A Survey on Membership Inference Attacks Against Machine Learning

Yang Bai[1,2], Ting Chen[1], and Mingyu Fan[1]
(Corresponding author: Mingyu Fan)
School of Computer Science and Engineering, University of Electronic Science and Technology of China, China[1]
No.2006, Xiyuan Road, West High Technology District, Chengdu 611731, China
No.30 Institute of CETC, China[2]
No.8, Chuangye Road, High Technology District, Chengdu 611731, China
Email:alicepub@163.com, brokendragon@uestc.edu.cn, ff98@163.com
(Received May 31, 2020; Revised and Accepted Feb. 10, 2021; First Online June 2, 2021)

Abstract

Nowadays, machine learning is widely used in various applications. However, machine learning models are vulnerable to various membership inference attacks (MIAs) that leak information on the individual records trained by these models. Although many studies focus on finding new attack methods or improving attack performance, how to characterize MIAs is not well studied. This paper focuses on MIAs and the defense mechanisms against them by analyzing a framework that allows the general decomposition of existing MIAs against machine learning systems. We investigate MIAs by multiple key elements related to the victim model, including the adversary's observation, the prior knowledge of attacks, the classification of the target model, and the learning frame of the target model. Then, we classify the adversary's prior knowledge into seven sub-classes to further analyze the existing attacks. After that, we survey defense mechanisms employed by existing models. Our work contributes to understanding: 1) What is the working mechanism of MIAs; 2) Which components should be considered during the design of an MIA.

Keywords: Analysis Framework; Defense; Membership Inference Attack; Machine Learning

1 Introduction

In recent years, machine learning has been widely used in privacy-sensitive applications, *e.g.*, image recognition [16,25,33,59], speech recognition [21], healthcare data management [6, 14].

In such applications, privacy threats should be considered when devising machine learning techniques, especially that the training data should be protected from leakage because the training data contains sensitive information such as patients' healthcare information, personal preference, personal photos. Recently, academic work has revealed a variety of privacy threats against machine learning.

Privacy issues [8,29–31,64,65] of machine learning techniques include model inversion [15, 24, 49, 54], model extraction [47,60], and membership inference [7,22,32,38,45, 46, 52, 53, 57, 62]. A model inversion attack tries to reconstruct the model's input from output information [15]. *e.g.* Fredrikson *et al.* [15] introduce a model inversion attack that infers sensitive features used as inputs to decision tree models. In a model extraction attack, an adversary obtains black-box access to one target model and attempts to learn a model that closely approximates to, or even matches the target model [60]. The malicious user can leverage a model extraction attack to avoid query charge from the machine learning service company. The membership inference attacker aims to infer whether a specific data record is in the training data set of the target model or not [57]. Such attack can leak the privacy of the training dataset. In this paper, we focus on the overvoew of membership inference attack (MIA) against machine learning.

MIA has been extensively studied in other research fields, such as genomics privacy [28, 54] and mobility privacy [49]. Shokri *et al.*'s [57] was the first work to apply MIA against machine learning. Since then, many MIAs were proposed [7, 22, 53]. ML-Leaks [53] proposed a generic attack by relaxing some assumptions to show that such attacks are very broadly applicable at LOGAN [22] and GAN-Leaks [7] propose MIAs towards the generative machine learning model.

Although many works aimed at analyzing privacy threats and defense in machine learning systems, there lacks studies about systematical analysis of MIA and comprehensive comparisons among various attacking approaches. Such an empirical study can help researchers to understand how these attacks happen, what constraint conditions these attacks face, and what capabilities the

attackers possess. With these motivations, we provide a general survey of MIA against machine learning. Our work contributes to understanding why MIA chooses the existing designs, what are the causal factors of MIA, and how is the researching progress of defensive methods.

In this work,firstly, we construct a comprehensive framework to analyze existing MIAs against machine learning systems, which concludes four aspects: The attack observations, the prior knowledge, the target model type, and the target frame. Then we conduct a deep analysis of the prior knowledge of existing MIA and classify them into three categories and seven sub-classes. In addition, we summarize the factors that cause MIA, and classify existing defense mechanisms preventing MIA against machine learning into three categories. We also discuss their applicability to different MIA approaches and their effectiveness at mitigating these attacks. Finally, we envision three notable trends in the research on MIA methods and mitigation, which are worthy of in-depth studies in future.

The remainder of this paper is organized as follows. We discuss the attack model and state-of-the-art attack of membership inference attack against machine learning in Section 2. Then we propose the analytical aspects of this attack in Section 3.In Section 4, we summarize the factors influencing the attack and retrospect the exist defending mechanisms. This paper concludes in Section 6.

2 Terms and Prior Works Related to MIA

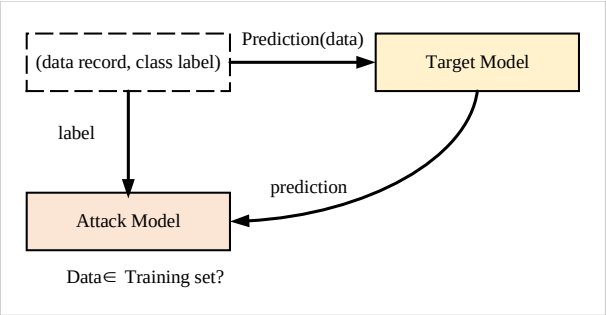

Figure 1: The working principle of MIA against machine learning

In this section, firstly, we introduce some terms that related to membership inference attack; Secondly, we give a brief review of prior works related to MIA by year-wise road map. The purpose of this section is to make us have a basic knowledge about MIA.

