Membership Inference Attacks on Machine Learning: A Survey

HONGSHENG HU and ZORAN SALCIC, The University of Auckland, New Zealand
LICHAO SUN, Lehigh University, USA
GILLIAN DOBBIE, The University of Auckland, New Zealand
PHILIP S. YU, University of Illinois at Chicago, USA
XUYUN ZHANG, Macquarie University, Australia

Machine learning (ML) models have been widely applied to various applications, including image classification, text generation, audio recognition, and graph data analysis. However, recent studies have shown that ML models are vulnerable to membership inference attacks (MIAs), which aim to infer whether a data record was used to train a target model or not. MIAs on ML models can directly lead to a privacy breach. For example, via identifying the fact that a clinical record that has been used to train a model associated with a certain disease, an attacker can infer that the owner of the clinical record has the disease with a high chance. In recent years, MIAs have been shown to be effective on various ML models, e.g., classification models and generative models. Meanwhile, many defense methods have been proposed to mitigate MIAs. Although MIAs on ML models form a newly emerging and rapidly growing research area, there has been no systematic survey on this topic yet. In this article, we conduct the first comprehensive survey on membership inference attacks and defenses. We provide the taxonomies for both attacks and defenses, based on their characterizations, and discuss their pros and cons. Based on the limitations and gaps identified in this survey, we point out several promising future research directions to inspire the researchers who wish to follow this area. This survey not only serves as a reference for the research community but also provides a clear description for researchers outside this research domain. To further help the researchers, we have created an online resource repository, which we will keep updated with future relevant work. Interested readers can find the repository at https://github.com/HongshengHu/membership-inference-machine-learning-literature.

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

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

ACM Reference format:
Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* 54, 11s, Article 235 (September 2022), 37 pages.
https://doi.org/10.1145/3523273

Dr Xuyun Zhang is the recipient of an ARC DECRA (project No. DE210101458) funded by the Australian Government. This work is supported in part by NSF under grants III-1763325, III-1909323, III-2106758, and SaTC-1930941.

Authors' addresses: H. Hu, Z. Salcic, and G. Dobbie, The University of Auckland, New Zealand; emails: hhu603@aucklanduni.ac.nz, {z.salcic, g.dobbie}@auckland.ac.nz; L. Sun, Lehigh University; email: lis221@lehigh.edu; P. S. Yu, University of Illinois at Chicago; email: psyu@uic.edu; X. Zhang, Macquarie University, Australia; email: xuyun.zhang@mq.edu.au.

0360-0300/2022/09-ART235 $15.00
https://doi.org/10.1145/3523273

1 INTRODUCTION

Machine learning (ML) has achieved tremendous results for various learning tasks, including image recognition [62], natural language processing [38], graph data applications [96], as well as advanced applications such as brain circuits analysis [113], healthcare analysis [136], and functionality of mutations in DNA [221]. Besides the powerful computational resources, the availability of large datasets is another key factor contributing to the success of ML [9]. As datasets can contain individuals' private information such as user speech, images, and medical records, it is essential that ML models should not leak privacy sensitive information about their training data. However, recent studies [16, 183, 234] have shown that ML models are prone to memorizing information of training data, making them vulnerable to several privacy attacks such as model extraction attacks [197], attribute inference attacks (also known as model inversion attacks) [47], property inference attacks [49], and membership inference attacks [181]. Model extraction attacks aim to duplicate the functionality of an ML model, i.e., an attacker tries to construct another model whose predictive performance is similar to the target ML model. Unlike model extraction attacks targeting the ML model, the attacker of an attribute inference attack, property inference attack, or membership inference attack focuses on inferring private information of the training data. More specifically, attribute inference attacks aim to infer sensitive attributes of a target data record given the output of a model and the information about the non-sensitive attributes. Property inference attacks aim to infer the global property of the training dataset. For example, given a malware classifier model whose training data consists of the execution traces of malicious and benign software, property inference attacks infer the property of the testing environment, which can be viewed as a property of the entire training dataset. **Membership inference attacks (MIAs)**, which are also the focus of this article, aim to infer members of the training dataset of an ML model. We discuss MIAs in detail subsequently.

