Kwok-Yan Lam. 2023. Membership Inference Vulnerabilities in Peer-to-Peer Federated Learning. In Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop. 1–5.
- [96] Soumya Banerjee, Sandip Roy, Sayyed Farid Ahamed, Devin Quinn, Marc Vucovich, Dhruv Nandakumar, Kevin Choi, Abdul Rahman, Edward Bowen, and Sachin Shetty. 2024. Mia-bad: An approach for enhancing membership inference attack and its mitigation with federated learning. In 2024 International Conference on Computing, Networking and Communications (ICNC). IEEE, 635–640.
- [97] Stacey Truex, Ling Liu, and Mehmet Emre Gursoy. 2021. Demystifying Membership Inference Attacks in Machine Learning as a Service. IEEE Transactions on Services Computing 14, 6 (2021), 2073–2089. https://doi.org/10.1109/TSC. 2019.2897554
- [98] Wei Yuan, Chaoqun Yang, Quoc Viet Hung Nguyen, Lizhen Cui, Tieke He, and Hongzhi Yin. 2023. Interaction-level Membership Inference Attack Against Federated Recommender Systems. In Proceedings of the Web Conference 2023. 1–10
- [99] Truc Nguyen, Phung Lai, Khang Tran, NhatHai Phan, and My T Thai. 2023. Active Membership Inference Attack under Local Differential Privacy in Federated Learning. In International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain (Proceedings of Machine Learning Research, Vol. 206). PMLR, 5714–5730.
- [100] Yanjun Zhang, Guangdong Bai, Mahawaga Arachchige Pathum Chamikara, Mengyao Ma, Liyue Shen, Jingwei Wang, Surya Nepal, Minhui Xue, Long Wang, and Joseph Liu. 2023. AgrEvader: Poisoning Membership Inference against Byzantine-robust Federated Learning. In Proceedings of the ACM Web Conference 2023. 2371–2382.
- [101] Gaoyang Liu, Zehao Tian, Jian Chen, Chen Wang, and Jiangchuan Liu. 2023. TEAR: Exploring Temporal Evolution of Adversarial Robustness for Membership Inference Attacks against Federated Learning. IEEE Transactions on Information Forensics and Security (2023).
- [102] Liwei Zhang, Linghui Li, Xiaoyong Li, Binsi Cai, Yali Gao, Ruobin Dou, and Luying Chen. 2023. Efficient Membership Inference Attacks against Federated Learning via Bias Differences. In Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses. 222–235.
- [103] Dan Feldman, Matthew Faulkner, and Andreas Krause. 2011. Scalable training of mixture models via coresets. Advances in neural information processing systems 24 (2011).
- [104] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications

- security. 308-318.
- [105] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. 2018. signSGD with Majority Vote is Communication Efficient and Fault Tolerant. In *International Conference on Learning Representations*.
- [106] Zilu Yang, Yanchao Zhao, and Jiale Zhang. 2023. FD-Leaks: Membership Inference Attacks Against Federated Distillation Learning. In Web and Big Data: 6th International Joint Conference, APWeb-WAIM 2022, Nanjing, China, November 25–27, 2022, Proceedings, Part III. Springer, 364–378.
- [107] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier* systems. Springer, 1–15.
- [108] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In Proceedings of the 26th international conference on world wide web. 173–182.
- [109] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 639–648.
- [110] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. 2021. Revisiting Membership Inference Under Realistic Assumptions. Proceedings on Privacy Enhancing Technologies 2021, 2 (2021).
- [111] Virendra J Marathe and Pallika Kanani. 2022. Subject Granular Differential Privacy in Federated Learning. arXiv preprint arXiv:2206.03617 (2022).
- [112] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy risks of securing machine learning models against adversarial examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 241–257
- [113] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership inference attacks by exploiting loss trajectory. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. 2085–2098.
- [114] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. 2018. Sparsified SGD with memory. Advances in Neural Information Processing Systems 31 (2018).
- [115] Aritra Dutta, El Houcine Bergou, Ahmed M Abdelmoniem, Chen-Yu Ho, Atal Narayan Sahu, Marco Canini, and Panos Kalnis. 2020. On the discrepancy between the theoretical analysis and practical implementations of compressed communication for distributed deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3817–3824.
- [116] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. 2018. Deep gradient compression: Reducing the communication bandwidth for distributed training. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.
- [117] Wei Dai, Yi Zhou, Nanqing Dong, Hao Zhang, and Eric Xing. 2019. Toward Understanding the Impact of Staleness in Distributed Machine Learning. In *International Conference on Learning Representations*.
- [118] Yusuke Tsuzuku, Hiroto Imachi, and Takuya Akiba. 2018. Variance-based Gradient Compression for Efficient Distributed Deep Learning. In 6th International Conference on Learning Representations, ICLR 2018.
- [119] Chong Fu, Xuhong Zhang, Shouling Ji, Jinyin Chen, Jingzheng Wu, Shanqing Guo, Jun Zhou, Alex X Liu, and Ting Wang. 2022. Label inference attacks against vertical federated learning. In 31st USENIX Security Symposium (USENIX Security 22). 1397–1414.
- [120] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. 2019. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. arXiv preprint arXiv:1912.11279 (2019).
- [121] Dimitris Stripelis, Umang Gupta, Nikhil Dhinagar, Greg Ver Steeg, Paul M Thompson, and José Luis Ambite. 2022. Towards Sparsified Federated Neuroimaging Models via Weight Pruning. In *Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health.* Springer, 141–151.
- [122] Xiaoyong Yuan and Lan Zhang. 2022. Membership inference attacks and defenses in neural network pruning. In 31st USENIX Security Symposium (USENIX Security 22). 4561–4578.
- [123] Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security 13, 5 (2017), 1333–1345.
- [124] Ruinian Li, Yinhao Xiao, Cheng Zhang, Tianyi Song, and Chunqiang Hu. 2018. Cryptographic algorithms for privacy-preserving online applications. *Math. Found. Comput.* 1, 4 (2018), 311–330.
- [125] Xuefei Yin, Yanming Zhu, and Jiankun Hu. 2021. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–36.
- [126] Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In 2017 IEEE symposium on security and privacy (SP). IEEE, 19–38.
- [127] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical secure aggregation for privacy-preserving machine learning. In proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 1175–1191.