2.1 Terms

Membership inference attack (MIA). It shows in Figure 1 that Membership inference attacks aim to determine whether a given data point was present in the training data used to build a model [66]. Membership inference violates the privacy of both the individual participants involved in the model training and the owner of the training dataset [62]. This type of attack has been extensively studied in the adjacent area of genomics, and recently this attack is introduced in the context of machine learning [57]. In an MIA, the adversary attempts to infer whether a candidate data record is included in the training dataset of a target model. The adversary maybe given a candidate data record, or they can input some data point to target model and get out the query result. What's more, they might know some other background knowledge about the target model and training dataset. This attack then becomes a binary classification problem [57]. For each candidate record, there are two possible classes: The class "member" means that the candidate data is a member of the target model's training dataset, and the class "non-member" means otherwise. Thus, the adversary tries to establish a binary classifier to solve this problem.

Target model. In the MIA, the trained machine learning model will be treated as the adversary's target model.

Candidate data record. In the MIA, the candidate data records denote that a set of data sample which may belong to the target model's training dataset.

Shadow model. The shadow model is used to imitate the behavior of the target model, which is used in the black-box attack to obtain more information about the target model. During the attack, the adversary generates a shadow model by crafted shadow model training samples. Shadow models are models with the same architecture as the target model [45].

Attack model. Attack model is a binary classifier model used to infer the candidate data records whether are the member of the target model's training data. In other words, the adversary's attack process is the process of building an accurate attack model.

Machine Learning as a Service (MLaaS). MLaaS is an array of services that provide machine learning tools as part of cloud computing services. MLaaS helps clients benefit from machine learning without the cognate cost, time, and risk of establishing an in house internal machine learning team. Infrastructural concerns such as data pre-processing, model training, model evaluation, and ultimately, predictions, can be mitigated through MLaaS.

2.2 Prior Works Related to MIA

Figure 2 shows that the year-wise road map of MIAs against machine learning system. We describe the prior works ralated MIA by the timeline of this research direction.

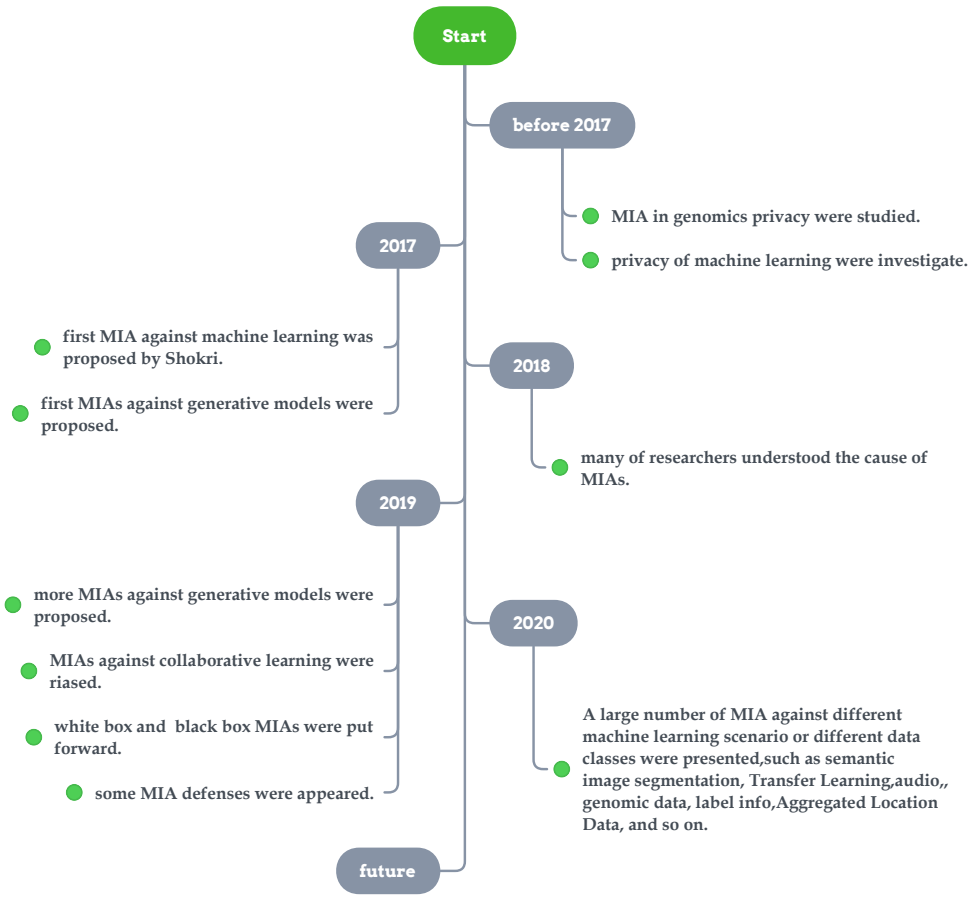

Figure 2: The year-wise road map of MIA against machine learning

Before 2017. Membership inference originated from genomics privacy related research [3, 28, 54], and then with the development of machine learning privacy research some related notions of privacy had appeared. Fredrikson *et al.* [15] demonstrated how the confidence information returned by many machine learning ML classifiers enables new model inversion attacks that could lead to unexpected privacy issues. Tramer *et al.* [60] explore model extraction attacks that could subvert model monetization, violate training-data privacy, and facilitate model evasion.

Year of 2017. 2017 should be the first year of MIA against machine learning.Because Shokri *et al.* [57] proposed the first MIA against machine learning, and they invented the shadow model technique to construct the attack models. as machine learning model includes discriminator and generator, Shokri *et al.*s' work only focused on MIA against discrimination model,but had not studied MIA in generation model. In LOGAN [22] the first MIA against generative models was presented. In this paper, Hayes *et al.* putted forward MIA against several state-of-the-art generative models, *e.g.*, Deep Convolutional GAN (DCGAN), Boundary Equilibrium GAN (BEGAN), and the combination of DCGAN with a Variational Autoencoder (DCGAN+VAE). The LOGAN introduces a full black-box attack model and a discriminator-accessible attack model against GANs. But the assumption of discriminator-accessible is the most knowledgeable but unrealistic setting because the discriminator in GAN is not always accessible in practice.

