Defenses to Membership Inference Attacks: A Survey

LI HU, ANLI YAN, HONGYANG YAN, JIN LI, TENG HUANG, YINGYING ZHANG, and CHANGYU DONG, Guangzhou University, China
CHUNSHENG YANG, National Research Council Canada, Canada and Guangzhou University, China

Machine learning (ML) has gained widespread adoption in a variety of fields, including computer vision and natural language processing. However, ML models are vulnerable to membership inference attacks (MIAs), which can infer whether access data was used in training a target model, thus compromising the privacy of training data. This has led researchers to focus on protecting the privacy of ML. To date, although there have been extensive efforts to defend against MIAs, we still lack a comprehensive understanding of the progress made in this area, which can often impede our ability to design the most effective defense strategies. In this article, we aim to fill this critical knowledge gap by providing a systematic analysis of membership inference defense. Specifically, we classify and summarize the existing membership inference defense schemes, focusing on optimization phase and objective, basic intuition, and key technology, and we discuss possible research directions of membership inference defense in the future.

CCS Concepts: • **Security and privacy** → *Privacy protections;*

Additional Key Words and Phrases: Membership inference, privacy defense, privacy attack, Machine learning

ACM Reference format:
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. *ACM Comput. Surv.* 56, 4, Article 92 (November 2023), 34 pages.
https://doi.org/10.1145/3620667

1 INTRODUCTION

Recent advances in complex machine learning models and computing infrastructure, coupled with the availability of massive data, have facilitated the application of machine learning in everyday life. For example, in computer vision, machine learning has gained widespread adoption in face recognition, object detection, image classification, and so on. The success of ML has recently propelled leading internet companies such as Google and Amazon to adopt **machine learning as a service (MLaaS)**, which provides training services for data owners to train ML models for different

This work was supported by National Natural Science Foundation of China (Nos. 62102107, 62002074), National Natural Science Foundation of China for Joint Fund Project (No. U1936218) and Basic Innovation Project for Full-time Postgraduates of Guangzhou University (No. 2021GDJC-D18).
Authors' addresses: L. Hu and J. Li (Corresponding author), Institute of Artificial Intelligence, Guangzhou University, Guangzhou, Guangdong, China and The Guangdong Provincial Key Laboratory of Blockchain Security, Guangzhou University, Guangzhou, Guangdong, China; e-mails: hl_27@e.gzhu.edu.cn, jinli71@gmail.com; A. Yan, H. Yan, T. Huang, Y. Zhang, and C. Dong, Institute of Artificial Intelligence, Guangzhou University, Guangzhou, Guangdong, China; e-mails: anli_yan2021@163.com, hyang_yan@gzhu.edu.cn, huangteng1220@buaa.edu.cn, zhangyingying@e.gzhu.edu.cn, changyu.dong@gmail.com; C. Yang, Institute of Artificial Intelligence, Guangzhou University, Guangzhou, Guangdong, China and National Research Council Canada, Canada; e-mail: chunsheng.yang@nrc.gc.ca.
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 the author(s) 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 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0360-0300/2023/11-ART92 $15.00
https://doi.org/10.1145/3620667

applications. These models are then either published as a prediction **application programming interface (API)** and accessed in a black-box fashion or a set of parameters in a white-box fashion. Despite its success being overshadowed by different study domains, ML models trained with sensitive data remember the sensitive information of the data and may pose a privacy threat to the data owner when they are published and used. Typically, the data used to train ML models often contain sensitive user information such as clinical records, location traces, personal photos, and so on. Prior work unveiled that ML models are vulnerable to various privacy inference attacks, e.g., model extraction attack, attribute inference attack, and membership inference attack [57].

In this investigation, we focus on **membership inference attack (MIA)**. Through this attack, the attacker can infer the member information about the training set of the target model, which may cause serious privacy leakage issues. For example, if a machine learning model is trained on data collected from patients with a disease, then the attacker can immediately know the health status of the victim by knowing whether the victim's data is part of the model training data. MIA for machine learning is first proposed by Shokri et al. [64]. They can judge whether the access data belongs to the training data of the model only according to the prediction vector output by the model. Subsequent related works have also gradually expanded the work of MIA, which has been successfully applied in many fields, such as biomedicine. Considering the privacy harm caused by MIAs, a large number of studies have proposed different **m**embership **i**nference **d**efenses **(MIDs)** from different angles according to the reasons for the successful implementation of MIAs to resist MIAs while maintaining the utility of the target ML model.

There are many surveys that summarize the works of membership inference attack [3, 25, 40, 42, 70, 76, 81]. Among them, Reference [25] is the first very comprehensive survey on **m**embership **i**nference **(MI)**. The authors classify the existing attack methods from various perspectives, discuss the working principles of MIAs, and introduce the existing defense methods to mitigate MIAs. Furthermore, it summarizes the open-source implementation of most existing evaluation metrics and datasets, discusses the challenges of MI from both attack and defense aspects, and points out potential research opportunities for future research. Despite this work providing a very comprehensive overview of MI, it does not focus on MID and does not conduct a comprehensive investigation into MID. We hope to further expand the overview of MID works on this basis. In addition, considering that most of the current defense works are carried out for the classification model in computer vision, we will take the classification model as the main line to analyze. Our ultimate goal is to systematically analyze the principles of various MID works so relevant workers can more clearly grasp the progress of MID works and the direction that can be further studied. The main contributions of this article are:

— Based on the attacker's knowledge, we conduct a comprehensive review of membership inference attacks in computer vision domain classification tasks, explore the evolution of such attacks in other domains, and gain a fresh perspective on the underlying principles of membership inference attacks.
— We establish a comprehensive and systematic framework for defending against membership inference attacks. Specifically, we first sort out the defense technology of membership inference and then delve into the three key dimensions of membership inference defense: pre-training, training, and inference phases of the target model. Through this analysis, we examine the existing defense mechanisms for membership inference attacks in computer vision domain classification tasks. Furthermore, we classify the defense mechanisms used in other fields and study the underlying principles of membership inference defense.
— Finally, we propose promising future research in the field of membership inference defense and recognize that there is vast potential for further development in this area.

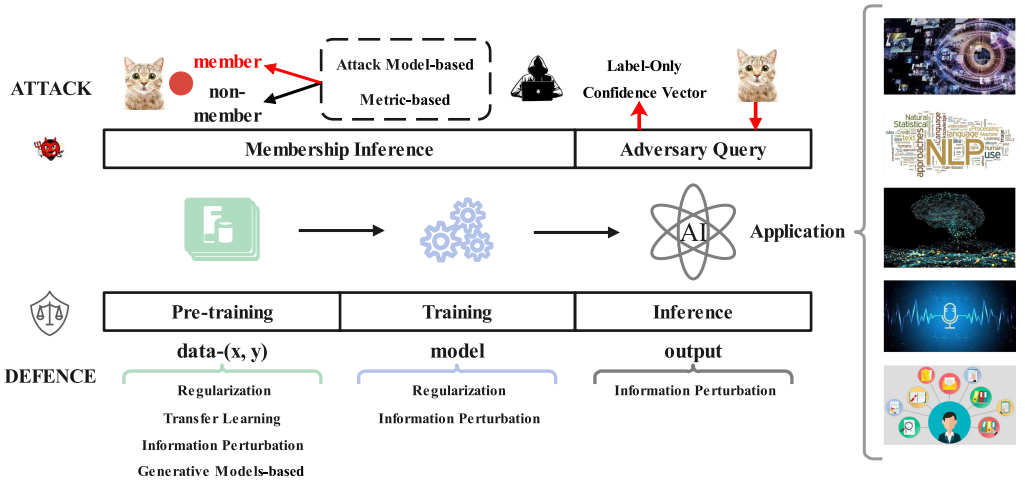

Fig. 1. The main content structure of this article about MIA and MID.

In this article, we put together the state-of-the-art and most important works on membership inference but focus on membership inference defenses. The main content structure of this article is shown in the Figure 1, and the corresponding organization structure is as follows: Section 2 formally defines the membership inference attack and combs the adversary knowledge possessed by the attacker in detail. In Section 3, we sort out the existing membership inference attacks against the visual domain classification model, briefly describe the work of other domain tasks, and finally summarize the principles of membership inference attacks. Section 4 classifies and describes the membership inference defense technology, which can be roughly divided into: Regularization, Transfer Learning, Information Perturbation, Generative Models-based. In Section 5, we review the membership inference defense works of the existing visual domain classification model according to the defense phase and describe it in detail in combination with the defense technology described in Section 4. In Section 6, we classify the membership inference defense works in other domain tasks. In Section 7, we outline the principles of membership inference defense. Finally, we look forward to the future research direction of membership inference defense and summarize the work of this article in Sections 8 and 9, respectively.

2 BACKGROUND

In this section, we first give the definition of membership inference attack based on ML models, then give a brief overview of the information that attackers may have during the execution of MIAs.

2.1 Definition of Membership Inference Attack

We formalize the definition of membership inference attack: Given a query instance x, x accesses the target model $f_{target}(\theta)$ trained on the training set D^{train}_{target}, the attacker can obtain the output $f_{target}(x;\theta)$ of the target model and judge whether the instance x belongs to the training set D^{train}_{target}. The definition points out that the membership inference attack focuses on the member information of x in D^{train}_{target} rather than the content of x and the attacker wants to know whether a specific x is in D^{train}_{target} rather than the whole D^{train}_{target}. These differences distinguish membership inference attacks from other privacy attacks in Reference [57].

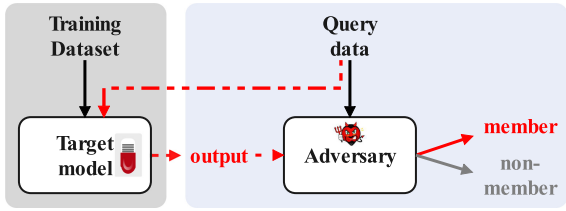

Fig. 2. The workflow of the MIA.

Figure 2 describes the workflow of membership inference attack. From the perspective of the attacker, the attacker accesses the prediction API provided by the machine learning service provider to obtain the prediction result/output $f_{target}(x;\theta)$ corresponding to the query data x. Then, the attacker combines with any public knowledge or background knowledge $\mathcal{K}$ about the target model $f_{target}(\theta)$ and builds an attack algorithm $\mathcal{A}$. Then, the attack algorithm $\mathcal{A}$ can be used to launch a membership inference attack in real time. According to the definition in Reference [95], we describe the membership inference attack as a binary classification task. The attacker's goal is to classify whether an instance x is used to train the victim model. Formally, we describe it as:

$$\mathcal{A} : (x, f_{target}(\theta), \mathcal{K}) \rightarrow \{0, 1\}, \tag{1}$$

where 0 means x is not a member of $f_{target}(\theta)$'s training dataset D^{train}_{target} and 1 otherwise.

According to the adversary knowledge about the target model and the main feature for executing the attack, the attacker's strategy for executing the membership inference attack is different. We describe it in the following sections.

2.2 Adversary Knowledge

Adversary knowledge refers to the information about the target model that an attacker can access, as well as the knowledge that an attacker has about the query data. According to the different adversary knowledge possessed by the attacker, the attacker's ability is also different. To better sort out the work of existing membership inference attacks, we introduce the adversary knowledge of the attacker in detail according to the following categories: Data Knowledge, Model Knowledge, Training Knowledge, and Output Knowledge. As shown in the Figure 3, depicting the adversary knowledge from these four dimensions can systematically describe the adversary ability.

(1) **Data Knowledge.** The data knowledge is described in two parts: one is the knowledge of the target model training dataset D^{train}_{target}, and the other part is the knowledge of a query data x from attacker. The knowledge about the target model training data that is accessible to the attacker may be part of the training data or the distribution of the target model training data. In most MIA works, it is assumed that the attacker can obtain the distribution knowledge of D^{train}_{target}, i.e., the attacker can have a dataset $D^{'}$ that is identically distributed with D^{train}_{target}. Whether the dataset $D^{'}$ intersects with D^{train}_{target} determines the difficulty of the attack. If there is an intersection, then it is assumed that the attacker has some data of D^{train}_{target}. Generally, it is assumed that there is no intersection between $D^{'}$ and D^{train}_{target}. For the knowledge of the query data possessed by the attacker, it needs to be analyzed from the perspective of whether the attacker has used the label of the query data. Normally, we use the unlabeled data to access the prediction API provided by the machine learning service provider and obtain the corresponding output results. However, the attacker may also use the data with label to access the target model just to obtain the member information of the query data.

Fig. 3. The illustration of adversary knowledge.