111:34 Bai et al.

[128] Suhel Sayyad. 2020. Privacy preserving deep learning using secure multiparty computation. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 139–142.

- [129] Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Helen Möllering, Thien Duc Nguyen, Phillip Rieger, Ahmad-Reza Sadeghi, Thomas Schneider, Hossein Yalame, et al. 2021. SAFELearn: secure aggregation for private federated learning. In 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 56–62.
- [130] Khac-Hoang Ngo, Johan Östman, Giuseppe Durisi, and Alexandre Graell i Amat. 2024. Secure Aggregation Is Not Private Against Membership Inference Attacks. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 180–198.
- [131] Jiqiang Gao, Boyu Hou, Xiaojie Guo, Zheli Liu, Ying Zhang, Kai Chen, and Jin Li. 2021. Secure aggregation is insecure: Category inference attack on federated learning. *IEEE Transactions on Dependable and Secure Computing* (2021).
- [132] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.
- [133] Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. 1978. On data banks and privacy homomorphisms. Foundations of secure computation 4, 11 (1978), 169–180.
- [134] Yang Bai and Mingyu Fan. 2021. A method to improve the privacy and security for federated learning. In 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). IEEE, 704–708.
- [135] Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. 2020. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020)*.
- [136] Xiang Ma, Haijian Sun, Rose Qingyang Hu, and Yi Qian. 2022. A New Implementation of Federated Learning for Privacy and Security Enhancement. In GLOBECOM 2022-2022 IEEE Global Communications Conference. IEEE, 4885–4890.
- [137] Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. 2022. Vertical Federated Learning. arXiv preprint arXiv:2211.12814 (2022).
- [138] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. Knowledge-Based Systems 216 (2021), 106775.
- [139] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.
- [140] Huadi Zheng, Haibo Hu, and Ziyang Han. 2020. Preserving user privacy for machine learning: Local differential privacy or federated machine learning? *IEEE Intelligent Systems* 35, 4 (2020), 5–14.
- [141] Ulfar Erlingsson, Ilya Mironov, Ananth Raghunathan, and Shuang Song. 2019. That which we call private. arXiv preprint arXiv:1908.03566 (2019).
- [142] Daniel Bernau, Günther Eibl, Philip W Grassal, Hannah Keller, and Florian Kerschbaum. 2021. Quantifying identifiability to choose and audit  $\epsilon$  in differentially private deep learning. *Proceedings of the VLDB Endowment* 14, 13 (2021), 3335–3347.
- [143] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. 2013. Differential privacy for functions and functional data. *The Journal of Machine Learning Research* 14, 1 (2013), 703–727.
- [144] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2023. Investigating Membership Inference Attacks under Data Dependencies. In 36th IEEE Computer Security Foundations Symposium, CSF 2023, Dubrovnik, Croatia, July 10-14, 2023. IEEE, 473-488.
- [145] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In 2021 IEEE Symposium on security and privacy (SP). IEEE, 866–882.
- [146] Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. 2020. LDP-Fed: Federated Learning with Local Differential Privacy. In Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking.
- [147] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. 2018. Membership Inference Attack against Differentially Private Deep Learning Model. *Trans. Data Priv.* 11, 1 (2018), 61–79.
- [148] Marco Avella-Medina. 2021. Privacy-preserving parametric inference: a case for robust statistics. *J. Amer. Statist. Assoc.* 116, 534 (2021), 969–983.
- [149] Mengjiao Zhang and Shusen Wang. 2021. Matrix sketching for secure collaborative machine learning. In *International Conference on Machine Learning*. PMLR, 12589–12599.
- [150] Yuanyuan Xie, Bing Chen, Jiale Zhang, and Di Wu. 2021. Defending against Membership Inference Attacks in Federated learning via Adversarial Example. In 2021 17th International Conference on Mobility, Sensing and Networking (MSN). 153–160. https://doi.org/10.1109/MSN53354.2021.00036
- [151] Yuchen Yang, Haolin Yuan, Bo Hui, Neil Gong, Neil Fendley, Philippe Burlina, and Yinzhi Cao. 2023. Fortifying Federated Learning against Membership Inference Attacks via Client-level Input Perturbation. In 2023 53rd Annual

- IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, 288-301.
- [152] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems* 30 (2017).
- [153] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*. PMLR, 5650–5659.
- [154] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. 2021. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In 28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021. The Internet Society.
- [155] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2020. Membership inference attacks and defenses in supervised learning via generalization gap. arXiv preprint arXiv:2002.12062 3, 7 (2020).
- [156] Lixin Fan, Kam Woh Ng, Ce Ju, Tianyu Zhang, Chang Liu, Chee Seng Chan, and Qiang Yang. 2020. Rethinking privacy preserving deep learning: How to evaluate and thwart privacy attacks. Federated Learning: Privacy and Incentive (2020), 32–50.
- [157] Shahbaz Rezaei and Xin Liu. 2021. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7892–7900.
- [158] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604 (2018).
- [159] Lingchen Zhao, Lihao Ni, Shengshan Hu, Yaniiao Chen, Pan Zhou, Fu Xiao, and Libing Wu. 2018. Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications. IEEE, 2087–2095.
- [160] Rui Wang, Heju Li, and Erwu Liu. 2021. Blockchain-based federated learning in mobile edge networks with application in internet of vehicles. arXiv preprint arXiv:2103.01116 (2021).
- [161] Yang Zhao, Jun Zhao, Linshan Jiang, Rui Tan, Dusit Niyato, Zengxiang Li, Lingjuan Lyu, and Yingbo Liu. 2020. Privacy-preserving blockchain-based federated learning for IoT devices. IEEE Internet of Things Journal 8, 3 (2020), 1817–1829.
- [162] Bin Gu, Zhiyuan Dang, Xiang Li, and Heng Huang. 2020. Federated doubly stochastic kernel learning for vertically partitioned data. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2483–2493.
- [163] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. 2019. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*. PMLR, 634–643.
- [164] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. 2020. Local model poisoning attacks to byzantine-robust federated learning. In Proceedings of the 29th USENIX Conference on Security Symposium. 1623–1640.
- [165] Di Cao, Shan Chang, Zhijian Lin, Guohua Liu, and Donghong Sun. 2019. Understanding distributed poisoning attack in federated learning. In 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 233–239.
- [166] Leixia Wang, Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Kai Huang. 2024. LDP-Purifier: Defending against Poisoning Attacks in Local Differential Privacy. In *International Conference on Database Systems for Advanced Applications*. Springer, 3–18.
- [167] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2938–2948.
- [168] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. 2020. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*.
- [169] C-L Chen, Leana Golubchik, and Marco Paolieri. 2020. Backdoor Attacks on Federated Meta-Learning. In 34th Conference on Neural Information Processing Systems.
- [170] Haoyang Li, Qingqing Ye, Haibo Hu, Jin Li, Leixia Wang, Chengfang Fang, and Jie Shi. 2023. 3DFed: Adaptive and Extensible Framework for Covert Backdoor Attack in Federated Learning. In 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 1893–1907.
- [171] Ming Yang, Hang Cheng, Fei Chen, Ximeng Liu, Meiqing Wang, and Xibin Li. 2023. Model poisoning attack in differential privacy-based federated learning. *Information Sciences* 630 (2023), 158–172.
- [172] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs.. In USENIX security symposium, Vol. 16. 601–618.
- [173] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. 2022. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In 31st USENIX Security Symposium (USENIX Security 22). 4525–4542.