Year of 2018. Many researchers focused on understanding the cause of MIAs. Long *et al.*'s work [37]investigate and analyze membership attacks to understand why and how they succeed. And based on those understanding, they proposed Differential Training Privacy to estimate the privacy risk. In paper [38] reported a study that discover overfitting to be a sufficient but not a necessary condition for MIA to succeed, more specifically, they demonstrated that even a well-generalized model contains vulnerable instances subject. Yeom *et al.*'s [66] examined the effect that outfitting and influence have on the ability of an attacker to learn information about the training data from machine learning models, either through training set membership inference or attribute inference attacks. the cause factors will be discuss in Section 4.1. ML-Leaks [53] investigate the assumptions what a MIA requires. And they relaxes some assumptions of Shokri *et al.*'s work [57], such as the number of shadow models, the knowledge of the target model structure, and the target model's dataset information. This work reveals that attacks with relaxed assumptions are very broadly applicable at low cost and thereby pose a more severe risk than previ-

ously thought.

Year of 2019. In 2019, there were four main research advances in this direction. First of all, more MIAs against generative models were proposed. Hilpercht *et al.* [26] proposed a MIA based on Monte Carlo integration that applicable to all generative models. Chen *et al.* [7] explored the MIA against GANs and present the first taxonomy of MIAs in four classes, which included full black-box generator, partial black-box generator, withe-box generator and accessible discriminator. Secondly, MIAs against collaborative learning were raised. Melis *et al.* [41] proposed inference attacks against collaborative machine learning system. Nasr *et al.* [45] designed inference algorithms for both centralized and federated learning. thirdly, more white-box and black-box MIAs were put forward. Sablayrolles *et al.* [51] proposed a optimal inference strategy, the result showed that black-box attacks will perform as well as withe-box attacks in this optimal asympotic setting. Next, some MIAs defenses were appeared. Jia *et al.* [32] proposed MenGuard which adds noise to each confidence score vector by the target classifier to guarantees against black-box MIA. Rahimian *et al.* [50] studied the effect of Differential Privacy-Stochastic Gradient Descent(DP-SGD) to defense the MIAs. The systematic investigation of defenses against MIAs will be introduced in Section 4.

Year of 2020. A large number of MIA against different machine learning scenario or different data classes were presented. He *et al.* [23] show structural outputs of segmentation have severe risks of leaking membership, and present the first work on MIAs against semantic segmentation models while the prior works focus on classification models. As machine learning algorithms are used to process wireless signals, Shi *et al.* [55] presents how to leak privacy information from a wireless signal classifier by launching an over-the air MIA. Li *et al.* [36] investigate a MIA when the target model only provides the predicted label. Zhang *et al.* [68] propose LocMIA which allows adversaries to launch MIAs against aggregated LOCation data by train a binary classifier to infer whether a specific victim's location data involved in the aggregation group.

All of above, existing methods mainly study on finding out new attack approaches, improving the attack's performance, or proposing efficient mitigation methods against MIAs. But, none of the existing studies focus on the comprehensive analysis of MIA. Along with the development of machine learning privacy issues, more and more researchers pay attention to this field and the requirement of further research in this area increased. It is necessary to analyze the MIA from various angles to understand MIA better.

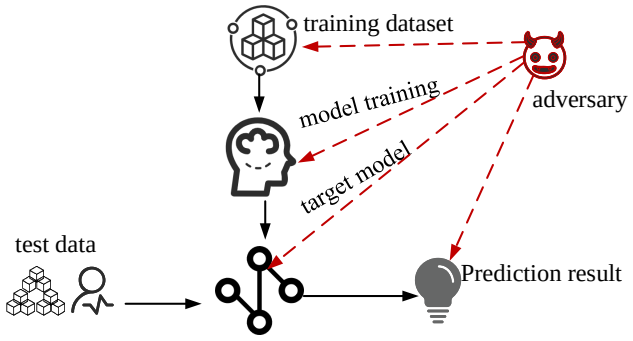

Figure 3: The elements considered by attacker to launch a MIA against machine learning

3 Analytical Framework of Membership Inference

3.1 Factors Considered in our Framework

As the processes of machine learning related to the general keywords include training dataset, model training, machine learning model, and prediction result. From the MIAs attackers view, their adversarial capability refers to the control-ability of these elements. In general, the adversary can condider several aspects for designing a MIA against machine learning system.

1) From the adversary's view, whether the adversary has knowledge about the training dataset, and what the training dataset background knowledge the attacker knows is the consideration elements.

2) During the model training, whether the model is a stand-alone or collaborative learning style, and whether the attacker is a bystander or one of the participants should be taken into consideration.

3) The adversary requires to think clearly about that the attacker can use what observation with the target model, the target model is a generative model or a discriminative model, and what detail prior information about dose the model the attacker has.

4) Considering with the prediction result, whether the adversary has the querying capability to get correspond prediction result with input data is one of the most important assumptions.

Understanding these aspects and developing an analysis structure serves a twofold purpose. First, it provides greater insight into previous researches, facilitating common ground comparison between different approaches. Second, it provides insights into the detailed design choices for MIA approaches which can contribute to the future research of membership inference attack against machine learning and the defense against the attack.