MIAs on ML models aim to infer whether a data record was used to train a target ML model or not. MIAs can raise severe privacy risks to individuals. For example, by identifying the fact that a clinical record has been used to train a model associated with a certain disease, MIAs can infer that the owner of the clinical record has the disease with a high chance. A recent report [193] published by the **National Institute of Standards and Technology (NIST)** specifically mentions that an MIA determining an individual was included in the dataset used to train the target model is a confidentiality violation. Moreover, such privacy risks caused by MIAs can lead to commercial companies who wish to release **machine learning as a service (MLaaS)** to violate privacy regulations. For example, Veale et al. [202] mention that MIAs on ML models increase their risks of being classified as private personal information under the **General Data Protection Regulation (GDPR)** [213]. The concept of MIAs is first proposed by Homer et al. in Reference [70], where they demonstrate an attacker can leverage the published statistics about a genomics dataset to infer the presence of a particular genome in this dataset. Also, recent papers [157, 158] have shown the feasibility of MIAs on location data. Besides the MIAs on such databases, Shokri et al. [181] proposed the first MIAs on several classification models in the context of ML. They demonstrate that an attacker can identify whether a data record was used to train a neural network-based classifier or not, solely based on the prediction vector of the data record (which is also known as black-box access to the target ML model). Since then, there have been an increasing number of studies that investigate MIAs on various ML models, including regression models [57], classification models [181], generation models [60], and embedding models [182]. Meanwhile, a large body of work proposes different membership inference defenses from different perspectives to defend against MIAs while preserving the utility of the target ML models.

Given the importance of data privacy protection and the successful applications of ML models in various domains, both academia and industry are interested in the privacy of ML models. In this article, we contribute the first study summarizing different membership inference attacks and defenses on ML models and establish taxonomies based on various criteria for the relevant research communities. There are many surveys that summarize different attacks on ML models [35, 84, 114, 117, 137, 155, 166, 168, 176, 190, 228]. Among them, a line of surveys [168, 176, 190] focuses on adversarial attacks [192], which can lead to severe security risks in critical ML application domains such as self-driving cars, health care, and cybersecurity. For example, in Reference [168], the authors comprehensively summarize and shed a light on the risks of the latest studies on adversarial attacks in the domain of cybersecurity. Another line of surveys [35, 114, 137, 166, 228] focuses on privacy attacks, which can breach personal information that violates the privacy of an individual. These survey papers either investigate the privacy issues in a specific paradigm like federated learning [84] or provide general discussions [35] on different privacy attacks such as attribute inference attacks, property inference attacks, and membership inference attacks. The existing surveys of privacy attacks [35, 84, 114, 137, 166, 228] have mentioned MIAs with basic introductions to the concepts and shallow discussions of the methods. In contrast, our survey on MIAs differs from them significantly in scope and depth. Instead of covering all the privacy attacks, we focus only on MIAs, given that they have emerged recently and are of great interest to the research community due to their high likelihood of compromising the privacy of training data. Unlike the existing reviews that select a very limited number of publications related to MIAs, e.g., only eight references are included in Reference [114], we conduct a comprehensive search and include more than 100 related works in this survey. Our survey offers deeper discussion on the concepts, theories, methods, categorization with taxonomies, and visions of future research directions. The main contributions of this article are:

1. **Comprehensive Review.** To the best of our knowledge, this is the first work to provide a comprehensive review of membership inference attacks and defenses on ML models. We summarize most, if not all, the published and pre-print works (over 100 papers) before September 2021. In this work, we establish novel taxonomies for membership inference attacks and defenses, respectively, according to various criteria.
2. **Taxonomies of Membership Inference on ML Models.** There are already over a hundred papers published in this domain. A list of all papers can help but is not good enough for readers to quickly understand the similarity and differences among membership inference attacks and among membership inference defenses. To this end, we categorize all existing works of MIAs based on different target ML models, adversarial knowledge, attack methods, training algorithms, and task domains, respectively. For membership inference defenses, we categorize all existing works based on different techniques. More details of the taxonomies are given in Figures 8 and 9.
3. **Challenges and Future Directions.** Membership inference attacks on machine learning models is an active and ongoing area of research. Based on the literature reviewed, we have discussed the challenges yet to be solved and proposed several promising future directions for membership inference attacks and membership inference defenses, respectively, to inspire interested readers to explore this field in more depth.
4. **Datasets and Metrics.** To help researchers conduct empirical studies on membership inference attacks and defenses, we summarize most, if not all, the datasets and metrics that have been used in previous work. This aims to pave the way for the community to build a good benchmark in this area for future empirical analysis and in-depth technical understanding.