(2) **Model Knowledge.** For attackers, the available model knowledge can also be divided into two parts: one is the model structure knowledge of the target model, and the other is the model parameter knowledge of the target model. The structural knowledge of the target model includes the type of neural network, the number of layers, the type of activation function, and so on. With this structural knowledge of the target model, attackers can train a model that achieves similar performance to the target model by leveraging additional data. Generally, the attacker can only access the target model through the API of the target model, i.e., the attacker does not know the model parameter of the target model but can deduce the model parameter knowledge through the model stealing attack. According to whether the attacker knows the model parameters of the target model, we can divide the attacks into black box MIAs and white box MIAs. The white box MIA assumes that the attacker knows the structure and parameters of the target model, which means that the attacker can obtain the output information of data at any layer of the target model.

(3) **Training Knowledge.** Training knowledge refers to the knowledge about training algorithm, including the type of optimization algorithm, the number of training steps, the setting of optimization algorithm, and so on. These training knowledges reveal how the target model is trained. In most MIA settings, attackers can retrain models with same performance to the target model if they know training knowledge and model structure knowledge.

(4) **Output Knowledge.** The output knowledge is the result of the attacker accessing the prediction API provided by the machine learning service provider. For attackers, the available output knowledge can be divided into full output knowledge, partial output knowledge, and label-only knowledge. The full output knowledge means that the attacker can obtain the complete confidence value vector corresponding to the access data, while the partial output knowledge only displays the partial confidence value vector, such as the maximum three confidence values. Label-only knowledge is an extreme case. The attacker can only know

the predicted label corresponding to query data, which provides the attacker with the least output information.

In most MIA works, it is usually assumed that the attacker possesses the distribution knowledge of the target model training dataset, model structure, training knowledge, and output knowledge of the target model. The difficulty of MIA execution can be divided according to the amount of knowledge possessed by the attacker. In the next section, we divide and briefly describe the existing related attack schemes from difficult-to-execute to easy-to-execute.

3 ATTACKS OF MEMBERSHIP INFERENCE

In this section, we classify and describe the MIA works in the field of image classification according to the output knowledge obtained by the attacker after accessing the API. Then, we introduce the MIA works in other fields, i.e., text and voice. Finally, we analyze the principles of existing membership inference attacks.

3.1 Attack Approaches of Membership Inference

Hu et al. [25] described the existing MIA works according to whether the attack model is trained during the attack process, which can be divided into neural network-based attacks and metric-based attacks. To better understand the attacker's capabilities, we classify and describe the MIA works in the field of image classification according to the output knowledge obtained by the attacker after accessing the API. Table 1 presents the papers analyzed in terms of output knowledge, attack algorithm, model parameter knowledge, and the knowledge of attacker's query data.

(1) Attacks with label knowledge in black-box scenarios

Label-based attack means that the attacker only gets the label, $\hat{y} = f_{target}(x; \theta)$, after accessing the API of target model. This is most likely to happen in real-world scenarios, such as when the model provides a face recognition service that will tell you directly who the person is, and an attacker can easily access this information to perform MIAs.

The first label-based MIA is proposed by Yeom et al. [95], whose main idea is that if the target model can correctly predict an input instance (x, y), then the attacker infers the instance (x, y) as a member, otherwise, the attacker infers it as a non-member. The intuition of the attack is that the target model can correctly predict its training dataset, but the generalization of the target model may be poor on the test dataset, so it can use this difference to perform MIAs. This attack is generally regarded as a simple attack, and the subsequent relevant literatures [12, 39, 66] have taken it as a baseline to compare the performance of their proposed attacks. In addition, referring to the correctness of the model prediction, Choquette-Choo and Li [12, 41] designed a new measurement method. For a given instance (x, y), they tried to measure the distance from the model decision boundary $dist_{f(\theta)}(x, y)$, when $\hat{y} \neq y$, and let $dist_{f(\theta)}(x, y) = 0$. When $\hat{y} = y$, the method of adversarial example generation should be used to find the adversarial example (x', y') with the smallest Euclidean distance from the instance (x, y), now $y' \neq y$, and let $dist_{f(\theta)}(x, y) = \sqrt{(x - x')^2 + (y - y')^2}$. When $dist_{f(\theta)}(x, y) > \tau$, we judge (x, y) to be a member, where τ can be obtained by constructing a shadow model that simulates the behavior of the target model.

Moreover, Choquette-Choo et al. [12] also designed a label-based MIA using neural network. They used disturbed version samples $(x_i, y)(i = 1, 2, \ldots, n)$ of instance (x, y) to extract more subtle member information. During the attack, (x_i, y) is first obtained by means of data augmentation. Then, x_i is input into the target model to obtain its corresponding prediction label $\hat{y}_i$. When its prediction label is equal to its corresponding label (i.e., $\hat{y}_i = y$), the disturbed version data signal b_i is marked as 1 (i.e., $b_i = 1$), otherwise, b_i is marked as 0 (i.e., $b_i = 0$) to obtain an n-dimensional disturbed signal vector $(b_1, b_2, \ldots, b_n)$. When we train the attack model, n-dimensional

disturbance signal vector is used as the data feature for the attack model. When the corresponding data (x, y) of the disturbed signal vector is training data, the n-dimensional disturbed signal vector $(b_1, b_2, \ldots, b_n)$ is marked as 1 (i.e., member). When the corresponding data (x, y) of the disturbed signal vector is test data, the n-dimensional disturbed signal vector $(b_1, b_2, \ldots, b_n)$ marked 0 (i.e., non-member). The generation of training data for the attack model still needs the assistance of the shadow model. The main idea of this label-based MIA using a neural network is to extract fine-grained information about classifier decision boundaries by combining multiple queries on disturbed version data. By evaluating the robustness of the target model to disturbing data with different inputs, the data with high robustness can be inferred as members.

Recently, Zhang et al. [104] pointed out that label-based MIAs rely on the different robustness of members and non-members in the target model. Its main goal is to find the disturbance that can distinguish members and non-members, which is independent of the task of finding the minimum disturbance. If different adversarial disturbance directions are used, then the gap between members and non-members may be different. Therefore, they proposed a new scheme to adjust the adversarial disturbance direction through label smoothing to enhance the existing label-based MIAs. In fact, we can see that although label-based MIAs acquire little knowledge after accessing the API of the target model, they require the attacker to have labeled query data to launch the attack.

(2) Attacks with partial output knowledge in black-box scenarios

MIAs with partial output knowledge in the black-box scenario is the classical MIA proposed by Shokri et al. [64]. The attacker can access the target model and obtain the first k larger values of the predicted confidence $P_x(Top\,k)$ $(P_x = f_{target}(x;\theta))$ of the model output. Here, the attacker can execute MIAs with partial output knowledge without the label of query data.

In Reference [64], Shokri et al. trained an attack model to distinguish between member and non-member by using $P_x(Top\,3)$ as the input of the attack model. To construct the attack model, multiple shadow models are trained to imitate the behavior of the target model. The main idea is that the more shadow models, the more training knowledge is provided for the attack model, and the more accurate the attack model is. What needs to be noted is the dataset (including the training dataset and the test dataset) of the shadow model and the training dataset of the target model should be subject to the same distribution without intersection. Dataset between shadow models can intersect. Subsequently, Salem et al. [59] made improvements on this basis. They relaxed the requirements on the structure of the shadow model, i.e., the data of the training shadow model and the number of shadow models. They only need to train one shadow model whose structure is arbitrary, and the training data of the shadow model does not need to obey the same distribution as the training data of the target model. Under this assumption, the primary objective of their attack is to utilize the shadow model for capturing the membership states of the data in the training dataset, rather than replicating the behavior of the target model.

Meanwhile, Salem et al. [59] also designed a method to determine whether the access data is a member by using the maximum value of the model output confidence. The experiment shows that the maximum confidence value can achieve very high attack performance. The main idea is that if the maximum predicted confidence of instance x is greater than the preset threshold τ, the attacker deduces instance x as a member; otherwise, the attacker infers that x is a non-member. The intuition of the attack is that the target model is trained by minimizing the predicted loss of the training data, which means that the maximum predicted confidence should be close to 1 for the training data. The selection method about threshold τ is to generate a sample of random points in the feature space of the target data points and input it to the target model to obtain the corresponding output prediction vector. The random point can be considered as a non-member point, and the first t bit of the maximum value in the prediction vector can be a good threshold. In

their work, they chose to use a single threshold for all class labels. Subsequently, Song et al. [66] refined this attack by setting different thresholds for the different class labels.

In addition, the attack based on prediction differential distance was initially proposed in Reference [29]. Experiments show that this attack method can achieve better attack performance and defeat the most advanced defense system. The main idea is to move the instance (x, y) from the member dataset to the non-member dataset. If the difference distance calculated after the data in the two sets are input into the target model becomes smaller, then the instance (x, y) is presumed to be a member data, otherwise, the attacker infers it as a non-member data. The intuition of the attack is that for two disjoint sets, moving data from one side to the other affects the spatial distance between the two sets.

(3) Attacks with full output knowledge in black-box scenarios

Once the attacker obtains the full predicted confidence values P_x ($P_x = f_{target}(x; \theta)$), they can use some statistical information from the target model to execute MIAs, such as the average loss of model training data, prediction entropy, and prediction difference distance. In this way, MIAs can be successfully executed without relying on the attack model. At the same time, attackers can also use unlabeled query data to execute attacks, but Song et al. [66] showed that labeled query data can achieve stronger attack effects.

Yeom et al. [95] proposed that the attacker could judge whether the query data belonged to a member by calculating the predicted loss value of the access data and showed that the attack only needed less computing resources and background knowledge to achieve the same performance as the neural network-based attack proposed by Shokri et al. [64]. The main idea is that if the predicted loss of instance (x, y) is less than the average loss of all training samples, then the attacker infers that (x, y) is a member, otherwise, the attacker infers that (x, y) is a non-member. The intuition of attack is that the target model trains by minimizing the predicted loss of its training sample, so the predicted loss of the training sample should be smaller than the input loss not used in the training process.

The difference in prediction entropy distribution between training data and test data was initially presented in Reference [64] to explain the existence of member privacy risks. The main idea is that if the predicted entropy of instance (x, y) is less than the preset threshold, the attacker classifies (x, y) as a member, otherwise, the attacker infers it as a non-member. The intuition of this attack is that the prediction entropy distribution between training data and test data is very different, and the prediction entropy of the target model for its test data is usually larger than training data. Subsequently, literature [59] proved the effectiveness of using prediction entropy to carry out MIA. On this basis, Song et al. [66] proposed a scheme based on entropy with different thresholds for different class label. Besides, they also proposed a modified prediction entropy attack. They believed that prediction entropy does not contain any information about real labels, which may lead to the misclassification of members and non-members.

In addition, Carlini et al. [4] pointed out that directly setting the threshold based on loss attack implicitly assumes that the loss of all samples is *a priori* loss of equal proportions. The inclusion or exclusion of a sample has a similar effect on the model as any other sample. The only confident assessment that can be made of this attack is that the sample of high losses is non-members, but the judgment of members is invalid. Therefore, they proposed to perform membership inference attack as a likelihood ratio test. By calculating per-example hardness scores and estimating likelihood to predict whether a sample is a member of the training set. Subsequently, this work was improved by Wen et al. [85], who used adversarial tools to directly optimize query samples to obtain discriminative and diverse queries, which were used to achieve more precise membership inference than this method. Based on the hypothesis testing framework designed by Ye et al. [94], a

model of indistinguishability game was proposed, and the interpretation of the attack success rate of different game is provided. In Reference [45], Liu et al. were the first to exploit the member information of the target model throughout the training process to improve the attack performance. They showed that the sample loss evaluated on the target model during different training periods can be used to distinguish members from non-members, and their attack performance achieves a high true positive rate at a low false positive rate.

(4) Attacks in the white-box scenarios

In the white-box MIAs, the attacker can obtain all knowledge except the training data of the target model and can completely contact the target model. In other words, for instance x, the attacker can not only obtain all the information corresponding to its prediction confidence P_x, but also can see the intermediate calculation process of the input target model. In general, white-box membership inference attacks are stronger than black-box membership inference attacks. Because in the former context, the attacker knows more information about the target model. However, Sablayrolles et al. [58] showed that the white-box inference attack can not provide more information than the black-box setting. By assuming the distribution of parameters, the author deduced the optimal strategy of membership inference. It is proved that the best attack only depends on the loss function, so the black-box attack is as good as the white-box attack.