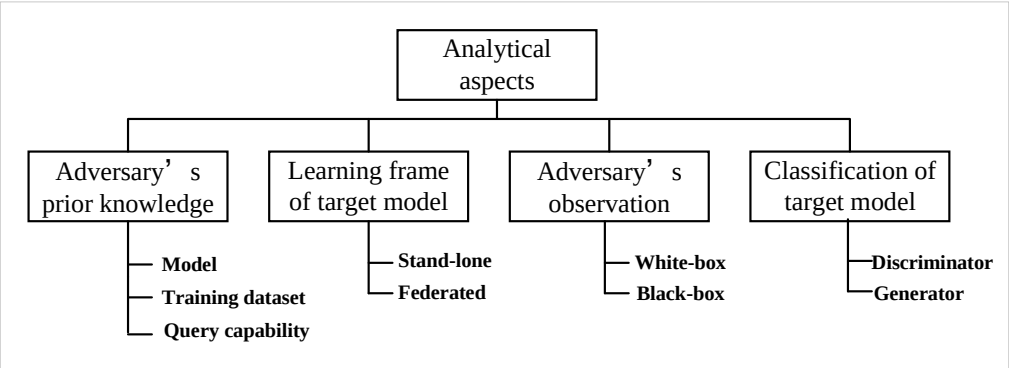

Figure 4: Analytical aspects of detail information

3.2 Analytical Framework

With all considerations mentioned above, we propose an analytical framework which includes adversary's prior knowledge, learning frame of target model, adversary's observation, classification of the target model, as shown in Figure 4. Then, we make definitions of these aspects and then study them in detail at the rest of this section. Lastly, we summary MIAs with our analytical framework in Section 3.7.

Adversary's prior knowledge. In an MIA, the more prior knowledge the adversary has, the stronger the adversary's capabilities are. On the contrary, the less prior knowledge the adversary has, the weaker the attacker's capabilities are. Several characteristics related to the adversary's capabilities. He *et al.*'s work [24] propose that the prior knowledge including three aspects: Knowledge of target model, knowledge of the training dataset, and the capability of the model querying. The prior knowledge will be comprehensive analyzed in Section 3.3.

Learning frame of the target model. The learning frame of the target model has two types, stand-alone learning frame, and federated learning frame. The adversary can different attack approaches with different learning frame.The collaborative learning will be discuss in Section 3.4 comparing with the stand-alone one.

Adversary's observation. In paper [45], they define that the adversary's observations of machine learning algorithms are what constitute the inputs for the MIA. The attack observation can be classified into the black-box and the white-box which will be discussed in Section 3.5.

Classification of the target model. There are two types of target model: Discriminative target model and the generative target model. Both of them suffer from MIA. The MIA against discriminative model and the generative model will be analyze in Section 3.6.

3.3 Adversary's Prior Knowledge

In this part, we consider the attacker's prior knowledge with completion coverage aspects, and then introduce a classification method of adversary's prior knowledge. In an MIA, the prior knowledge means the adversary's capabilities which have an impact on the attack results.

Previous works introduce several characteristics related to the adversary's capabilities. Salem *et al.* [53] studied three attacks with different prior knowledge consists of target model structure, training data distribution. He at el.'s work [24] classify the prior knowledge in three categories, *e.g.* target model, training dataset, the capability of model querying. This method covers all aspects related to MIA. Thus, we survey the previous works, by considering the prior knowledge into three aspects proposed by He *et al.* [24], *i.e.*, knowledge of target model, knowledge about the training dataset, the capability of model query. For clarity, we summarize the notations in Table 1.

Knowledge of Target Model (M). In this aspect, the adversary may obtain information about the target model, including the parameters, the structure, the type, machine learning as a service (MLaaS), and mode type, termed by M_1, M_2, M_3 and M_4 respectively.

M_1: Model parameters. Some researches [7, 22, 45] assumed that attacker knows some model parameters. The adversary can download the description of the model through MLaaS cloud systems [20, 42]. The method in [1] shows that an honest-but-curious server can partially recover participants' data points from the shared gradient updates. Paper [7] proposed an attack by which the attacker has access to the parameters of the generator.

M_2: Model structure. In paper [22,45,56], the adversary obtained knowledge of model structure.

M_3: MLaas platform. Several works [7,38,53,57,62] considered that the adversary can use the same MLaaS platform with target model.

M_4: Model type. Shokri *et al.* [57] and Hayes *et al.* [22] set the type of target model as one of the prior knowledge.

Table 1: Definition of prior knowledge's sub-classes

Knowledge Types	Symbol	Definition
Model	M_1	The parameters of the target model
	M_2	Can access or know the structure of the target model
	M_3	Use the same MLaaS platform with target mode
	M_4	The type of target model
Training Dataset	D_1	Know some properties about the target model's training dataset, such as the distribution, the size, or the value
	D_2	A dataset which includes model's training dataset
Query Ability	Q	Can query the machine learning model

Knowledge about the Training Dataset (D). Training data set is a set of examples used to initially fit the parameters (e.g. weights of connections) of the machine learning model. Each training example is represented by an array or vector, consists of pairs of an input vector and the corresponding output vector. There are many public dataset commonly used for machine learning model training, such as CIFAR-10,CIFAR-100, MNIST, Texas, Purchase-10, Purchase-100, Hospital, Location, News, and so on. The Knowledge about the training dataset means that the adversary has some information about the training dataset. the training dataset info includes the following two classes.

D_1: The attacker knows some property information about the target training dataset, such as the distribution, size or value. Long *et al.* [38] exploit datasets, which sampled from the same space as the terget training set but not containing the target record, to build the shadow model. Shokri *et al.* [57] say that they have some background knowledge about the target model's training dataset, but disjoint from the tarining dataset. Salem *et al.* [53] assume that the adversary has a dataset which comes from the same underlying distribution as the training data for the target model or perform model extraction to approximate the target model. Hayes *et al.* [22] give an assumption that the adversary knows the size of the training set, but not know how data-points are split into training and test sets.

D_2: The adversary obtains a dataset which includes model's training dataset.Nasr *et al.* [46] reveal that in a realistic setting, the probability distribution of data points and the probability distribution over the member of the training set are not directly and accurately available to the adversary. They assume a dataset known by the attacker which is the subset of the target training set. Hayes *et al.* [22] introduce a white-box attack in which the attacker has a dataset containing data-points used to train the target model.

The capability of Model Query (Q). This kind of prior knowledge means that the adversary whether can query the target model or not. In papers [7, 15, 22, 24, 37, 38, 52, 53, 57], the authors study that the adversary can query the learning model (Q).