(5) **Online Updating Resource.** We create an open-source repository[1] that includes most, if not all, the relevant work. This repository provides all paper links and released code links to help researchers interested in this area. As a small number of surveyed papers are only available in pre-print, authors are welcome to update us when the full publication information becomes available. We will keep updating the repository with new work in this domain in the future. We hope this open-source repository can shed light on future research about membership inference analysis on ML models.

The rest of the article is organized as follows: Section 2 introduces ML preliminaries. In Section 3, we introduce the existing attack approaches and provide taxonomies to categorize the released papers. In Section 4, we discuss why MIAs can work on ML models. Section 5 provides taxonomies for membership inference defenses. Section 6 summarizes datasets, metrics, and open-source implementation of popular approaches. We discuss the challenges and propose the future directions in Section 7. We conclude this article in Section 8.

2 PRELIMINARIES ABOUT MACHINE LEARNING

To help the audience understand the context of machine learning where membership inference attacks are performed, we introduce the basic preliminaries of machine learning. It is worth noting that this is not a comprehensive introduction, and interested readers can refer to References [2, 51, 139] for a systematic introduction.

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by learning from data [139]. Generally, ML algorithms can be divided into two categories, i.e., supervised learning and unsupervised learning, depending on the information provided by the training data and the different learning tasks.

Supervised Learning. A supervised ML model aims to learn a general rule that maps inputs to outputs from a labeled dataset [170]. Let $D_{train} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$ be a training dataset, where N is the number of data instances, $\boldsymbol{x}$ is a feature vector, and y is a label. An ML model is a function $f(\boldsymbol{x}; \boldsymbol{\theta})$ that takes as input $\boldsymbol{x}$ and outputs $y = f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are parameters that are learned from D_{train}. When y is discrete, the learning task of $f(\boldsymbol{x}; \boldsymbol{\theta})$ is called classification. When y is continuous, the learning task of $f(\boldsymbol{x}; \boldsymbol{\theta})$ is called regression.

Training Supervised ML Models. A well-trained supervised ML model should have a small expectation loss for the data it works on. However, as we do not know the true distribution of data, we cannot calculate the model's expected risk. A realistic approach to train supervised ML models is **Empirical Risk Minimization (ERM)** [201]. The core idea is to measure the model's performance on a known training dataset. For a given dataset D_{train}, ERM tries to find the parameters $\boldsymbol{\theta}^{*}$ that minimize the following objective function:

$$\min \mathcal{R}_{D_{train}}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(y^{(n)}, f(\boldsymbol{x}^{(\boldsymbol{n})}; \theta)), \tag{1}$$

where $\mathcal{L}(\cdot, \cdot)$ is a loss function. An iterative optimization algorithm called ***stochastic gradient descent*** **(SGD)** [171] is usually used to find the best parameters θ^{*}. The SGD algorithm follows:

$$\theta_{t+1} = \theta_t - \alpha \frac{\partial \mathcal{R}_{\mathcal{D}}(\theta)}{\partial \theta}, \tag{2}$$

$$\frac{\partial \mathcal{R}_{\mathcal{D}}(\theta)}{\partial \theta} = \frac{1}{K} \sum_{n=1}^{K} \frac{\partial \mathcal{L}\left(y^{(n)}, f\left(\boldsymbol{x}^{(n)}; \theta\right)\right)}{\partial \theta}, \tag{3}$$

[1] https://github.com/HongshengHu/membership-inference-machine-learning-literature.

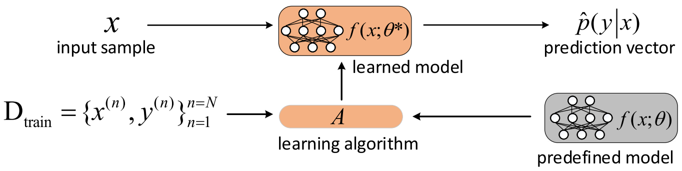


Fig. 1. A typical deep learning process for classification models.

where K is the batch size, θ_t are iterative parameters in the tth time, and α is the learning rate. Training is finished when the model converges to a local minimum.