In Reference [51], Nasr et al. implemented a white-box MIA, which uses the gradient calculated by the middle layer of the model, the output of the middle layer, the confidence of the model output, and the label of its corresponding output to distinguish training samples from non-training samples. This work showed that the inference scheme obtains higher attack accuracy than black-box MIA. However, in this scheme, it is assumed that the attacker knows part of the data of the training dataset, which is different from the general assumption of MIAs. Leino et al. [39] considered relaxing this assumption to achieve an effective white-box attack without accessing the training data of the target model. Recently, on the basis of label-based MIAs, Grosso et al. [17] performed white-box access to the target model and judged the member attributes by calculating the amount of perturbation required to change the predicted results of the access data. In addition, Cohen et al. [14] introduced a new white-box MIA, which can be applied to any ML model. The core idea is that training samples have a direct impact on the loss of test samples. To quantify this effect, they used influence function to determine how the data points in the training set affect the prediction of the target model for a given test sample. This measure quantifies the impact of the small weight rise of a specific training point on the loss of a test point in the empirical error of the target model. Given a sample point, they can infer whether the sample belongs to the training set by calculating its self-influence function score and querying the prediction label output from the target model.

3.2 Membership Inference Attacks on Different Domains

Artificial intelligence is a branch of computer science. It attempts to understand the essence of intelligence and produce a new intelligent machine that can respond in a similar way to human intelligence. Since the birth of artificial intelligence, its theory and technology have become more and more mature, and its application fields have also expanded. However, the privacy security of artificial intelligence has also been put forward in various fields, such as membership inference attacks. Starting from the initial visual image classification task, a growing number of works have studied membership inference attacks in other application fields. Recently, Hu et al. [25] classified the work of MIAs in various fields, such as vision, natural language processing, audio, graph, and recommendation systems. This article mainly analyzes the defense works of membership inference. Therefore, we only give a brief overview of the attack works in other fields.

In the field of computer vision, at present, many MIAs focus on image classification models. Relevant works also analyze the membership inference risk of Pruning Neural Networks, Model

Table 1. Summary of Papers on MIA in the Field of Image Classification, Including Information of their Assumptions about Output Knowledge (with Label Knowledge/with Partial Output Knowledge/with Full Output Knowledge), Attack Algorithm (Neural Network based/Metric based), Model Parameter Knowledge (Black-box/White-box) and Whether Query Data has Labels

Ref.	Year	output knowledge			attack algorithm		model parameter knowledge		whether query data has label	
		with label knowledge	with partial output knowledge	with full output knowledge	neural network based	metric based	black-box	white-box	Yes	No
[64]	2017		•	•	•	•	•		•	•
[95]	2018	•		•		•	•		•	
[59]	2018		•		•	•	•			•
[51]	2019			•	•			•	•	
[58]	2019			•		•	•		•	
[39]	2020				•			•	•	
[41]	2020	•				•	•		•	
[12]	2021	•			•	•	•		•	
[29]	2021		•			•	•		•	
[66]	2021			•		•	•		•	
[4]	2021			•		•	•		•	
[94]	2021			•		•	•		•	
[104]	2022	•				•	•		•	
[17]	2022	•				•		•	•	
[14]	2022	•				•		•	•	
[45]	2022			•		•	•		•	
[85]	2022			•		•	•		•	

Including Information of their Assumptions about Output Knowledge (with Label Knowledge/with Partial Output Knowledge/with Full Output Knowledge), Attack Algorithm (Neural Network based/Metric based), Model Parameter Knowledge (Black-box/White-box) and Whether Query Data has Labels

Explanations, Algorithmic Fairness, Adversarial Examples, Deep Transfer Learning, Causal Learning, and other models. In addition, according to different scenarios, attacks against federated learning scenarios have emerged one after another. With the attack on the classification model, the subsequent related works also gradually attack the generation model and image segmentation model.

With the gradual development of MIA in the field of computer vision, relevant works have also carried out MIA risk analysis for natural language processing domain, graph domain, audio domain, and recommender system domain. The MIA work of natural language processing domain mainly focuses on the tasks of text classification, text generation, word embedding, and masked language. It can be seen that as long as the model has access and output, there may be MIA privacy risk. With the hot development of graph neural network, relevant researches are prompted to the exploration of its privacy and security issues. In graph domain, the researches on MIA mainly focus on knowledge graphs, node classification, and graph classification tasks. In graph domain, the researches on MIA mainly focus on the speech recognition task. In recent years, recommender systems have achieved good performance and become one of the most widely used web applications. However, recommender systems are often trained on highly sensitive user data, so potential data leakage of recommender systems may lead to serious privacy concerns. Zhang et al. [100] quantified the privacy disclosure of recommendation systems from the perspective of MIAs.

With the rapid development of artificial intelligence in various fields, when it comes to the model of privacy data training, there are naturally privacy and security issues. Research on members' privacy risks has also appeared in various fields. Since the attack works have appeared one after another, the defense works against the attacks are inevitable. We will describe the defense works in various fields in detail in Section 4 and Section 5.

3.3 Mechanisms of Successful Execution of MIAs

Why deep learning models are susceptible to MIAs is the basis for implementing MIAs and defending against MIAs. We can summarize the reasons for the successful of MIAs as follows: overfitting of the target model, the unique impact of the training set, and other properties of the target model.

(1) **Overfitting of the target model.** At present, most works on MIA are based on target model overfitting. When the target model is in the overfitting state, its performance in the training set and test set is different. By means of this difference, the attacker accesses the target model and determines whether the query data is a member. Yeom et al. [95] theoretically analyzed the connection between overfitting and MIAs. The experiment shows that overfitting is a sufficient but unnecessary condition for the successful of MIAs. In a nutshell, there are other reasons besides overfitting causes that make the target model vulnerable to MIAs. Experiments in Reference [64] show that overfitting models have significant differences in the probability distribution of the output of member data and non-member data. Furthermore, Different types of machine learning models and different datasets also show different vulnerabilities to MIAs. Leino et al. [39] showed a new view on how overfitting leads to information leakage of membership. They believe that the training data memorized by the overfitting classification model is not only reflected in the output behavior of the model, but also in the inner layer of the model.

(2) **The unique impact of the training set.** In Reference [46], Long et al. showed that the non-overfitting model is also vulnerable to MIAs. In the generalized model, the information leakage from member is caused by the unique impact of specific instances in the training set on the learning model. This unique impact affects the output of the model about single or multiple inputs, provides useful information for predicting models from other instances, and brings noise with unique features. The model regularization method that restrains overfitting can reduce the noise introduced by training examples, but their unique impact cannot be completely eliminated, especially those essential for the prediction power of the model. Noise addition techniques based on the concept of differential privacy can reduce the impact of each training instance, but at the same time it will reduce the prediction accuracy of the model. When the training set is not representative, that is, the distribution of the training set is different from the test set, the model developed by the training set cannot fit the dataset well for prediction, so the training set of the model can be easily distinguished from the test set and the MIA can be successful.

(3) **Other properties of the target model.** Tan et al. [68] showed that the generalization performance of the model can be improved by increasing the trained model parameters or the number of epochs trained. However, it is at the cost of reducing privacy, which means that the success of MIAs is not only caused by the overfitting of the model, but also related to other attributes of the model. In addition, some literatures [16, 76] explored the robustness of different types of classical ML models with respect to MIAs and the impact of hyperparameters on attack effectiveness. The experiment shows that most classical machine learning models are not vulnerable to MIAs, but tree-based models are vulnerable to MIAs. Therefore, it is necessary to consider the difference of MIA privacy risk in different models. And some

literatures [72, 107] explored the impact of model fairness on MIA privacy risk and showed that the privacy risk from MIA is different under different fairness.

4 DEFENSE TECHNOLOGIES OF MEMBERSHIP INFERENCE

To better understand membership inference defense works, before sorting out the existing defense works, we first describe the relevant defense technologies. The existing defense technologies can be summarized into the following categories: Regularization, Transfer Learning, Information Perturbation, and Generative Models-based.

4.1 Regularization

Regularization technology is the general term for a series of technologies to improve the model performance, which reduces the degree of overfitting to improve the generalization ability of the model. Many papers [39, 59, 64, 95] pointed out that overfitting of the target ML model is a major factor in the success of MIAs, hence regularization technology can certainly defend against MIAs [37, 50, 59, 82].

There are many regularization techniques, which can be divided into two categories: one is in the process of training, such as L2-norm regularization, early-stopping, dropout, and adversarial training, and the other is in the process of data processing, such as data augmentation, model stacking, label smoothing, and so on. However, regularization technologies are often difficult to achieve. The regularization method changes the internal parameters of the target model, which also affects the output distribution of the target model. To not damage the usability of the model, a reasonable regularization setting is required. Note that data augmentation usually refers to the random generation of new training features, such as rotation, clipping, deformation, color transformation, mix-up, and so on. However, data augmentation can also be achieved by adding noise. In this article, we classify this method into noise perturbation method.

4.2 Transfer Learning

Transfer learning [84] is a learning process that uses the similarities between different fields, tasks, or distributions to apply the knowledge learned in the old field to the new field. It can alleviate the problem that there are few or no labels for tasks in new fields. In this process, the old field and the new field are not required to be subject to **independent and identically distributed (i.i.d.)**. To protect data privacy, relevant works [20, 53, 80, 98] combined knowledge transfer and differential privacy. For membership inference attacks, knowledge transfer can be used to protect member privacy of target data.

Recent studies [13, 27, 28, 32, 62, 69, 106] showed that knowledge transfer can be used to train the model with member privacy. By reducing the access to the target data and replacing the target data with similar but different data, it prevents attackers from inferring the privacy of target data members and provides a better tradeoff between members' privacy and classification accuracy. According to the types of transfer learning, we introduce the existing transfer learning defense schemes from the perspectives of model transfer and feature transfer.

Knowledge Distillation. Knowledge distillation **(KD)** [22] uses the output of large teacher model to train smaller student model. It allows smaller student model to have similar accuracy to their teacher model [15]. With the idea of knowledge distillation, some papers [13, 32, 62, 69, 106] proposed different defense schemes, and experiments show that the idea of knowledge distillation can achieve a better tradeoff between model privacy and utility.

Domain Adaptation. Domain adaptation **(DA)** [84] is a representative approach in transfer learning. This method can transfer knowledge from the source domain to different but related target domains and promote the implementation of tasks in the target domain by extracting the

shared representation of the source domain dataset and the target domain dataset. The shared representation has the basic common features of the two datasets. For example, the knowledge extracted from the cat and dog image dataset (source domain) collected on Instagram is used to improve the classification task of cat and dog images (target domain) intercepted in animated movies. On the basis of domain adaptation, Huang et al. [27, 28] proposed related defense methods and effectively reduced the privacy risk of MIAs.

4.3 Information Perturbation

Information perturbation is prevalently leveraged to resist membership inference attacks, which protects sensitive information by adding customized noise. Existing works based on information perturbation are broadly categorized into three classes: differential privacy, output perturbation, and data perturbation.

Differential Privacy. Differential privacy [18] **(DP)** expresses the deterministic output as a probability by adding noise perturbations to the real data, and such randomization perturbations process does not cause serious damage to model utility. Specifically, given two adjacent datasets D and D' that differ by only one piece of data, the results of a mechanism with differential privacy for D and D' should ideally be indistinguishable and the utility of the randomization mechanism should not be severely compromised. Differential privacy is defined as follows:

Definition 1 (Differential Privacy). A random mechanism M preserves (ϵ, δ) differential privacy with domain D and range R. For any two neighboring datasets $d, d' \in D$ and any output $S \subseteq R$, we have:

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\epsilon} \Pr\left[\mathcal{M}\left(d'\right) \in S\right] + \delta, \tag{2}$$

where $\mathcal{M}(d)$ and $\mathcal{M}(d')$ represent the output of algorithm $\mathcal{M}$ on datasets d, d', respectively. $\Pr[\cdot]$ is the output probability of the algorithm $\mathcal{M}$. ϵ is the privacy budget, which is used to control the level of privacy protection. The smaller the ϵ is, the stronger the privacy protection capability is. δ is another privacy budget, as introduced by Dwork et al. [19], representing the probability that the tolerable privacy budget exceeds ϵ. If δ equals 0, then we call $\mathcal{M}$ satisfies ϵ-differential privacy.

Many papers [9, 10, 12, 29, 30, 33, 39, 49, 50, 52, 56, 73–75, 88, 103, 105] analyzed the defense capability of differential privacy against MIAs. Differential privacy provides theoretical guarantees for protecting member privacy for individual samples. It can be used as a defense mechanism against MIAs on different task models, regardless of whether the adversary is in a black-box or white-box setting. Although the privacy protection capability of differential privacy has wide applicability and effectiveness, its drawback is that it is difficult to achieve the tradeoff between the model utility and privacy. This problem has been mentioned in many papers [33, 39, 56].

Output Perturbation. A membership inference attack is one that can be successfully implemented using only the model's output. It is common to believe that the attack can be successfully defended by directly influencing the model output. Randomly altering the model output could, however, inevitably reduce the model's usefulness. Therefore, it is necessary to reasonably perturb the model output to protect the privacy of members as much as possible without affecting the utility of the model. In [34, 90], the authors successfully weakened the performance of MIAs by perturbing the confidence scores. This method is simple to implement and does not require retraining the target model. However, to not affect the utility of the target model, we need to use a suitable method to find the appropriate noise, and this method is invalid for attackers who only access the output label.