3.4 The Frame of Target Model: Stand-Alone vs. Federated Learning

The learning frame of the target model has two major types, stand-alone learning one and federated learning one. It depends on whether all the training data is available in one place, or the training data is distributed among multiple parties [45]. The adversary has different attack approaches with different learning frames.

Stand-alone learning frame. In this setting, the target model is trained in one place, it means centralized training wherein all the training data is available in one place. Under the stand-alone learning frame, the adversary has two points of view to launch the MIA. First, the attacker can observe the process of the model updating. Second, an adversary can attack a final trained model. The latter method has been studied more than the former one, in previous works.

Federated learning frame. Comparing with the stand-alone learning frame, the federated learning frame has a distributed structure. The federated learning aims at training a machine learning algorithm on multiple local datasets contained in local participants. We illustrate the federated learning frame in Figure 5. In the federated learning frame, the central server orchestrates the different steps of the algorithm. First of all, the central server transmits the initial model to several distributed participants. Then, the participant uses local training datasets and optimizers (such as stochastic gradient descent optimizer) to train the local

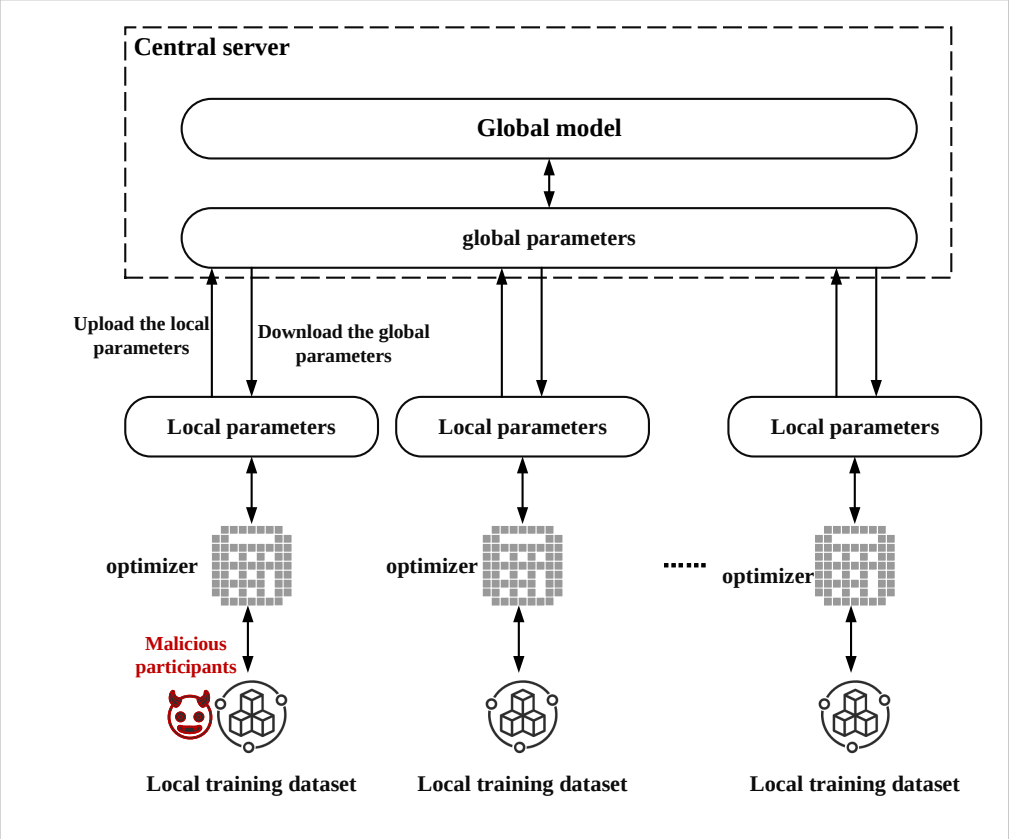

Figure 5: Federated learning frame

model. After that, participants upload their local parameters to the central server. The central server uses a specific method (such as computing average value) to transform these parameters into the global parameters. Finally, the central server generates a global model with global parameters. Federated learning constructs a global model using multiple rounds. In one round, it begins with downloading the global parameters from the central server and ends with uploading the local parameters to the central server. Collaborative training continues until the global model converges [45].

In papers [24, 41, 45], they propose an attack against federated learning, wherein the attacker is one of the participates who can observe the global parameters and craft his adversarial parameter updates to gain more information about the information of other participants' training dataset.

3.5 Observation: Black-Box vs. White-Box

White-box. A white box attack means that the attacker has access to the full model, notably its architecture and parameters. With such information, the adversary can reconstruct the target model and even the training dataset. Hayes *et al.*'s work [22] propose a white-box attack against the generative model. In their white-box scenario, the attacker relies on internal access to the target model instead of training an attack model. In paper [45], they introduce an extension of existing black-box membership to the white-box setting which uses the same attack on all of the activation functions of the model.

Black-box. In this setting, the adversary only has the capability of model querying but can neither access the model's parameters, nor the model structure. It means that the attacker can only query the target model by inputting data points to obtain output results. LOGAN [22] proposes a black-box attack with no auxiliary knowledge and a black-box attack with limited auxiliary knowledge. In ML-Leak [53], the authors present an attack whose adversary has black-box access to the target model, but the attacker not able to extract the membership status from the target model. Thus, the adversary trains a shadow model to mimic the behavior of the target model and relies on the shadow model to obtain the ground-truth membership to train the attack model. In paper [47], they introduce a black-box attack against the deep neural network (DNN) classifiers by crafting adversarial examples without knowledge of the classifier training data or model. In paper [57], the author defined a black-box attack in which the attacker used the given data record to query the target model in the black-box observation.