Unsupervised Learning. An unsupervised ML model aims to extract features and patterns from unlabeled data or labeled data without access to the labels [68]. Recently, generative models, which aim to learn how to generate samples from the underlying data distribution, are gaining increasing attention as a typical unsupervised learning method. There are two typical generative models, **Generative Adversarial Networks (GANs)** [52] and **Variational Autoencoders (VAEs)** [95].

Training Unsupervised ML Models. We briefly introduce how to train GANs and VAEs because current MIAs on unsupervised learning models mainly target GANs and VAEs.

A GAN consists of two competing neural network modules, a generator $\mathcal{G}$, and a discriminator $\mathcal{D}$, which are trained to compete against each other. The generator takes the latent variable $\boldsymbol{z}$ and generates samples $\mathcal{G}_{\theta_{\mathcal{G}}}(\boldsymbol{z})$ that approximate the data distribution of D_{train}. The discriminator receives samples from D_{train} and the generated samples, and it is trained to learn the difference between them. The discriminator essentially is a binary classifier that determines whether $\boldsymbol{x}$ was taken from D_{train} or $\mathcal{G}$. After training, $\mathcal{G}$ can receive different $\boldsymbol{z}$ and generate synthetic samples. As both the generator and discriminator are neural networks, SGD is usually used for training GANs. SGD tries to find the parameters $\theta^*_{\mathcal{G}}$ of GANs following the objective function:

$$\min\nolimits_{\theta_{\mathcal{G}}} \max\nolimits_{\theta_{\mathcal{D}}} \quad \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}}\left[\log\left(\mathcal{D}_{\theta_{\mathcal{D}}}(\boldsymbol{x})\right)\right] + \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}}\left[\log\left(1-\mathcal{D}_{\theta_{\mathcal{D}}}\left(\mathcal{G}_{\theta_{\mathcal{G}}}(\boldsymbol{z})\right)\right)\right], \tag{4}$$

where $\theta_{\mathcal{G}}$ and $\theta_{\mathcal{D}}$ are parameters of the generator and the discriminator in GANs, P_{data} is the distribution of D_{train}, while $P_{\boldsymbol{z}}$ is the distribution of the latent variable $\boldsymbol{z}$.

VAEs aim to find an approximate posterior distribution function $q(\boldsymbol{z}|\boldsymbol{x})$ that parameterizes the latent distribution $P_{\boldsymbol{z}}$ according to the input data [95]. A VAE consists of an encoder and a decoder. The encoder maps data into a latent space, while the decoder maps the encoded latent representation back to the data space. At the beginning, the prior distribution of the latent variable $P_{\boldsymbol{z}}$ is defined as a unit normal distribution. Accordingly, the encoder and decoder are trained jointly such that the output of the decoder minimizes a reconstruction error between the parametric posterior and the true posterior measured by the **Kullback-Leibler (KL)** divergence [99]. Formally, to train VAEs, SGD tries to find the parameters θ^*_{de} following the objective function:

$$\min_{\theta_{\text{en}},\theta_{\text{de}}} -\mathbb{E}_{q_{\theta_{\text{en}}}(\boldsymbol{z}|\boldsymbol{x})}\left[p_{\theta_{\text{de}}}(\boldsymbol{x}\mid\boldsymbol{z})\right] + KL\left(q_{\theta_{\text{en}}}(\boldsymbol{z}\mid\boldsymbol{x})\|P_{\boldsymbol{z}}\right), \tag{5}$$

where $q_{\theta_{\text{en}}}(\boldsymbol{z}\mid\boldsymbol{x})$ and $p_{\theta_{\text{de}}}(\boldsymbol{x}\mid\boldsymbol{z})$ are the encoder and the decoder, and θ_{en} and θ_{de} are their parameters, $KL(\cdot\|\cdot)$ is the KL divergence, and $P_{\boldsymbol{z}}$ is the distribution of $\boldsymbol{z}$.

3 MEMBERSHIP INFERENCE ATTACKS ON MACHINE LEARNING MODELS

In this section, we first give a general definition of MIAs on ML models and then introduce adversarial knowledge, attack approaches, and target models of MIAs.

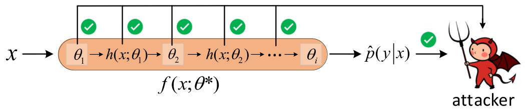

x
$\theta_1 \to h(x;\theta_1) \to \theta_2 \to h(x;\theta_2) \to \cdots \to \theta_i$
$\hat{p}(y|x)$
$f(x;\theta^*)$
attacker

Fig. 2. Overview of white-box membership inference attacks.