Data Perturbation. In the work of defending against membership inference, its purpose is to protect the member information of target model training data. The more direct way is to hide the member information of the training data by adding perturbations to the data. Unfortunately,

the flaw of this method is that it has an uncontrollable impact on the utility of data. To reduce the loss of model utility as much as possible, it is necessary to set the noise function carefully. In addition, we can make specific modifications to the training data for key features of the target model, making it difficult for an attacker to distinguish the distribution of prediction vectors for member and non-member data.

4.4 Generative Models-based

The goal of the generative model is to generate a batch of data. Assuming that the true distribution of these data is $p_{data}(x)$, the generative model wants to learn a distribution $p_{model}(x)$ to estimate $p_{data}(x)$. From this model, we can sample or generate new data identical to the distribution $p_{data}(x)$. The existing generation models can be divided into the following three types: **variational autoencoder (VAE)** based on the idea of variation, **generative adversarial network (GAN)** based on the perspective of the adversarial game, and **energy-based model (EBM)**.

With the rapid development of generative models, the samples generated by these models basically follow the same distribution as the original training data and achieve good diversity. In view of the development advantages of the generation model, relevant works [7, 10, 26, 54, 73] generated replaceable data for the training data of the target model with the generation ability of the generation model to reduce the leakage of member information of the training data of the target model. Extensive experimental results show that an alternative dataset can help decouple the direct relationship between the original training data and the output of the target model. Meanwhile, the overall characteristics can still be retained to train the effective model.

5 DEFENSES OF MEMBERSHIP INFERENCE AT DIFFERENT PHASE

To better understand the root causes of the effectiveness of various defense schemes in the field of image classification, we decided to break down each phase of the target model to further analyze the critical role of existing defense works. Note that all defense works also correspond to the defense technologies classified in Section 4. In this section, we analyze defense works from three stages: the pre-training phase of the target model, the training phase of the target model, and the inference phase of the target model. Moreover, we combine defense technology and phase analysis of defense works to better understand the principles of MID works.

5.1 Defenses at Pre-training Phase of Target Model

The defense works at the pre-training phase is to process the training data feature and label, i.e., (x, y), to defend against MIA. We describe the defense works carried out in this stage from three aspects: preprocessing feature, preprocessing label, and preprocessing feature and label. More details of the summary are listed in Table 2.

5.1.1 Preprocessing Feature. The defense work on preprocessing feature mainly realizes the protection of member information by perturbing or replacing data features. It is worth noting that after processing feature, the coupled label does not change. Existing related works can be divided into four categories according to the technologies used: (1) data deformation with regularization technology; (2) to perturb the characteristics of feature with the help of transfer learning technology; (3) to add noise to the data; (4) or use the generative ability of GAN to generate a surrogate dataset for the training set of the target model.

(1) With the help of Regularization

To obtain a distribution that is indistinguishable from the test data, Yin et al. [96] combined data of a certain class with data of other classes in proportion and maintained the label of this class to

Table 2. Summary of Papers on MID during the Pre-training Phase of Target Model

Phase	Optimization Object	Defense Technology	Defense Method	Year	Ref.
Pre-training	Preprocessing feature	Regularization	Data Augmentation	2020	[37]
				2021	[36, 96]
		Transfer Learning	Domain Adaptation	2021	[27, 28]
		Information Perturbation	Data Perturbation	2018	[101]
		Generative Models-based	GAN	2018	[73]
				2021	[54]
				2022	[26]
			EBM	2021	[7]
			AE/VAE+DP	2018	[10]
	Preprocessing label	Regularization	Label Smoothing	2020	[37]
				2021	[36]
			Model Stacking	2018	[59]
		Transfer Learning	Knowledge Distillation	2021	[13, 69, 106]
				2022	[32]
	Preprocessing feature and label	Regularization	Data Augmentation - Mix-up	2020	[40]
				2021	[11]
		Transfer Learning	Knowledge Distillation	2019	[62]
				2022	[47]
		Information Perturbation	Data Perturbation	2019	[79]
		Generative Models-based	GAN	2021	[83]
			VAE+DP	2022	[91]

form a new dataset. In this way, the feature of target data can be perturbed, and then the defense against MIAs can be achieved. In addition, in References [36, 37], Kaya et al. analyzed a variety of data augmentation methods and showed that the appropriate use of data augmentation methods that randomly generate new training features can increase the accuracy of the model and reduce the privacy risk of members.

(2) With the help of Transfer Learning

One of the cores of using transfer learning to defend against MIAs is domain adaptation. On the basis of domain adaptation, Huang et al. [28] proposed DAMIA, leveraging domain adaptation as a defense to counter MIAs. There are two optimization objectives in the domain adaptive training process. One is to minimize the classification loss of the target domain. The classification loss ensures fine classification performance by updating the weights of the feature extractor and classifier. The other is to ensure that the characteristics of the source domain and the target domain are similar. Three-domain adaptation approaches can be used, namely, Discrepancy-based approach, Adversarial-based approach, and Retractive-based approach. With Discrepancy-based approach, DAMIA's optimization formula is as follows:

$$Loss = Loss_C\left(X_L, y\right) + \lambda D^2_{\mathcal{MMD}}\left(X_S, X_T\right), \tag{3}$$

where $Loss_C\left(X_L, y\right)$ denotes classification loss on the available labeled data, X_L, and the ground truth labels, y, and $D^2_{\mathcal{MMD}}\left(X_S, X_T\right)$ denotes the distance between the source data, X_S, and the

target data, X_T. The hyperparameter λ determines how strongly we would like to confuse the domains. Domain adaptation requires a dataset of the source domain and a dataset of the target domain. To protect the privacy of the target domain, we need to find a source domain dataset that is different from but related to the target domain. Note that, before training, the label of the target domain data needs to be removed to synchronize with the training process of domain adaptation. Therefore, shared representations do not cause privacy leakage of images collected from animated movies.

Considering the difficulty of source domain data collection, Huang et al. [27] subsequently proposed NoiseDA, which uses the noise-adding mechanism to construct a feature based on the target domain feature to replace the source domain feature. This carefully crafted feature is different from the target feature, so NoiseDA can be effectively applied to defend against MIAs. The intuition of the domain-based adaptation defense scheme is that the two datasets involved are mixed and confounded. In this way, MIAs can be successfully resisted, and the generated shared representation can also well solve the task corresponding to the sensitive dataset, since the shared representation contains features from the sensitive dataset. Moreover, this defense overhead is trivial, because no additional phases/algorithms are involved once the model is released.

(3) With the help of Information Perturbation

Perturbing data features is a more direct defense method, but the performance loss caused by perturbation needs to be strictly controlled. To protect the member information of the target model training data, a fuzzy function was introduced in Reference [101] and applied to the training data. The fuzzy function adds random noise to the training data and enhances the dataset with updated samples. This results in sensitive information about the attributes of a single sample or the statistical attributes of a group of samples being hidden. Meanwhile, by designing a fuzzy function, the training model of a fuzzy dataset can still achieve high accuracy.

(4) With the help of Generative Models-based

The core of generative model-based technology is to use the synthetic data of generative model as the training dataset of target model. Similar to other defense techniques, this approach aims to safeguard privacy without compromising utility. In terms of privacy assurance, the original training data is replaced by the generated data of generative model, which avoids the attacker's contact with the original data from the data source. In terms of utility assurance, the existing generative model is used to generate high-quality synthetic data, that is, to ensure diversity and fidelity of the synthetic data.

In Reference [26], Hu et al. proposed a defense scheme DMIG from the source of privacy leakage. This scheme first uses the data generated by a GAN to replace the training data of the target model and then utilizes the synthetic data to train a surrogate model that can substitute for the target model. The surrogate model can protect the target model training data from membership inference attacks. To ensure the utility of the data generated by GAN trained with a small dataset, this article adopts different GAN models and special training techniques for different types of data. By evaluating the defense performance of the existing attack schemes on different datasets and comparing with other MID works, it shows that DMIG provides a significantly better tradeoff between member privacy and classification accuracy. In addition, Paul et al. [54] generated a substitute dataset with the help of GAN, which allows medical data sources (such as hospitals) to provide a replacement dataset, i.e., synthesized from original images to external agents (modelers). It obtains better defense performance combined with other defense methods. Furthermore, the authors also proposed a new metric, the P1 score, to measure the tradeoff between utility and privacy. Aiming at the security of complex real-world data, Triastcyn et al. [73] used GAN to

generate privacy-protected artificial data samples and introduced a new framework for statistically estimating the potential privacy loss of published data. Experiments show that this method can generate high-quality label data and can be used to successfully train and verify the supervision model. Finally, it is proved that this method significantly reduces the vulnerability of such models to model inversion attacks.

With the generating power of EBM, Chen et al. [7] proposed a **joint energy-based model (JEM)**. It uses the implicit generation ability of the model to sample the data that are independent and identically distributed with the training data. This scheme uses newly generated data and retraining or fine-tuning skills to obtain updated models that can suppress the attacker's membership advantage to a negligible level and maintain acceptable accuracy of the classifier. To improve learning efficiency and generate data with a privacy guarantee and high practicability, Chen et al. [10] proposed a **differentially private autoencoder-based generative model (DP-AuGM)** and a **differentially private variational autoencoder-based generative model (DP-VaeGM)**. Experiments show that DP-AuGM can effectively defend against model inversion, membership inference, and GAN-based attacks, and DP-VaeGM is also robust to membership inference attacks. The above works are to defend against MIAs by generating data with no additional processing of its coupled label. It can be seen that defense based on this method needs to carefully adjust the details in the training process to ensure the utility.

5.1.2 Preprocessing Label. The work of preprocessing labels mainly realizes the protection of member information by perturbing and replacing the label vector. It is worth noting that after processing the label, the coupled data feature is unchanged. Existing related works can be divided into two categories: (1) to soften the label vector of data with the help of regularization technology; (2) to extract data features onto soft labels with the help of transfer learning technology.

(1) With the help of Regularization

Label smoothing [67] is another regularization method that flattens one-hot labels. During deep neural network training, we usually use one-hot labels to calculate the cross-entropy loss. However, this way has the problem of prompting the model optimization process to only consider the loss of correct label positions of training samples, while ignoring the loss of incorrect label positions. This training approach leads to a model that performs very well on the training set but poorly on the test set. Label smoothing adds noise to the one-hot label, as depicted in Equation (4), thereby providing a degree of error tolerance ε to label y to mitigate the extremeness of the training optimization objective. Kaya et al. [37] applied this technology to defend membership inference attacks, and the experimental results show that label smoothing has a certain defense effect.

$$P_i = \begin{cases} 1, \text{ if } (i = y) \\ 0, \text{ if } (i \neq y) \end{cases} \Rightarrow P_i = \begin{cases} (1-\varepsilon), \text{ if } (i = y) \\ \frac{\varepsilon}{K-1}, \text{ if } (i \neq y) \end{cases}. \tag{4}$$

From another perspective, model stacking is an ensemble learning framework that organizes multiple weak classifiers in a hierarchical structure. Specifically, the first layer is generally composed of multiple base classifiers, whose input is the original dataset, and then the output of the first layer classifier is used as the input of the second layer classifier. Note that the output of the classifier in the last layer is taken as the output of the final model. For the basic model of the first layer, it is better to use a strong model. The number of models should not be too small to ensure that the model has higher performance. In contrast, the base model of the last layer can use a simple classifier, which prevents the model from overfitting. In Reference [59], Salem et al. used a two-layer model stacking architecture to protect the member privacy of the target model and

showed through experiments that model stacking can effectively defend against MIAs. However, this method is time-consuming, because multiple base models need to be trained to obtain the final target model. The main idea of this defense method is to train multiple base models with disjoint original data so the final target model integrates the advantages of multiple base models. It is equivalent to training different parts of the target model, which help to reduce the overfitting of the target model.

(2) With the help of Transfer Learning

Knowledge distillation, another special case of transfer learning, is used as the technical support for preprocessing label. Zheng et al. [106] proposed two algorithms, namely, **complementary knowledge distillation (CKD)** and **pseudo complementary knowledge distillation (PCKD)**. In CKD, the reference data for knowledge distillation are all from private training data, but their soft labels are generated by different teacher model. Note that the reference data at this time must be non-training data of the teacher model. However, it is time-consuming and expensive for CKD to train a set of teacher models. If the time cost is reduced by selecting a small k, then it leads to the loss of the utility of the teacher model. To alleviate this issue, Zheng et al. [106] further proposed PCKD. PCKD reduces the amount of training data of each teacher model and uses the average value of model output as the soft label of reference data. Although PCKD and CKD have different ways of generating soft labels for reference data, the soft target knowledge of transmitted data comes from their complementary sets. In order to avoid that each teacher model cannot obtain good utility due to too small training set, PCKD also adopts pre-training technique. Therefore, PCKD relaxes the limitation of complementary sets and can train models with better usability.