Among the two observations discussed above, the adversary under white-box setting is the most knowledgeable setting and the black-box observation has the least back-

ground knowledge. Therefore, white-box attacks are more powerful than black-box attacks. However, black-box attacks cannot be substituted by white-box attacks because the former is easier to apply in practice. For example, in a machine learning as a service (MLaaS) system, the attacker always has no knowledge about the target model's internal information, they do not know the model algorithm, have no knowledge about the model structure or the model parameter, but just has the capability to query the target model, so comparing with white-box MIAs, the black-box attacks are the most reasonable observation.

3.6 Classification of Target Model: Discriminative Model vs. Generative Model

Generally, machine learning models include discriminative models and Discriminative models. Both of them suffer from MIA and there are many previous works related to this aspect. In this section, we introduce the membership inference attack against discriminative models and generative models.

Discriminative models. Given the feature (x) of a data point and the corresponding label, discriminative models attempt to predict feature x by learning a discriminative function (x, y); The function takes in input x and outputs the most likely label y. It means that the discriminative models discriminate between different kinds of data points. However, discriminative models are not able to explain how the datapoints might have been generated [22]. Membership inference against discriminative deep learning models has attracted many studies [1, 3, 4, 27, 38, 41, 57, 66]. This kind of target model can provide confidence value about the data point which would help infer out the membership of the training dataset.

Generative models. Generative models describe how does the data generated by learning the joint probability distribution of $p(X, Y)$, which gives a score to the configuration determined together by pairs (x, y) [22]. Compared with the discriminative model, the membership inference attack against generative models has been less well-studied. As the generator cannot directly return the confidence value about the overfitting of data records, it's more challenging for the adversary in this scenario. With the generator model widely used in many applications, such as [2, 17, 18, 35, 40, 63], membership inference attacks against the generative model gained researchers' increasing attention. In the work [22], the authors use generative models to learn information about the target generative model, thus created a local copy of the target model for membership inference. In paper [7], the author proposed a taxonomy of membership inference attack against generative adversarial networks (GANs).

3.7 Summary

The summary of MIAs with analytical factors, which mentioned above, is provided by Table 2. In the existing work, researchers do more research on black-box membership inference attacks than white-box one. Among them, in the research of white-box, it is not necessary to know the conditions for query ability and information about training dataset; The attacker even only needs to know the M_1 condition to successfully obtain the membership information of the training dataset. For the MIA scenario of the discriminator, the attacker needs information related to the model and training dataset to assist in the attack. When one kind of the model or training dataset is missing, the query ability of the target model is needed to supplement the information. The above conditions are not necessary in the attack against the generator. The attacker can realize the MIAs on the generator through one kind of condition among the model attributes and training dataset properties. In the previous work, there are more attacks on stand-alone machine learning than attack against federated learning scenarios. The existing attacks against federated learning scenarios are usually white-box attacks, and at the same time, both information related to the model and training dataset are required to complete the MIAs. While the assumptions for stand-alone attacks will be more flexible.

4 Defenses

In this section, we discuss factors which influence the MIA. Then, we survey the defense mechanisms employed by existing privacy-preserving achievements and IMA defensive methods. Based on different implementation techniques, we classify the defense strategies into three categories: Generalization techniques, cryptography methods, adversarial method. In addition, we introduce prior work with these categories.

4.1 Factors Influence MIA

The factors influence MIA means that have on the advantage of adversaries who attempt to infer specific facts about the data used to train machine learning models [66]. Shokri *et al.* [57] show that overfitting is a sufficient condition for MI attack. The result in [38, 45] reveals that even well-generalized machine learning models might leak much information about their training data. Thus, it means that overfitting provides more information than necessary for MIA [38, 66]. Long *et al.*'s work [38] demonstrates that some training instances have unique impacts on the learning models, which will cause MI attacks. Shokri *et al.*'s work [57] finds that besides overfitting, the structure and type of the model also contribute to the vulnerable to MIA. Nasr *et al.* [45] show that model structure, gradients, and training size can also impact the learning model.

Table 2: Summary MIAs with our analytical aspects

Previous works	AO	Prior Knowledge				TMT	TF
		Model		TD	QA		
Long et al.'s [18]	B	M_3	D_1	Q	D		S
Shokri's [19]	B	M_3, M_4	D_1	Q	D		S
Nasr's-1 [20]	W	M_1, M_2	D_2	N	D		F
Nasr's-2 [20]	W	M_1, M_2	D_2	N	D		F
ML-Leaks-1 [22]	B	M_3	D_2	N	D		S
ML-Leaks-2 [22]	B	N	N	Q	D		S
ML-Leaks-3 [22]	B	N	N	Q	D		S
LOGAN-1 [23]	W	M_1, M_2, M_4	N	N	G		S
LOGAN-2 [23]	B	N	D_1, D_2	Q	G		S
LOGAN-3 [23]	B	N	D_1, D_2	Q	G		S
Gan-Leaks-1 [24]	B	M_3	N	Q	G		S
Gan-Leaks-2 [24]	B	N	D_1	N	G		S
Gan-Leaks-3 [24]	W	M_1	N	N	G		S
Nasr's et al.'s [26]	B	N	D_2	Q	D		S
Truex et al.'s [27]	B	M_3	N	Q	D		S
Melis et al.'s[33]	B	M_4	D_1	N	D		F
Long et al.'s[31]	B	N	D_1, D_2	Q	D		S

AO: attack observation, TD: training dataset, QA: query ability, TMT: target model type, TF: target frame; M_1, M_2, M_3, M_4, D_1, D_2, Q are the 7 sub-classes of adversary's prior knowledge, which defined in table I; N: not need; W: white-box; B: black-box; D: discriminator; G: generator. S: stand-alone; F: federated learning.

4.2 Generalization Techniques

As overfitting is an important reason why machine learning models leak information about their training datasets, generalization techniques such as dropout [53, 56, 58] can help degrade overfitting and strengthen privacy guarantees in neural networks [28] by randomly dropping out connections between neurons. While model stack [53] suitable for all machine learning models, independent of the classifier used to build them. The paper [57] uses standard regularization to overcome overfitting in machine learning.