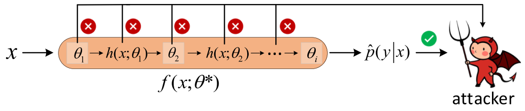

x
$\theta_1 \to h(x;\theta_1) \to \theta_2 \to h(x;\theta_2) \to \cdots \to \theta_i$
$\hat{p}(y|x)$
$f(x;\theta^*)$
attacker

Fig. 3. Overview of black-box membership inference attacks.

3.1 Definition of Membership Inference Attacks

To better illustrate the definition of MIAs, we introduce a typical learning process of an ML model. Figure 1 shows the typical learning process of a **deep neural network (DNN)** classifier. We use a learning algorithm $\mathcal{A}$ to train a predefined classifier $f(\boldsymbol{x};\theta)$ using dataset $D_{\text{train}} = \{(\boldsymbol{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$. Once the training process is finished, the learned model $f(\boldsymbol{x};\theta^*)$ can be used to make predictions for unseen data. The definition of MIAs on ML models is as follows: Given an exact input $\boldsymbol{x}$ and access to the learned model $f(\boldsymbol{x};\theta^*)$, an attacker infers whether $\boldsymbol{x} \in D_{\text{train}}$ or not.

3.2 Adversarial Knowledge

The attacker can receive different amounts of information to attack ML models. In this section, we first introduce adversarial knowledge and then introduce black-box and white-box MIAs.

There are two kinds of knowledge that are useful for an attacker to implement MIAs on ML models, i.e., knowledge of training data and knowledge of the target model. Knowledge of training data refers to the distribution of training data. In most settings of MIAs, the distribution of training data is assumed to be available to an attacker of MIAs. This means the attacker can obtain a shadow dataset containing data records from the same data distribution as the training records. This assumption is reasonable, because the shadow dataset can be obtained by statistics-based synthesis when the data distribution is known and model-based synthesis when the data distribution is unknown [181]. To conduct a non-trivial MIA, it is often assumed that the shadow dataset and the training dataset are disjoint. Knowledge of the target model refers to how the target model is trained (i.e., the learning algorithm) and the target model's architecture and learned parameters. Based on adversarial knowledge, we can characterize the dangerous levels of existing attacks.

White-box Attack. Under this setting, an attacker can get all information and use it to attack a target ML model. The information includes the distribution of training data, how the target model is trained, and the architecture and the learned parameters of the target model.

Black-box Attack. In this case, an attacker can only have black-box access to a target model. The attacker is given information limited to training data distribution and black-box queries on the target model. For example, the attacker queries the target classifier (if the target model is a classification model) and only gets prediction output of the input record.

We refer to Nasr et al.'s way of depicting white-box and black-box MIAs [147] and draw Figures 2 and 3 to illustrate the concepts of these two types of attacks and their differences for better visual understanding. Figures 2 and 3 show white-box and black-box MIAs on a target ML model, respectively (assuming the target model is a DNN classification model). In both figures, a green tick indicates availability, and the red cross indicates unavailability. In the white-box setting, an attacker has full access to the target classifier and obtains all information, including the learned

Table 1. Three Types of Black-Box Membership Inference Attacks Based on Different Information Provided by the Prediction Vector

Prediction Output	Description
Full confidence scores	The attacker queries an input record and obtains all confidence scores returned by the target classifier. Based on this, the attacker can obtain the predicted label of the target record. Thus, the attacker can further calculate the input's prediction loss (e.g., cross-entropy loss), because this attacker knows the predicted label.
Top-K confidence scores	The attacker queries an input record and obtains only top-K confidence scores returned by the target classifier. For example, the attacker only receives the probabilities of the most likely three classes (assuming the total number of classes is much larger than three).
Prediction label only	The attacker queries an input record and obtains only the predicted label returned by the target classifier. In this case, the attacker is given the most limited knowledge.

parameters of the classifier, the prediction vector, and intermediate computations of internal layers when querying an input record. However, in the black-box setting, the attacker has black-box access to the classifier and only receives the prediction vector of the input record. When the target models are classifiers, based on the different information provided by the prediction vector, black-box MIAs can be further divided into three categories, as shown in Table 1.