Subsequently, Tang et al. [69] put forward a framework similar to CDK, SELENA. SELENA has two main components. The first component is called Split-AI. Split-AI first divides the training data into random subsets, and the union of each subset is the entire dataset. Then, Split-AI trains a model on each subset. In the prediction phase, Split-AI always selects a model that the query data is its non-member data to predict. The authors demonstrate that the Split-AI architecture can defend against a large number of member inference attacks, but is difficult to defend against label-only attacks. Therefore, the author designed the second component Self-Distillation in SELENA to prevent stronger label-only attacks. First, the Self-Distillation component queries Split-AI with training samples to obtain the coupled prediction vector. Then, these prediction vectors are used as the soft label of the training set to train a protected model. In the inference stage, the protected model only needs to perform a calculation on the sample of each query, so its overhead is much lower than the Split-AI component. In addition, the protected model can not only protect the classical single query MIAs, but also prevent the adaptive MIAs.

At the same time, Jarin et al. [32] proposed a new MIA defense, MIAShield. The key point of the working principle of MIAShield is to weaken the strong member signal from the target sample by excluding the target sample in advance during prediction without affecting the utility of the model. For this purpose, the authors used the oracle model based on the confidence of the members and the learning of the members to evaluate and exclude a set of members first. In practical application, MIAShield divides training data into disjoint subsets and trains a model set using each subset. The discontinuity of subsets ensures that a target sample belongs to only one subset, isolating the samples and facilitating the realization of preemptive exclusion targets. Experiments show that MIAShield effectively alleviates MIAs (close to random guess). Compared with the state-of-the-art defense, it achieves a better tradeoff between privacy and utility and maintains the flexibility to adaptive opponents.

5.1.3 Preprocessing Feature and Label. In the work of feature and label preprocessing, it is necessary to adjust the label vector corresponding to the data feature after processing to achieve the

performance of defending against MIAs. The existing related work can be divided into four categories: (1) to update the label vector according to the processed data after processing the data features, i.e., it depends on regularization technology; (2) to use an auxiliary dataset combined with transfer learning technology to extract the knowledge learned by the model into soft labels; (3) to perturb important features of training data and update coupled labels, it relies on perturbation technology; (4) to generate surrogate data by using GAN, and to update the labels of surrogate data according to its peculiarities.

(1) With the help of Regularization

In Reference [40], Li et al. first showed the simple numerical relationship between the generalization gap (the difference between training accuracy and test accuracy) and the vulnerability of the classifier to MIAs and suggested that the defense against MIAs can be realized by deliberately reducing the training accuracy to match the accuracy of the test. The author achieves this by regularizing the training loss function and introducing a new regularizer. The new regularization function is a permutation invariant function (set function) that will force the classifier to train to match the output experience distribution corresponding to the training data and the validation data. To measure the difference between the two empirical distributions, the authors used **maximum mean difference (MMD)**. However, using MMD alone tends to reduce the accuracy of training and testing. To solve this problem, the author combined MMD with mix-up technology. Experiments show that the proposed method can not only achieve the effect of resisting MIAs, but also ensure that the test accuracy is not affected. Moreover, it can be trained on large neural networks, and there is no extra calculation cost in inference stage.

Then, Chen et al. [11] proposed **Enhanced Mixup Training (EMT)**, an improved defense scheme. Because mix-up technology retains the linear relationship of samples, it is still susceptible to the influence of MIAs. Therefore, the author improved it by recursively mixing the training data in the training process. Specifically, the EMT benefits from recursive hybrid training, which uses the designed enhanced hybrid items to mix training data in the training process. Compared with the existing defense, EMT fundamentally improves the accuracy and generalization of the target model, thus effectively reducing the risk of MIAs. In theory, EMT corresponds to a specific type of data-adaptive regularization, which leads to better generalization.

(2) With the help of Transfer Learning

In Reference [62], Shejwalkar and Houmansadr proposed a defense technology, called **Distillation for Membership Privacy (DMP)**. DMP requires a private training dataset and an unlabeled reference dataset. DMP first trains an unprotected teacher model $f(\theta_{\text{up}})$ and uses it to label data instances in the unlabeled reference dataset. Then, DMP selects data instances with low prediction entropy in the labeled reference dataset to train the protected model $f(\theta_{\text{p}})$ that is used to provide the service, and its optimization objective is as follows:

$$\theta_{\text{p}} = \underset{\theta_{\text{p}}}{\operatorname{argmin}} \frac{1}{|X_{\text{ref}}|} \sum_{(\mathbf{x},\hat{\mathbf{y}}) \in \left(X_{\text{ref}}, f(X_{\text{ref}}, \theta_{\text{up}})\right)} \mathcal{L}_{\text{KL}}(\mathbf{x}, \hat{\mathbf{y}}), \tag{5}$$

$$\mathcal{L}_{KL}(\mathbf{x}, \hat{\mathbf{y}}) = \sum_{i=0}^{c-1} \hat{\mathbf{y}}_i \log\left(\frac{\hat{\mathbf{y}}_i}{f(\mathbf{x}_i; \theta_{\text{p}})}\right). \tag{6}$$

The intuition in choosing the protected model training dataset is that the samples are easily classified and not significantly influenced by the members of the private training dataset. DMP avoids direct access to private training datasets with the help of protected models, thus significantly reducing the leakage of member information. Experimental results show that DMP achieves

a state-of-the-art balance between member privacy and classification model accuracy. In addition, Mazzone et al. [47] showed that not all knowledge distillation algorithms can effectively defend against MIA. In the knowledge distillation algorithm without adjusting temperature parameters and Kullback-Leibler optimization, the author combined confidence score masking to achieve the tradeoff between privacy and utility flexibly in various datasets for both black-box and white-box MIAs. However, Jagielski et al. [31] showed that the membership of the training data of the unprotected model could still be inferred through the access to the protected model, so it can be seen that the defense performance of DMP may still need to be improved.

(3) With the help of Information Perturbation

To resist MIAs against MLaaS, Wang et al. [79] proposed a framework, MIASec, that can ensure the indistinguishability of training data. The core idea of MIASec is to reduce the dynamic range of important features in training data. Specifically, MIASec reduces the difference between the model results of the training data and the test data by modifying the important eigenvalues in the training data, thus effectively protecting the training data while maintaining the stability of the model accuracy. A large number of experiments using real data on the machine learning model trained by offline neural networks and online MLaaS show that MIASec can effectively defend against membership inference attacks.

(4) With the help of Generative Models-based

In Reference [91], Yang et al. designed a **Privacy-Preserving Generative Framework (PPGF)** against MIAs, which generates synthetic data through VAE to meet differential privacy requirements. Note that they are working directly with the raw data rather than adding noise to the model output or tampering with the training process of the target model. Specifically, first, the source data is mapped to the latent space through the VAE model to obtain the latent code. Then the latent code is processed with noise to meet the measurement of privacy. Finally, the synthetic data is reconstructed by VAE's decoder. Moreover, during training, they used latent codes and a couple of labels to train a classification model to label synthetic data. Experimental results show that the machine learning model trained with the newly generated synthetic data can effectively resist MIAs and maintain high utility. In addition, Webster et al. [83] used GANs to generate the surrogate data of target model training data and used the target model for annotation to construct a new dataset to train the surrogate model.

5.2 Defenses at Training Phase of Target Model

During the training process of the target model, the loss function or model parameters can be adjusted to defend against MIAs. In general, regularization techniques can be applied to improve the generalization ability of the target model. In addition, **differential privacy (DP)** technique can also be used to perturb member signals by adding noise to the gradient during training. According to the optimization objective and the defense technology employed, this article categorizes and describes the defense works performed during the model training phase. More details of the summary are listed in Table 3.

5.2.1 Adjust the Loss Function of the Model. In the defense work of adjusting the model loss function, regularization can be implemented by introducing the idea of adversarial training, adding a common regularization term, or the optimization objective can be redesigned according to the defense goals. Among them, L_2-norm is one of the most commonly used regularization methods. Some papers [12, 64, 97] used L_2-norm standard regularization to penalize large model parameters, adding $\lambda \sum_j \omega_j^2$ to the loss function of the model, where ω_j is the parameter of the model. The

Table 3. Summary of Papers on MID during the Training Phase of Target Model

Phase	Optimization Object	Defense Technology	Defense Method	Year	Ref.
Training	Loss Function	Regularization	Adversarial Training	2018	[50]
				2021	[24]
			L2-Regularization	2017	[64]
				2020	[37, 97]
				2021	[12]
			Optimization Training	2021	[6]
				2022	[89]
	Model Parameter	Regularization	Dropout	2020	[37]
			Pruning	2021	[82]
	Gradient of the Loss Function	Information Perturbation	DPSD	2016	[1]
				2018	[10, 56, 88]
				2019	[33, 75]
				2020	[30, 99]
				2021	[55]
			SGLD	2020	[86]

regularization effect can be adjusted by changing the value of λ. The larger λ is, the more obvious the regularization effect during training.

In Reference [50], Nasr et al. studied the min-max privacy game between optimization of target model and MIAs, namely, **adversarial regularization (AR)**, which is equivalent to adding a new regularization term to the training process. Specifically, while training the target model, they also optimize the attack model. When updating the target model, we aim to minimize the prediction loss and MIA performance of the model. When updating the attack model of MIA, the goal is to maximize the attack performance of MIA, which is formalized as follows:

$$\underbrace{\min_{f}(L_D(f) + \lambda \underbrace{\max_{h} \mathrm{G}_{f,D,D'}(h)}_{\text{optimal inference}})}_{\text{optimal privacy-preserving classification}}, \tag{7}$$

where L is the loss function of target model f on dataset D, and G is the gain function. The internal maximization is to find the strongest attack model h for given target model f on member dataset D and non-member dataset D'. The external minimization is to find the strongest defensive target model f for given h, hoping to find the balance point of min-max game. The parameter λ controls the importance of optimizing classification accuracy and membership privacy. Experimental results show that AR can make the target model effectively defend against MIAs with negligible damage in classification performance. Since the mechanism can effectively prevent the model from overfitting, the prediction distribution of the model can not be distinguished between the training data and the non-training data to effectively defend against MIAs.

In addition, Hu et al. [24] pointed out that AR is an adversarial training algorithm, which is essentially the same as training the generator of GAN. Many variants of GAN can generate higher-quality samples and provide more stable training than GAN. However, it has not been investigated whether the ideas of these variants of GAN can improve the effectiveness of AR. To this end, an **enhanced adversarial regularization (EAR)** based on **Least Square GANs (LSGANs)** is proposed. The experimental results show that EAR is superior to AR, which provides stronger defense capability while maintaining the same prediction accuracy as the classifier to be protected.

Recently, Chen et al. [6] proposed a defense scheme called RelaxLoss, which has an easier-to-learn objective based on a training framework. In RelaxLoss, a repeated training strategy is run

to balance privacy and utility. It consists of two steps: (1) If the model is poorly trained, i.e., the current loss is greater than the target average α, then we run a normal gradient descent step; (2) Otherwise, gradient ascent or posterior planarization step is used. In this process, instead of pursuing to minimize the training loss of the target model to zero, the authors relax the average loss of the target to a threshold no lower than α. As long as the average loss of the current batch is less than α, the gradient rising step is adopted. However, when the target loss is relaxed, it may lead to incorrect prediction. To solve this problem, the author encouraged a large gap between the prediction scores of the ground truth class and other classes by flattening the target post score of the non-ground truth class. From another perspective, Xu et al. [89] developed a lightweight and fine-grained neuron-level regularization that simultaneously guided and coordinated final output neurons and hidden neurons (or intermediate features) to produce output confidence score distributions that are indistinguishable between the training set and the test set.

5.2.2 Adjust the Number of Parameters in the Model. The work of adjusting the number of model parameters is mainly to protect the member information by reducing the complexity of the model through regularization techniques, which can reduce the degree of overfitting of the model and make the model learn the general characteristics of member data. Common methods are dropout [23] and model parameter pruning [102].

In Reference [59], Salem et al. evaluated the effectiveness of using dropout in the input and hidden layers of neural networks to defend against MIAs. Experimental results show that dropout enables the model to defend against MIAs, but its defense ability is weak and depends on the dropout rate. A large dropout rate reduces the success rate of MIAs, but also reduces the utility of the model. Therefore, it is necessary to choose an appropriate dropout rate to balance the utility of the target model and the privacy of the training data. In addition, spatial dropout is a variant of dropout method proposed by Tompson et al. [71] in the field of images. The normal dropout randomly sets some elements to zero, while the spatial dropout randomly sets some regions to zero. Kaya et al. [37] evaluated the impact of dropout and spatial dropout on MIAs. They found that spatial dropout can reduce model overfitting to a greater extent, improve the generalization ability of the model more effectively, and thus can more effectively defend MIAs.