4.3 Cryptography Methods

Homomorphic encryption. He *et al.* [24] use homomorphic encryption to encrypt the input in the collaborative learning scenario, so the sensitive information will not be leaked. A drawback of homomorphic encryption is inefficiency [24].

Differential privacy. Differential privacy has been regarded as a strong privacy standard [9–13]. The paper [61] presents a differentially private GANs model which includes a Gaussian noise layer in the discriminator if a generative adversarial network to make the output and the gradients differentially private with respect to the training data. The paper [4] uses the differentially-private stochastic gradient descent algorithm (DP-SGD) to prevent memorization. Salem *et al.*'s work [52] adds noise to the posterior for each queried sample and also adds noise sampled from a uniform distribution to the posteriors. The result shows that the method drops the attack accuracy to a certain degree. Especially, adding noise is hard work against a multi-sample reconstruction attack. In [67], the researcher introduces a data obfuscation function and applies it to the training data before feeding them to the model training task. By doing so, sensitive information about both the properties of individual samples and the statistical properties of a group of samples will be hidden. Jia *et al.* [32] propose to add noise to each confidence score vector predicted by the target model to turn the confidence score vector into an adversarial example, which can mislead the adversary's classifier to make random guessing at member and non-member.

4.4 Adversarial Method

In [46], Nasr *et al.* put forward a Min-Max Game which designs an adversarial training algorithm that minimizes the prediction loss of the model as well as the maximum gain of the inference attacks. This strategy, which can guarantee membership privacy acts also helped to generalize the target model. Jia *et al.* [32] proposes a method based on adversarial examples to mislead the attack model. There are many methods to find adversarial examples [5, 19, 34, 39, 43, 44, 48]. These adversarial methods may be exploited as defense strategies in the future.

5 Future Research Direction

The current MIA methods have the following problems: On the one hand, building MIAs requires many preconditions, such as: Information about data, model or query ability, which is unreasonable in actual scenarios; On the other hand, the current defense methods cannot have protective effects on various MIAs. Therefore, in the future, we can study MIA in realistic scenarios, approach the real world by reducing assumptions, and study effective general protection frameworks for MIA to solve these problems. Considering the current challenges and existing solutions, we expect that the research of MIA will be advanced in the following aspects.

Membership inference attack against federated learning frame would attract more attentions of researchers. Along with the widely application of machine learning, for obtaining better performance of model training, the learning frame gradually changed from stand-alone learning to the collaborative learning. Thus, there are much more sensitive data that would be used as the federated training dataset, such as location data, personal medical records, personal characteristic data, healthcare data. Such sensitive data would increase the adversary's interests. To study the attack methods under this scenario, and devise defensive strategies to mitigation these vulnerabilities has academic and application values.

Threats based on membership inference attack would be raised. In paper [22], the author indicated that membership inference attacks often act as a gateway to further attacks. The attacker can firstly infers whether the target data is a part of the training dataset, and then link up with other attacks, (e.g. profiling, property inference, which leak additional information about the victim, or other further attacks. Hence, the subsequent attacks after the launching of MIA would be studied in the future.

Another valuable topic for research is to find out a fully effective defensive methodology to cope with different attack approaches. The application of MI methods in security defense scenarios will draw more attention. Shokri *et al.*'s work [56] uses the membership inference method as defense mechanism.

6 Conclusions

The study of Membership Inference attack(MIA) against machine learning is quite young field. This research direction has attracted attention of scholars and offers a number of opportunities for future exploration. For researchers just entering MI attacks and defenses against machine learning, we provided an in-depth introduction to this research field in its current state. For active researchers in the field, this paper not only provide a structured and comprehensive survey, but also as fundamental knowledge for the future researches in this area.

In this artical, we summarize the year-wise road map of MIAs against machine learning. and then, we construct a comprehensive framework to analyze the existing MIAs against machine learning systems, classifies the adversary's prior knowledge into seven sub-classes, overviews the factors that influence the attacks; Next we analyze the prior works with our framework, and give out a systematic comparison in Section 3.7. Further More, we characterizes existing defense mechanisms for MIA against machine learning in to three categories. Lastly, we give out the future research direction in this field.

Acknowledgments

This study was supported by a grant from the National High Technology Research and Development Program of China(863 Program)(No.2009AA01Z435).

References

[1] Y. Aono, T. Hayashi, L. Wang, S. Moriai, *et al.*, "Privacy-preserving deep learning: Revisited and enhanced," in *International Conference on Applications and Techniques in Information Security*, pp. 100–110, 2017.

[2] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, "Generalization and equilibrium in generative adversarial nets (gans)," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 224–232, 2017.

[3] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership privacy in microrna-based studies," in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pp. 319–330, 2016.

[4] N. Carlini, C. Liu, J. Kos, Ú. Erlingsson, and D Song, "The secret sharer: Measuring unintended neural network memorization & extracting secrets," *Machine Learning*, 2018. arXiv:1802.08232.

[5] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (SP'17)*, pp. 39–57, 2017.

[6] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[7] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-leaks: A taxonomy of membership inference attacks against gans," *Machine Learning*, 2019. arXiv:1909.03935.

[8] M. Y. Chen, C. C. Yang, and M. S. Hwang, "Privacy protection data access control," *International Journal Network Security*, vol. 15, no. 6), pp. 411–419, 2013.

[9] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, pp. 1–19, 2008.

[10] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, 2006.

[11] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, pp. 371–380, 2009.

[12] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[13] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60, 2010.

[14] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[15] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

[16] T. T. Gao, H. Li, and S. L. Yin, "Adaptive convolutional neural network-based information fusion for facial expression recognition," *International Journal of Electronics and Information Engineering*, vol. 13, no. 1, pp. 17-23, 2021.

[17] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, 2016.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Machine Learning*, 2014. arXiv:1412.6572.