Compared to white-box MIAs, an attacker of black-box MIAs gets limited information to attack the target ML model. However, if the black-box MIAs can work, then they would be more dangerous compared with white-box MIAs, because the attacker can breach the membership privacy with limited knowledge. The black-box MIAs on classification models where an attacker is only given the knowledge of a prediction label is the most dangerous attack among all MIAs, because the attacker can attack the model with the most limited knowledge. Most existing works study black-box MIAs on classification models with the knowledge of full confidence scores. There are many opportunities to study white-box attacks and black-box attacks with different levels of knowledge.

3.3 Membership Inference Attack Approaches

Machine learning (ML) models such as DNNs are often overparameterized, which means that they have sufficient capacity to memorize information about their training dataset [16, 144, 183, 234]. Moreover, the training datasets are finite in size, and ML models are trained over multiple (often tens to hundreds) epochs on the same instances repeatedly. Consequently, ML models exhibit a different behavior on training data records (i.e., members) versus test data records (i.e., non-members), and also in the model's parameters that store statistically correlated information about specific data records in their training dataset [144, 147, 181]. For example, a classification model would classify a training data record to its true class with a high confidence score while classifying a test data record to its true class with a relatively small confidence. These different behaviors of ML models enable an attacker of MIAs to build attack models to distinguish members from non-members of the training dataset. Based on the construction of the attack model, there are two major types of MIA approaches, i.e., binary classifier-based attack approaches and metric-based attack approaches.

3.3.1 Binary Classifier-based Membership Inference Attacks. Essentially, a binary classifier-based MIA involves training a binary classifier, which can distinguish a target model's behavior of its training members from the non-members. The challenge is how to train such a binary classifier. An effective technique called **shadow training** proposed by Shokri et al. [181] is the first and perhaps the most widely used approach for training a binary classifier-based MIA. The main idea is an attacker can create multiple shadow models to mimic the behavior of the target model, because the attacker is assumed to know the structure and the learning algorithm of the target model. For these shadow models, the attacker has their training datasets and test datasets, and thus can

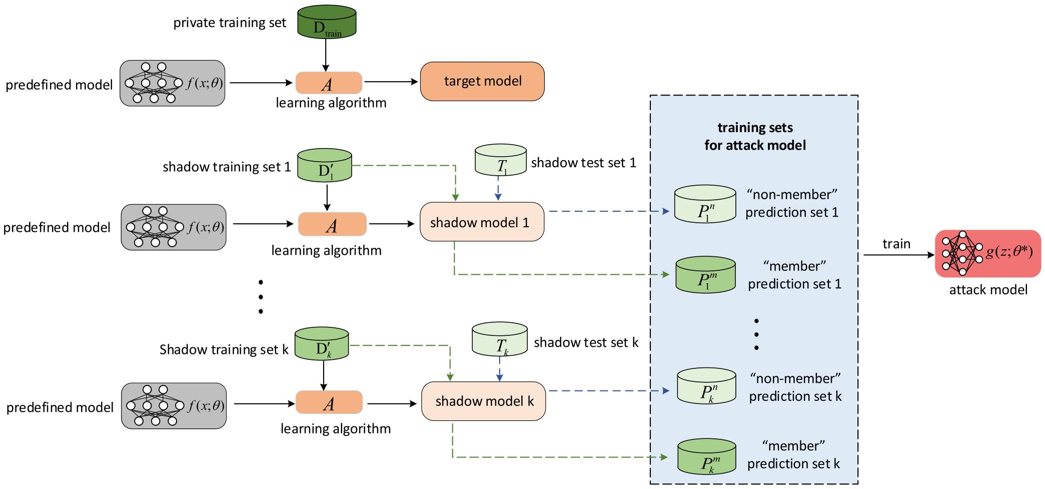
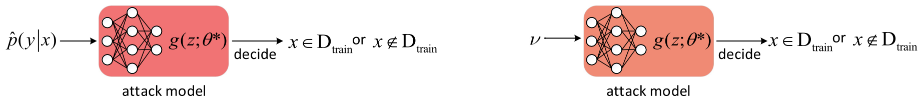


Fig. 4. Overview of the shadow training technique.

construct a dataset containing features and ground truth of membership of the training and test data records. Based on the constructed dataset, the attacker can train the binary classifier-based attack model.