In Reference [82], Wang et al. also noticed that DNN is vulnerable to MIAs due to over-parameterization, so they proposed a new defense against MIAs called MIA-Pruning. MIA-Pruning is a pruning algorithm that finds subnets from a fully over-parameterized random network to optimize data privacy and model efficiency. By using MIA-Pruning, the performance of the new model is comparable to the original model in the case of greatly reducing the parameters, and it can effectively defend MIAs. Formally, for the model f, the optimization objective of MIA-Pruning can be formulated as follows:

$$\begin{aligned} &\underset{\{\mathbf{W}_i\},\{\mathbf{b}_i\}}{\operatorname{argmin}} \mathcal{L}\left(f\left(\{\mathbf{W}_i\},\{\mathbf{b}_i\};x\right),y\right) \\ &\quad \text{s.t. } \mathbf{W}_i \in \{\mathbf{W}_i \mid \operatorname{card}\left(\mathbf{W}_i\right) \le n_i\} \\ &\qquad \{n_i\} = \underset{\{n_i\}}{\operatorname{argmin}} \max_{f_A} G_{f(\{\mathbf{W}_i\},\{\mathbf{b}_i\})}\left(f_A\right), \end{aligned} \tag{8}$$

where $\{\mathbf{W}_i\}$ and $\{\mathbf{b}_i\}$ are the weights and biases of each layer, and $\mathcal{L}\left(f\left(\{\mathbf{W}_i\},\{\mathbf{b}_i\};x\right),y\right)$ is the loss of the classification model f. $\operatorname{card}(W_i)$ is the cardinality of weights in each layer, which returns the number of non-zero weights. n_i is the desired number of non-zero weights for each layer, which regularizes the strength of compression. Experimental results show that MIA-Pruning outperforms DP in defending against MIAs. DP may significantly impairs model accuracy, whereas the loss of model accuracy due to MIA-Pruning is negligible. Moreover, the combination of MIA pruning and min-max game can further protect the privacy of the model.

5.2.3 Affect the Gradient of the Loss Function. The defense works of affecting the gradient of the loss function mainly utilize information perturbation techniques. Adding noise to the gradients of the training process to achieve DP is a more general defense against MIAs.

In general, DP can achieve excellent privacy guaranteeing, but at the same time, it will seriously damage the utility of the model. This type of approach is usually implemented through **DPSGD (Differential Privacy Stochastic Gradient Descent)**, a differential privacy optimizer proposed by Abadi et al. [1]. It ensures that the model training process is differentially private by clipping gradients and adding noise. Taking Reference [56] as an example, Rahman et al. systematically studied the performance of MIAs against differential privacy models and showed that differential privacy models can achieve the promise of privacy protection against powerful attackers with low model efficiency. However, when they provide a model with acceptable utility, the model exposes vulnerability to MIAs. Some literatures [30, 33, 75, 99] have also analyzed the impact of different factors on the privacy-preserving capability of DP and shown that some properties of the dataset, such as bias or data correlation, play a key role in determining the effectiveness of DP as a MIAs privacy-preserving mechanism.

In Reference [55], Rahimian et al. showed that because DPSGD adds noise at each step of training, the speed of model training becomes slower, and the privacy budget ϵ also gradually increases to a large value with the increase of training rounds. Therefore, they proposed DP-logits, which only adds DP noise to the logits of the sample prediction output. Experimental results show that DP-Logits can effectively resist MIAs, and the required privacy budget ϵ is lower than DPSGD. Meanwhile, in the work of exploring the vulnerability of machine learning to MIAs, Chen and Xie et al. [10, 88] evaluated the effectiveness of DPSGD as a defense mechanism and showed that MIAs can be more effectively defended when generative models are used in combination with DP. In addition, Wu et al. [86] established a theoretical framework to analyze the information leakage of the model trained by **SGLD (Stochastic Gradient Langevin Dynamics)**. In this framework, the authors demonstrated that for a model trained with SGLD, the member privacy leakage can be bounded within the desired range of a uniform constant and observed that SGLD can prevent DNN from overfitting in some cases. The key idea of SGLD is to apply stochastic optimization to Langevin dynamics, which is achieved by injecting appropriate Gaussian noise into the gradient estimates of mini-batch samples in the training dataset.

5.3 Defenses at Inference Phase of Target Model

The defense works at the inference phase of the target model are mainly to use information perturbation technology to perturb the confidence vector of the model output so member information can be hidden. More details of the summary are listed in Table 4.

5.3.1 Adjust the Confidence Vector of the Model Output. Some works have explored how to defend against MIAs by adjusting the confidence vector output from the model. It is worth noting that this method does not affect the utility of the model, but can only defend against MIAs that depend on the confidence vector output from the model. There are two methods to resist MIAs in this way. One is to add adversarial noise to the confidence vector, the other is to use adversarial optimization.

Jia et al. [34] proposed the first defense against black-box MIAs with a formal utility loss guarantee, MemGuard. The working process of MemGuard can be roughly divided into two stages. In the first phase, MemGuard looks for a well-crafted noise vector and converts the confidence vector into a adversarial example, which causes the attack model to fail to distinguish between members and non-members. In the second stage, MemGuard adds a carefully selected noise vector with a certain probability to a confidence vector satisfying a given effective loss budget. Note that MemGuard does not modify the target model, but only adds noise to the confidence score vector predicted by the target model and ensures that the addition of noise does not affect the

Table 4. Summary of Papers on MID during the Inference Phase of Target Model

Phase	Optimization Object	Defense Technology	Defense Method	Year	Ref.
Inference	Confidence Vector	Information Perturbation	Output Perturbation - Adversarial Noise	2019	[34]
				2020	[90]
				2022	[43]
			Output Perturbation - Adversarial Optimization	2020	[92]
				2022	[93]

classification of the model, which can effectively reduce the success rate of MIAs based on the prediction confidence distribution under the black box to the level of random guess. However, Song et al. [66] evaluated the defense performance of MemGuard in metrics-based MIAs and found that a model using MemGuard defense still had a high MIAs success rate. In Reference [90], Xue et al. also proposed an adversarial disturbance defense method, AEPPT. Unlike MemGuard, AEPPT uses a gradient-based approach to counter disturbances and adds them to the prediction of the target model by multiplying the disturbance by the random step size. As a result, AEPPT can defend against more powerful attackers than MemGuard, with stronger security, wider versatility, and faster generation of adversarial predictions.

In addition, Yang et al. [92] proposed a purification framework to defend against inference attacks by purifying predictive scores. The framework takes the predicted score of the target model as input to generate a purified version of the output and makes it meet two defense objectives: (1) to prevent model inversion attacks and (2) to prevent MIAs. The intuitionistic significance of this purification framework is that it reduces the discreteness of the confidence score vector predicted by the target classifier on both members and non-members. It helps to reduce the sensitivity of the prediction to the changes of input data and reduces the resolution of the confidence score vector between members and non-members. Note that while reducing its dispersion, the purification framework shows a negligible distortion to the original confidence score, which preserves useful information for the prediction. Recently, Yang [93] proposed a modified Purifier consisting of a confidence reformer and a label swapper, which comprehensively studies MIA from the perspectives of individual shape, statistical distribution, and prediction label. Additionally, to alleviate the prediction difference between training samples and non-training samples, Liu et al. [43] applied order-preserving and utility-preserving obfuscation to the prediction vector, which can guarantee the indistinguishability of prediction vector without affecting the prediction performance and can not recover the original prediction vector for well-informed attackers.

6 MEMBERSHIP INFERENCE DEFENSES ON DIFFERENT DOMAINS

As MIA work has evolved in various fields, MIDs for other tasks or scenarios have been introduced. At present, MID work for other tasks or scenarios can be divided into federated learning scenario, image generation task, image translation task, text generation task, and wireless signal classification and Beacon service. The categorization of the MIDs in other tasks/scenarios is listed in Table 5.

6.1 MIDs in Federated Learning Scenario

Federated Learning (FL), while providing a robust privacy solution by preventing private data from leaving the data owner's local device, is still a serious threat from MIAs. To defend against MIAs in FL, some work relies mainly on information perturbation technology.

For example, Liu and Naseri et al. [44, 49] proposed to use local and central differential privacy to reduce the privacy risk of members. However, this defense scheme increases noise in each update,

Table 5. The Categorization of the MIDs in Other Tasks/Scenarios

Other Tasks	Phase	Optimization Object	Defense Technology	Defense Method	Year	Ref.
Federated Learning Scenario	Training	Gradient of the Loss Function	Information Perturbation	DP	2020	[44, 49]
Image Generation Model	Pre-training	Preprocessing feature and label	Information Perturbation	Data Perturbation	2021	[8]
	Training	Loss Function	Regularization	Adversarial Training	2019	[48]
				Lipschitz-Regularization	2019	[87]
Image Translation Model	Pre-training	Preprocessing label	Transfer Learning	Knowledge Distillation	2022	[2]
Text Generation Model	Pre-training	Preprocessing feature and label	Information Perturbation	Data Perturbation	2022	[35]
Wireless Signal Classification and Beacon Service	Inference	Confidence Vector	Information Perturbation	Output Perturbation - Adversarial Noise	2022	[63]
		Model's Response	Information Perturbation	Output Perturbation - Flip Values	2021	[77]

resulting in a significant sacrifice in the classification accuracy of FL. To effectively alleviate this problem, Lee et al. [38] proposed an independent neural network, **Digestive Neural Network (DNN)**, which works in concert with a collaborative network for training. DNN extracts the features of a given small batch into completely different domains to delete the private information of the data. Small batches passing through DNN need to exclude the original information and must contain features useful for high classification accuracy. A collaborative network receives digested small batch data and optimizes its parameters to improve the classification accuracy of digested data. To improve the prediction accuracy of collaborative network, DNN and collaborative network need collaborative training. Simulation results show that DNN has good performance in the FL mechanism based on gradient sharing and weight sharing.

6.2 MIDs in Image Generation Task

At present, there are many MIA works for the classification model in the visual field, and there are a lot of defenses against MIAs with the help of the generation model. However, relevant studies show that generative adversarial networks have poor generalization ability and are easily affected by MIAs. Therefore, MIDs for GAN have also been launched in succession, mainly by regularization technology.

Chen et al. [8] attempted to improve the generalization of GAN from the perspective of privacy protection and designed a GAN framework, namely, **partition GAN (PAR-GAN)**, which consists of a generator and multiple discriminators trained on unconnected partitions of training data. The core idea of the PAR-GAN algorithm is to reduce the generalization gap by approximating the mixed distribution of all partitions of the training data. Theoretical analysis shows that PAR-GAN can achieve global optimization like the original GAN. The experimental results on simulated data and several commonly datasets show that PAR-GAN can improve the generalization ability of GAN and reduce the information leakage caused by MIAs. In addition, Mukherjee et al. [48] proposed a novel GAN framework (privGAN), which utilizes multiple generator discriminator pairs and a built-in opponent to prevent the model from overfitting the training set. Through the theoretical analysis of the optimal generator/discriminator, the consistency between privGAN and non-privGAN is proved. In a more practical scenario, it is verified that privGAN loss function is equivalent to adding regularization to prevent overfitting the training set.

In Reference [87], Wu et al. verified the generalization ability of GAN theoretically and experimentally and showed that the common goal of "narrowing the generalization gap" and "protecting member privacy" is to encourage the neural network to learn the characteristics of the group,

rather than remembering the characteristics of everyone, that is, the smaller the generalization gap, the less member information exposed in the training dataset. On the theoretical side, the authors leveraged the stability-based theory [61] to bridge the gap between differential privacy and the generalization and provide a new perspective from privacy protection to understand a number of recent techniques for improving the performance of GANs. On the experimental side, the authors quantitatively validated the relationship between the generalization gap and the information leakage of the training dataset. Results suggest that it is possible to design new variants of GAN from the perspective of building privacy-preserving learning algorithms, which can bring significant regularization effects while protecting the sensitive information of the training dataset.

6.3 MIDs in Image Translation Task

With the development of artificial intelligence, the image-to-image translation (image translation) model has attracted more and more attention, and its privacy problems have been noticed by relevant researchers. In Reference [60], Shafran et al. showed that the image translation model is vulnerable to the impact of MIAs, and it showed that the information perturbation technology can not protect the privacy of members in the image translation task. Alvar et al. [2] showed that they can resist MIAs on image translation models with the help of transfer learning technologies.