[20] Google, *AI Platform*, 2020. (`https://cloud.google.com/ai-platform`)

[21] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *Computation and Language*, 2014. arXiv:1412.5567.

[22] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Evaluating information leakage of generative models using generative adversarial networks," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, 2017.

[23] Y. He, S. Rahimian, B. Schiele, and M. Fritz, "Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation," *Computer Vision and Pattern Recognition*, 2019. arXiv:1912.09685.

[24] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 148–162, 2019.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[26] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 4, pp. 232–249, 2019.

[27] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pp. 603–618, 2017.

[28] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density snp genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, 2008.

[29] M. S. Hwang, E. F. Cahyadi, S. F. Chiou, and C. Y. Yang, "Reviews and analyses the privacy-protection system for multi-server," in *Journal of Physics: Conference Series*, vol. 1237, pp. 022091, 2019.

[30] M. S. Hwang and I. C. Lin, "Introduction to information and network security (in chinese)," *Mc Grew Hill. In Taiwan*, 4, 2011.

[31] M. S. Hwang, C. H. Wei, and C. Y. Lee, "Privacy and security requirements for RFID applications," *Journal of Computers*, vol. 20, no. 3, pp. 55–60, 2009.

[32] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pp. 259–274, 2019.

[33] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, vol. 2, 2015. (`https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf`)

[34] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *Computer Vision and Pattern Recognition*, 2016. arXiv:1607.02533.

[35] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European Conference on Computer Vision*, pp. 702–716, 2016.

[36] Z. Li and Y. Zhang, "Label-leaks: Membership inference attack with label," *Machine Learning*, 2020. arXiv:2007.15528.

[37] Y. Long, V. Bindschaedler, and C. A. Gunter, "Towards measuring membership privacy," *Cryptography and Security*, 2017. arXiv:1712.09136.

[38] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "Understanding membership inferences on well-generalized learning models," *Computer Science*, 2018. arXiv:1802.04889.

[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *Machine Learning*, 2017. arXiv:1706.06083.

[40] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *Sound*, 2016. arXiv:1612.07837.

[41] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *IEEE Symposium on Security and Privacy (SP'19)*, pp. 691–706, 2019.

[42] Microsoft, *Microsoft Azure Machine Learning*, 2020. (`https://azure.microsoft.com/en-us/services/machine-learning/`)

[43] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773, 2017.

[44] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

[45] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE Symposium on Security and Privacy (SP'19)*, pp. 739–753, 2019.

[46] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, 2018.

[47] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of ACM on Asia Conference on Computer and Communications Security*, pp. 506–519, 2017.

[48] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy (EuroS&P'16)*, pp. 372–387, 2016.

[49] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock knock, who's there? Membership inference on aggregate location data," *Proceedings of the 25th Network and Distributed System Security Symposium*, 2017. arXiv:1708.06145.

[50] S. Rahimian, T. Orekondy, and M. Fritz, "Differential privacy defenses and sampling attacks for membership inference," in *PriML Workshop (PriML'19)*, vol. 13, 2019. (`https://priml-workshop.github.io/priml2019/papers/PriML2019_paper_47.pdf`)

[51] A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 5558-5567, 2019.

[52] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," *Cryptography and Security*, 2019. arXiv:1904.01067.

[53] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *Computer Science*, 2018. arXiv:1806.01246.

[54] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool," *Nature Genetics*, vol. 41, no. 9, pp. 965, 2009.

[55] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers," in *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, pp. 61–66, 2020.

[56] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1310–1321, 2015.

[57] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (SP'17)*, pp. 3–18, 2017.

[58] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[59] J. X. Tong, H. Li, and S. L. Yin, "Research on face recognition method based on deep neural network," *International Journal of Electronics and Information Engineering*, vol. 12, no. 4, pp. 182–188, 2020.

[60] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *The 25th USENIX Security Symposium*, pp. 601–618, 2016.

[61] A. Triastcyn and B. Faltings, *Generating Differentially Private Datasets Using Gans*, 2018. (`https://openreview.net/pdf?id=rJv4XWZA-`)

[62] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service,"

IEEE Transactions on Services Computing, 2019. DOI:10.1109/TSC.2019.2897554.

[63] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *Sound*, 2016. arXiv:1609.03499.

[64] C. H. Wei, M. S. Hwang, and A. Y. H. Chin, "A secure privacy and authentication protocol for passive RFID tags," *International Journal of Mobile Communications*, vol. 15, no. 3, pp. 266–277, 2017.

[65] C. H. Wei, M. S. Hwang, and A. Y. H. Chin, "Security analysis of an enhanced mobile agent device for RFID privacy protection," *IETE Technical Review*, vol. 32, no. 3, pp. 183–187, 2015.

[66] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE 31st Computer Security Foundations Symposium (CSF'18)*, pp. 268–282, 2018.

[67] T. Zhang, Z. He, and R. B. Lee, "Privacy-preserving machine learning through data obfuscation," *Cryptography and Security*, 2018. arXiv:1807.01860.

[68] G. Zhang, A. Zhang, and P. Zhao, "Locmia: Membership inference attacks against aggregated location data," *IEEE Internet of Things Journal*, pp. 1-1, 2020. DOI: 10.1109/JIOT.2020.3001172.

Biography

Yang Bai is the Ph.D candidate of the University of Electronic Science and Technology of China (UESTC), China. Her research interest include machine learning, cloud computing, and information security.

Ting Chen received the Ph.D degree from the University of Electronic Science and Technology of China (UESTC), China, 2013. Now he is a professor with UESTC. His research interest include on blockchain, smart contract, program analysis, and information security.

Mingyu Fan received her Ph.d in Southwest Jiaotong University, and worked as a post-doc in Tsinghua University. Now she is a professor at the University of Electronic Science and Technology of China(UESTC), China. Her research interest is mainly in the area of communication engineering, computer science, and information security.