Figure 4 shows how to use shadow training to train a binary classifier-based attack model to implement MIAs on classification models. D_{train} is a private training dataset, which is used for training the target classifier using the learning algorithm $\mathcal{A}$. $D'_1, \ldots, D'_k$ are shadow training datasets that are disjoint from the private training dataset D_{train}. Each shadow training dataset contains data records coming from the same data distribution as the training members in D_{train}, because the attacker is assumed to know the distribution of training data. The attacker first trains k shadow models using shadow training datasets $D'_1, \ldots, D'_k$ and the learning algorithm $\mathcal{A}$. Each shadow model is trained in such a way to mimic the behavior of the target model. $T_1, \ldots, T_k$ are shadow test datasets that are disjoint from $D'_1, \ldots, D'_k$. The more shadow models, the more accurate the attack model can be, because more shadow models can provide more training fodder for the attack model [181]. When the shadow models finish training, the attacker queries each of the shadow models using its shadow training dataset and shadow test dataset to obtain the outputs, which are prediction vectors of each data record. For each shadow model, the prediction vector of each record in the shadow training dataset is labeled "member" and the prediction vector of each record in the shadow test dataset is labeled "non-member." Thus, the attacker can construct k "member" datasets and k "non-member" datasets, which jointly consist of the training datasets for the attack model. Finally, the problem of recognizing the complex relationship between members and non-members of the training dataset is converted into a binary classification problem. Because binary classification is a standard ML task, the attacker can use any state-of-the-art ML framework to build the attack model.

The shadow training technique can be used for training both white-box and black-box attack models. In both MIAs, the training procedure of the attack models is the same as shown in Figure 4. However, because adversarial knowledge available for an attacker of black-box and white-box MIAs is different, the attacker can collect different amounts of information about the training members and non-members under the different settings. In the black-box setting, the attacker only has black-box access to the target model, which means the attacker can only receive the prediction vector of an arbitrary input record when querying the target model. Thus, when querying



(a) Binary classifier based black-box MIAs.

(b) Binary classifier based white-box MIAs.

Fig. 5. Overview of binary classifier-based attack models in black-box and white-box settings. In the membership inference phase, the black-box attack model only takes the prediction vector $\hat{p}(y \mid \boldsymbol{x})$ as input and outputs the membership status of the data record. However, the white-box attack model can take the flat vector $\boldsymbol{\nu}$ containing much more information of the data record as input and outputs its membership status.

the shadow models using their own shadow training datasets and test datasets, the attacker only collects the prediction vectors of each data record. However, in the white-box setting, the attacker has full access to the target model, which means the attacker can observe the intermediate computations at hidden layers and the prediction vector of an arbitrary input record. Thus, in the white-box setting, when querying the shadow models, the attacker can collect prediction vectors in addition to the intermediate computations of each data record. Compared to black-box MIAs, the attacker of white-box MIAs gets much more information to build the attack model. Next, we show more details of how an attacker constructs the attack model in both settings.

Binary Classifier-based MIA in Black-box Setting. Datasets $P_1^{\mathrm{m}}, \ldots, P_k^{\mathrm{m}}$ are "member" datasets that contain prediction vectors of the data records in the shadow training datasets. Datasets $P_1^{\mathrm{n}}, \ldots, P_k^{\mathrm{n}}$ are "non-member" datasets that contain prediction vectors of the data records in the shadow test datasets. We denote a prediction vector as $\hat{p}\,(y \mid \boldsymbol{x})$, "member" as 1, and "non-member" as 0. Then, each "member" dataset and "non-member" dataset is represented as follows:

$$P_i^{\mathrm{m}} = \left\{\hat{p}\left(y \mid \boldsymbol{x}^{(t)}\right), 1\right\}_{t=1}^{N_i^{\mathrm{m}}}, \tag{6}$$

$$P_i^{\mathrm{n}} = \left\{\hat{p}\left(y \mid \boldsymbol{x}^{(t)}\right), 0\right\}_{t=1}^{N_i^{\mathrm{n}}}. \tag{7}$$