In Reference [2], Alvar et al. used the knowledge distillation method to alleviate MIAs in the image translation task. However, it is not easy to directly apply the knowledge distillation method to the image translation task. Because the sample size of image translation datasets is usually smaller than classification datasets, and the output of the image translation task has no entropy information, it is not feasible to apply the knowledge distillation method to the image translation task directly. To achieve a better tradeoff between utility and privacy, an **adversarial knowledge distillation (AKD)** method is proposed as a defense against MIAs in the image translation model by Alvar. This method combines knowledge distillation with adversarial training and protects the privacy of training samples by improving the generalization of the image translation model.

6.4 MIDs in Text Generation Task

Neural language models (NLMs)—systems trained to predict the next word in a sequence of text—have become the fundamental building blocks for numerous natural language processing tasks and domains. Unfortunately, Carlini et al. [5] pointed out that NLMs are vulnerable to MIAs. By generating sequences from the model and then scoring these sequences with different MIAs, the sequence with the highest score is classified as training data.

To solve this problem, Kandpal et al. [35] used the help of information perturbation technology and showed that the success of Reference [5]'s attack is mainly due to the repeated sequences found in the common network capture training dataset. To explain this reason, the authors showed the superlinear correlation between the speed of the language model regenerating the training sequence and the count of the sequence in the training set. It also showed that the existing methods for detecting the memory sequence have approximate accuracy on the non-repeated training sequence. Finally, it is found that the language model has high security against these types of privacy attacks after the method is applied to retrain the data.

6.5 MIDs in Wireless Signal Classification and Beacon Service

In addition to the MIA on classification models such as images and texts, it is also pointed out that the models for classifying wireless signals and genomic datasets are also vulnerable to MIAs. Considering the different data conditions, the effectiveness of defense methods is also different. Therefore, the relevant literatures have specifically studied MIDs for these two kinds of data with the help of information perturbation technology.

For wireless signal classification service, Yi et al. [63] first proposed an MIA for wireless signal classifier to expose member privacy of wireless data. On this basis, the author regards the defense problem as an optimization problem and uses shadow model for active defense. The main idea of defense is to add disturbance to the classification process so (1) the classification result is unchanged and (2) the MIA accuracy is low. The gradient search method is used to find the optimal disturbance in the process of classification. The simulation results show that the defense scheme can effectively defend against MIA.

Beacon service enables a user to query for the presence of particular minor alleles in an underlying private genomic dataset. While exposing such limited information may appear safe, it has been shown to be vulnerable to MIAs, because it allows users to issue queries for every region of the genome [65, 78]. A common approach to mitigate privacy risks in Beacon services is to flip the values in a subset of the query responses [78]. For example, the response to a particular allele does not exist, but it does exist in the dataset. However, not all methods offer privacy guarantees, and when they do, they are usually probabilistic. While minimizing the number of flipped queries is a standard measure of utility, none of the previous approaches provide a formal guarantee of optimality. In Reference [77], Rajagopal et al. used likelihood-ratio-test statistics to propose a novel framework, which can accurately dissect various ways of Beacon services. By controlling query methods and feedback strategies, the framework is superior to existing technologies in terms of privacy and practicality.

7 WHY MEMBERSHIP INFERENCE DEFENSES WORK

Membership inference attack exploits the different performance of member and non-member samples on the target model to infer whether a sample belongs to the training set. Therefore, the method to defend against membership inference attack must be to make it difficult for attackers to distinguish members from non-members according to the feedback of the target model, which is also the basic purpose of defense work. Existing defense works have different deep-seated reasons for the indistinguishable nature of members and non-members. We can summarize it into the following three defense principles and relate them in Table 6 with the attack principles in Section 3.3, the defense technologies in Section 4, and the defense phases with the defense optimization objectives in Section 5.

(1) **Avoid Overfitting.** As mentioned in Section 3.3, currently, the success of most MIAs is attributed to the overfitting of the target model. When the target model is in the overfitting state, its performance in the training set and the test set is different. Through this difference, the attacker can access the target model and determine whether the query data is a member or not. Therefore, many defense works achieve the defense of MIAs by reducing the degree of overfitting of the model. For example, some work through regularization tricks, some work through fine-grained controlling the training mode of the model, and so on.

(2) **Information Confusion.** As it has been said, existing MIA works are implemented based on the different performances between members and non-members. Therefore, to achieve the effect of resisting MIAs, the performance between members and non-members can be interfered. Defense based on this principle can be achieved by adding a small amount of noise to the three stages of the target model. However, while realizing defense MIAs, this defense method may significantly damage the utility of the target model, and the level of noise is difficult to control.

(3) **Information Isolation.** In the defense of MIAs, our basic goal is to protect the member privacy of the target model training data. Therefore, from another point of view, we can

Table 6. The Association between Defense Principle, Attack Principle, Defense Technology, Defense Phase, and Defense Optimization Goal

Principle of MIDs	Against MIAs Principle	Defense Technology	Defense Phase	Optimization Object
Avoid Overfitting.	• Overfitting of the target model	Regularization	Pre-training	Preprocessing feature
				Preprocessing label
				Preprocessing feature and label
			Training	Loss Function
				Model Parameter
Information Confusion	• Overfitting of the target model • The unique impact of the training set • Other properties of the target model	Information Perturbation	Pre-training	Preprocessing feature
				Preprocessing feature and label
			Training	Gradient of the Loss Function
			Inference	Confidence Vector
Information Isolation	• Overfitting of the target model • The unique impact of the training set • Other properties of the target model	Transfer Learning	Pre-training	Preprocessing feature
				Preprocessing label
				Preprocessing feature and label
		Generative Models-based	Pre-training	Preprocessing feature
				Preprocessing feature and label

substitute privacy-secure data for the original training data of the target model. Specifically, we can train a proxy model of the target model using privacy-secure data and publish the proxy model to provide the service. This method avoids the disclosure of member privacy by isolating the unprotected target model training data. In this process, the training data of the target model are all non-member data for the proxy model, so the membership attributes of the training data of the target model cannot be inferred according to the output of the proxy model to achieve the purpose of resisting membership inference attacks. However, this method also requires fine optimization to ensure the utility of the surrogate model.

8 FUTURE RESEARCH DIRECTIONS ON MEMBERSHIP INFERENCE DEFENSE

With the development and reform of deep learning technology and computing hardware architecture, artificial intelligence technology has made major breakthroughs in key tasks such as machine vision, speech recognition, and so on. However, membership inference attacks have also exposed privacy vulnerabilities in various domain models. In this section, we discuss several major challenges and potential research directions of membership inference defense to stimulate interested readers to explore this field more.

(1) **MIDs for Other Tasks/Scenarios.** At present, the membership inference defense of the classification model in the field of computer vision is relatively comprehensive. However, the defense work in other fields or tasks is not explored much, and the work in this gap needs to be further explored. Moreover, although there are many membership inference defense methods for classification models in the field of computer vision, whether they can be compatible with other tasks also needs systematic research.

(2) **Design MIDs from the Perspective of Member Features.** To date, most of the reasons for the success of membership inference attacks are attributed to the overfitting of the model, and the related attack work or defense work is to attack or defend by means of different performances of overfitting. However, what is the feature of model overfitting remains to

be explored, i.e., the nature of member features is still unclear. Although the literature [21] showed that data features can be decoupled into class features and member features, this work had only been verified on simple data, and the utility of data after removing member features is significantly reduced. It is not feasible to apply the work to defense directly, and the defense work from this point of view still needs to be improved.

(3) **Analyze the Privacy and Utility Implemented by MIDs.** Although there are many existing works on defending against membership inference attacks, there are not many works that can actually protect the member privacy of target model without compromising the utility of the model. Moreover, the work that satisfies this requirement does not inherently analyze the reasons for its implementation. There is still a lot of research space in the design and analysis of the work that can achieve privacy and ensure that the utility of the model is not damaged.

9 CONCLUSION

The large-scale and industrialized development of deep learning technology has formed a business pattern. In this pattern, data is widely used, yet at the same time, it also makes the privacy data of data holders face the risk of leakage. Although many researchers have carried out a series of studies on membership inference attacks and defenses and have carried out systematic classification and description, they have not carried out an independent systematic analysis on membership inference defenses.

This article first combs the work of membership inference attacks from the perspective of the amount of knowledge gained by the attacker and introduces the basic principles of membership inference attacks. We then give a brief overview of the existing defense technology. After that, taking the defense phase as the dividing point, it clearly shows the research progress and research ideas in the membership inference defense of the classification model in the field of computer vision. Subsequently, the defense work of missions in other fields is analyzed. Finally, the working principle and the challenge of membership inference defenses is discussed, and the potential direction of future research is pointed out. Through this comprehensive investigation, we hope to provide a solid foundation for the research on membership inference defenses.

REFERENCES

[1] 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 ACM SIGSAC Conference on Computer and Communications Security*. 308–318.

[2] Saeed Ranjbar Alvar, Lanjun Wang, Jian Pei, and Yong Zhang. 2022. Membership privacy protection for image translation models via adversarial knowledge distillation. *arXiv preprint arXiv:2203.05212* (2022).

[3] Yang Bai, Ting Chen, and Mingyu Fan. 2021. A survey on membership inference attacks against machine learning. *Management* 6 (2021), 14.

[4] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2021. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570* (2021).

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Oprea Alina, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security'21)*. 2633–2650.

[6] Dingfan Chen, Ning Yu, and Mario Fritz. 2021. RelaxLoss: Defending membership inference attacks without losing utility. In *Proceedings of the International Conference on Learning Representations*.

[7] Jiyu Chen, Yiwen Guo, Qianjun Zheng, and Hao Chen. 2021. Protect privacy of deep classification networks by exploiting their generative power. *Mach. Learn.* 110, 4 (2021), 651–674.

[8] Junjie Chen, Wendy Hui Wang, Hongchang Gao, and Xinghua Shi. 2021. PAR-GAN: Improving the generalization of generative adversarial networks against membership inference attacks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 127–137.

[9] Junjie Chen, Wendy Hui Wang, and Xinghua Shi. 2020. Differential privacy protection against membership inference attack on machine learning for genomic data. In *Proceedings of the Pacific Symposium on Biocomputing*. World Scientific, 26–37.
[10] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaarfar, and Haojin Zhu. 2018. Differentially private data generative models. *arXiv preprint arXiv:1812.02274* (2018).
[11] Zongqi Chen, Hongwei Li, Meng Hao, and Guowen Xu. 2021. Enhanced mixup training: A defense method against membership inference attack. In *Proceedings of the International Conference on Information Security Practice and Experience*. Springer, 32–45.
[12] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1964–1974.
[13] Rishav Chourasia, Batnyam Enkhtaivan, Kunihiro Ito, Junki Mori, Isamu Teranishi, and Hikaru Tsuchida. 2021. Knowledge cross-distillation for membership privacy. *arXiv preprint arXiv:2111.01363* (2021).
[14] Gilad Cohen and Raja Giryes. 2022. Membership inference attack using self influence functions. *arXiv preprint arXiv:2205.13680* (2022).
[15] Elliot J. Crowley, Gavin Gray, and Amos J. Storkey. 2018. Moonshine: Distilling with cheap convolutions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2893–2903.
[16] Gonzalo Martínez Ruiz de Arcaute, José Alberto Hernández, and Pedro Reviriego. 2022. Assessing the impact of membership inference attacks on classical machine learning algorithms. In *Proceedings of the 18th International Conference on the Design of Reliable Communication Networks (DRCN'22)*. IEEE, 1–4.
[17] Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. 2022. Leveraging adversarial examples to quantify membership information leakage. *arXiv preprint arXiv:2203.09566* (2022).
[18] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Proceedings of the International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.
[19] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC'06, New York, NY, USA, March 4-7, 2006. Proceedings 3*, Springer, 265–284.
[20] Y. X. Fu, Y. B. Qin, and G. W. Shen. 2019. Sensitive data privacy protection method based on transfer learning. *J. Data Acquisition Process* 34, 3 (2019), 422–431.
[21] Heonseok Ha, Jaehee Jang, Yonghyun Jeong, and Sungroh Yoon. 2022. Membership feature disentanglement network. In *Proceedings of the ACM on Asia Conference on Computer and Communications Security*. 364–376.
[22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
[23] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).
[24] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Yi Chen, and Xuyun Zhang. 2021. EAR: An enhanced adversarial regularization approach against membership inference attacks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'21)*. IEEE, 1–8.
[25] 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* 54, 11s (2022), 1–37.
[26] Li Hu, Jin Li, Guanbiao Lin, Shiyu Peng, Zhenxin Zhang, Yingying Zhang, and Changyu Dong. 2022. Defending against membership inference attacks with high utility by GAN. *IEEE Transactions on Dependable and Secure Computing* 20, 3 (2022), 2144–2157.
[27] Hongwei Huang. 2021. Defense against membership inference attack applying domain adaptation with addictive noise. *J. Comput. Commun.* 9, 5 (2021), 92–108.
[28] Hongwei Huang, Weiqi Luo, Guoqiang Zeng, Jian Weng, Yue Zhang, and Anjia Yang. 2021. Damia: leveraging domain adaptation as a defense against membership inference attacks. *IEEE Transactions on Dependable and Secure Computing* 19, 5 (2021), 3183–3199.
[29] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. 2021. Practical blind membership inference attack via differential comparisons. *arXiv preprint arXiv:2101.01341* (2021).
[30] Thomas Humphries, Matthew Rafuse, Lindsey Tulloch, Simon Oya, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2020. Differentially private learning does not bound membership inference. *arXiv preprint arXiv:2010.12112* (2020).
[31] Matthew Jagielski, Milad Nasr, Christopher Choquette-Choo, Katherine Lee, and Nicholas Carlini. 2023. Students parrot their teachers: Membership inference on model distillation. *arXiv preprint arXiv:2303.03446* (2023).
[32] Ismat Jarin and Birhanu Eshete. 2022. MIAShield: Defending membership inference attacks via preemptive exclusion of members. *arXiv preprint arXiv:2203.00915* (2022).

[33] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security'19)*. 1895–1912.
[34] 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 ACM SIGSAC Conference on Computer and Communications Security*. 259–274.
[35] Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. *arXiv preprint arXiv:2202.06539* (2022).
[36] Yigitcan Kaya and Tudor Dumitras. 2021. When does data augmentation help with membership inference attacks? In *Proceedings of the International Conference on Machine Learning*. PMLR, 5345–5355.
[37] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2020. On the effectiveness of regularization against membership inference attacks. *arXiv preprint arXiv:2006.05336* (2020).
[38] 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. *Comput. Secur.* 109 (2021), 102378.
[39] Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security'20)*. 1605–1622.
[40] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership inference attacks and defenses in classification models. In *Proceedings of the 11th ACM Conference on Data and Application Security and Privacy*. 5–16.
[41] Zheng Li and Yang Zhang. 2020. Label-leaks: Membership inference attack with label. *arXiv e-prints* (2020), arXiv–2007.
[42] Gaoyang Liu, Yutong Li, Borui Wan, Chen Wang, and Kai Peng. 2021. Membership inference attacks in black-box machine learning models. *J. Cyber Secur.* 6, 3 (2021), 15.
[43] Yaru Liu, Hongcheng Li, Gang Huang, and Wei Hua. 2022. OPUPO: Defending against membership inference Attacks With Order-Preserving and Utility-preserving obfuscation. *IEEE Transactions on Dependable and Secure Computing* (2022), 1–12.
[44] Yi Liu, Jialiang Peng, Jiawen Kang, Abdullah M. Iliyasu, Dusit Niyato, and Ahmed A. Abd El-Latif. 2020. A secure federated learning framework for 5G networks. *IEEE Wirel. Commun.* 27, 4 (2020), 24–31.
[45] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. 2022. Membership inference attacks by exploiting loss trajectory. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 2085–2098.
[46] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. 2018. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889* (2018).
[47] Federico Mazzone, Leander van den Heuvel, Maximilian Huber, Cristian Verdecchia, Maarten Everts, Florian Hahn, and Andreas Peter. 2022. Repeated knowledge distillation with confidence masking to mitigate membership inference attacks. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*. 13–24.
[48] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, and Juan Lavista Ferres. 2019. privGAN: Protecting GANs from membership inference attacks at low cost. *arXiv preprint arXiv:2001.00071* (2019).
[49] Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. 2020. Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy. *arXiv e-prints* (2020), arXiv–2009.
[50] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 634–646.
[51] 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 *Proceedings of the IEEE Symposium on Security and Privacy (SP'19)*. IEEE, 739–753.
[52] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. *arXiv preprint arXiv:2101.04535* (2021).
[53] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2016. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755* (2016).
[54] William Paul, Yinzhi Cao, Miaomiao Zhang, and Phil Burlina. 2021. Defending medical image diagnostics against privacy attacks using generative methods: Application to retinal diagnostics. In *Clinical Image-based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-preserving Machine Learning*. Springer, 174–187.
[55] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. 2021. Differential privacy defenses and sampling attacks for membership inference. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*. 193–202.
[56] 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.

[57] Maria Rigaki and Sebastian Garcia. 2020. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646* (2020).
[58] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the International Conference on Machine Learning*. PMLR, 5558–5567.
[59] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).
[60] 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.
[61] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2010. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* 11 (2010), 2635–2670.
[62] Virat Shejwalkar and Amir Houmansadr. 2019. Membership privacy for machine learning models through knowledge transfer. *arXiv preprint arXiv:1906.06589* (2019).
[63] Yi Shi and Yalin Sagduyu. 2022. Membership inference attack and defense for wireless signal classifiers with deep learning. *IEEE Transactions on Mobile Computing* 22, 7 (2022), 4032–4043.
[64] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 3–18.
[65] Suyash S. Shringarpure and Carlos D. Bustamante. 2015. Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* 97, 5 (2015), 631–646.
[66] Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security'21)*.
[67] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.
[68] Jasper Tan, Daniel LeJeune, Blake Mason, Hamid Javadi, and Richard G. Baraniuk. 2022. Benign overparameterization in membership inference with early stopping. *arXiv preprint arXiv:2205.14055* (2022).
[69] Xinyu Tang, Saeed Mahloujifar, Liwei Song, Virat Shejwalkar, Milad Nasr, Amir Houmansadr, and Prateek Mittal. 2021. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. *arXiv preprint arXiv:2110.08324* (2021).
[70] Gao Ting. 2022. Research progress and challenges of membership inference attacks in machine learning. *Oper. Res. Fuzziol.* 12, 1 (2022), 15.
[71] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 648–656.
[72] Shakila Mahjabin Tonni, Dinusha Vatsalan, Farhad Farokhi, Dali Kaafar, Zhigang Lu, and Gioacchino Tangari. 2020. Data and model dependencies of membership inference attack. *arXiv preprint arXiv:2002.06856* (2020).
[73] Aleksei Triastcyn and Boi Faltings. 2018. Generating artificial data for private deep learning. *arXiv preprint arXiv:1803.03148* (2018).
[74] 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 3rd ACM International Workshop on Edge Systems, Analytics and Networking*. 61–66.
[75] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Wenqi Wei, and Lei Yu. 2019. Effects of differential privacy and data skewness on membership inference vulnerability. In *Proceedings of the 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA'19)*. IEEE, 82–91.
[76] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing* 14, 6 (2019), 2073–2089.
[77] Rajagopal Venkatesaramani, Zhiyu Wan, Bradley A. Malin, and Yevgeniy Vorobeychik. 2021. Defending against membership inference attacks on Beacon services. *arXiv preprint arXiv:2112.13301* (2021).
[78] Zhiyu Wan, Yevgeniy Vorobeychik, Murat Kantarcioglu, and Bradley Malin. 2017. Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services. *BMC Med. Genom.* 10, 2 (2017), 87–100.
[79] Chen Wang, Gaoyang Liu, Haojun Huang, Weijie Feng, Kai Peng, and Lizhe Wang. 2019. MIASec: Enabling data indistinguishability against membership inference attacks in MLaaS. *IEEE Trans. Sustain. Comput.* 5, 3 (2019), 365–376.
[80] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S. Yu Philip. 2019. Private model compression via knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1190–1197.

[81] Lulu Wang, Peng Zhang, Zheng Yan, and Xiaokang Zhou. 2019. A survey on membership inference on training datasets in machine learning. *Cybersp. Secur.* 10, 10 (2019), 7.
[82] Yijue Wang, Chenghong Wang, Zigeng Wang, Shanglin Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. 2020. Against membership inference attack: Pruning is all you need. arXiv preprint arXiv:2008.13578 (2020).
[83] Ryan Webster, Julien Rabin, Loïc Simon, and Frédéric Jurie. 2021. Generating private data surrogates for vision related tasks. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21)*. IEEE, 263–269.
[84] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang. 2016. A survey of transfer learning. *J. Big Data* 3, 1 (2016), 1–40.
[85] Yuxin Wen, Arpit Bansal, Hamid Kazemi, Eitan Borgnia, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. Canary in a coalmine: Better membership inference with ensembled adversarial queries. *arXiv preprint arXiv:2210.10750* (2022).
[86] Bingzhe Wu, Chaochao Chen, Shiwan Zhao, Cen Chen, Yuan Yao, Guangyu Sun, Li Wang, Xiaolu Zhang, and Jun Zhou. 2020. Characterizing membership privacy in stochastic gradient Langevin dynamics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6372–6379.
[87] Bingzhe Wu, Shiwan Zhao, Chaochao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. 2019. Generalization in generative adversarial networks: A novel perspective from privacy protection. *Adv. Neural Inf. Process. Syst.* 32 (2019).
[88] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* (2018).
[89] Nuo Xu, Binghui Wang, Ran Ran, Wujie Wen, and Parv Venkitasubramaniam. 2022. NeuGuard: Lightweight neuron-guided defense against membership inference attacks. *arXiv preprint arXiv:2206.05565* (2022).
[90] Mingfu Xue, Chengxiang Yuan, Can He, Zhiyu Wu, Yushu Zhang, Zhe Liu, and Weiqiang Liu. 2020. Use the spear as a shield: A novel adversarial example based privacy-preserving technique against membership inference attacks. *arXiv preprint arXiv:2011.13696* (2020).
[91] Ruikang Yang, Jianfeng Ma, Yinbin Miao, and Xindi Ma. 2022. Privacy-preserving generative framework against membership inference attacks. *arXiv preprint arXiv:2202.05469* (2022).
[92] Ziqi Yang, Bin Shao, Bohan Xuan, Ee-Chien Chang, and Fan Zhang. 2020. Defending model inversion and membership inference attacks via prediction purification. *arXiv preprint arXiv:2005.03915* (2020).
[93] Ziqi Yang, Lijin Wang, Da Yang, Jie Wan, Ziming Zhao, Ee-Chien Chang, Fan Zhang, and Kui Ren. 2022. Purifier: Defending data inference attacks via transforming confidence scores. *arXiv preprint arXiv:2212.00612* (2022).
[94] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. 2021. Enhanced membership inference attacks against machine learning models. *arXiv preprint arXiv:2111.09679* (2021).
[95] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of the IEEE 31st Computer Security Foundations Symposium (CSF'18)*. IEEE, 268–282.
[96] Yu Yin, Ke Chen, Lidan Shou, and Gang Chen. 2021. Defending privacy against more knowledgeable membership inference attackers. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2026–2036.
[97] Zuobin Ying, Yun Zhang, and Ximeng Liu. 2020. Privacy-preserving in defending against membership inference attacks. In *Proceedings of the Workshop on Privacy-preserving Machine Learning in Practice*. 61–63.
[98] Yu FU and others. 2019. Multi-source data privacy protection based on transfer learning. *Computer Engineering & Science* 41, 4 (2019), 641.
[99] Bo Zhang, Ruotong Yu, Haipei Sun, Yanying Li, Jun Xu, and Hui Wang. 2020. Privacy for all: Demystify vulnerability disparity of differential privacy against membership inference attack. *arXiv preprint arXiv:2001.08855* (2020).
[100] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership inference attacks against recommender systems. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 864–879.
[101] Tianwei Zhang, Zecheng He, and Ruby B. Lee. 2018. Privacy-preserving machine learning through data obfuscation. *arXiv preprint arXiv:1807.01860* (2018).
[102] Tianyun Zhang, Shaokai Ye, Kaiqi Zhang, Jian Tang, Wujie Wen, Makan Fardad, and Yanzhi Wang. 2018. A systematic DNN weight pruning framework using alternating direction method of multipliers. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 184–199.
[103] Xinyang Zhang, Shouling Ji, and Ting Wang. 2018. Differentially private releasing via deep generative model (technical report). arXiv preprint arXiv:1801.01594 (2018).
[104] Zhaoxi Zhang, Leo Yu Zhang, Xufei Zheng, Bilal Hussain Abbasi, and Shengshan Hu. 2022. Evaluating membership inference through adversarial robustness. *arXiv preprint arXiv:2205.06986* (2022).

[105] Jingwen Zhao, Yunfang Chen, and Wei Zhang. 2019. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access* 7 (2019), 48901–48911.
[106] Junxiang Zheng, Yongzhi Cao, and Hanpin Wang. 2021. Resisting membership inference attacks through knowledge distillation. *Neurocomputing* 452 (2021), 114–126.
[107] Da Zhong, Haipei Sun, Jun Xu, Neil Gong, and Wendy Hui Wang. 2022. Understanding disparate effects of membership inference attacks and their countermeasures. In *Proceedings of the ACM on Asia Conference on Computer and Communications Security*. 959–974.

Received 22 July 2022; revised 27 March 2023; accepted 22 August 2023