For an binary classifier $g(\boldsymbol{z};\theta)$ (assuming the classifier is a DNN classifier), the attacker uses an SGD algorithm to find parameters θ^* that minimize the following objective function:

$$\mathcal{R}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}\left(\mathcal{I}(\boldsymbol{x}), g(\hat{p}(y \mid \boldsymbol{x}); \theta)\right), \tag{8}$$

where N is the total number of shadow data records, $\mathcal{L}(\cdot,\cdot)$ is a binary cross-entropy loss function, and $I(\cdot)$ is an indicator function as follows:

$$\mathcal{L}(y, p) = -(y\log(p) + (1-y)\log(1-p)), \tag{9}$$

$$\mathcal{I}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in P^{\mathrm{m}}, \\ 0 & \text{if } \boldsymbol{x} \notin P^{\mathrm{m}}. \end{cases} \tag{10}$$

After the binary classifier is trained, the attacker can use it as the attack model to implement MIAs on arbitrary data records. Figure 5(a) demonstrates a black-box MIA using the trained attack model. The binary classifier $g(\boldsymbol{z};\theta^*)$ takes the prediction vector $\hat{p}(y \mid \boldsymbol{x})$ of a data record as input and outputs whether this record is in D_{train} or not.

Binary Classifier-based MIA in White-box Setting. The attacker can get all the information to implement MIAs in a white-box setting. When querying a shadow model on an input record of the shadow datasets, the attacker can collect the prediction vector $\hat{p}(y \mid \boldsymbol{x})$, the intermediate computation $h(\boldsymbol{x};\theta_i)$ at each hidden layer, the loss $\mathcal{L}(y,\hat{p}(y\,|\,\boldsymbol{x}))$, and the gradient of the loss with respect to the parameters of each layer $\frac{\partial\mathcal{L}}{\partial\theta_i}$ of the input record. In this case, $P_1^m,\ldots,P_k^m$ are "member" datasets that contain the above computations of each data record in the shadow training sets, and $P_1^n,\ldots,P_k^n$ are "non-member" datasets that contain the computations of each data record in the shadow test datasets. The attacker then concatenates all the computations of each data record into a flat vector as follows:

$$\boldsymbol{v} = \left(\frac{\partial\mathcal{L}}{\partial\theta_1}, h\left(\boldsymbol{x};\theta_1\right),\ldots,\frac{\partial\mathcal{L}}{\partial\theta_i}, h\left(\boldsymbol{x};\theta_i\right), \hat{p}(y\,|\,\boldsymbol{x}), \mathcal{L}(y;\hat{p}(y\,|\,\boldsymbol{x}))\right). \tag{11}$$

Then, each "member" dataset and "non-member" dataset is represented as follows:

$$P_i^{\mathrm{m}} = \{\boldsymbol{v},1\}_{t=1}^{N_i^{\mathrm{m}}}, \tag{12}$$

$$P_i^{\mathrm{n}} = \{\boldsymbol{v},0\}_{t=1}^{N_i^{\mathrm{n}}}. \tag{13}$$

The structure of the binary classifier-based attack model in the white-box setting is usually different from that in the black-box setting, because the input of the attack model in the two settings is very different. Nevertheless, the attack model is a binary classifier. For a binary classifier $g(\boldsymbol{z};\theta)$, the attacker uses SGD to find parameters θ^* that minimize the following objective function:

$$\mathcal{R}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(I(\boldsymbol{x}\in P^{\mathrm{m}}), g(\boldsymbol{v};\theta)). \tag{14}$$

Figure 5(b) demonstrates a white-box MIA using the trained attack model. The binary classifier $g(\boldsymbol{z};\theta^*)$ takes the flat vector $\boldsymbol{v}$ of a data record as input and outputs whether this record is in D_{train} or not.

3.3.2 Metric-based Membership Inference Attacks. Unlike binary classifier-based MIAs relying on training a binary classifier to recognize the complex relationship between members and non-members, metric-based MIAs are more simple and less computational. Metric-based MIAs make membership inference decisions for data records by first calculating metrics on their prediction vectors. The calculated metrics are then compared with a preset threshold to decide the membership status of the data record. Based on different metric options, there are four major types of metric-based MIAs, i.e., prediction correctness-based, prediction loss-based, prediction confidence-based, and prediction entropy-based attacks. We denote a metric-based MIA as $\mathcal{M}(\cdot)$, which codes members as 1 and non-members as 0. We introduce the detailed metric-based attack approaches as follows. Each approach follows the reference of the first paper that proposes or uses this attack approach.

Prediction Correctness-based MIA [227]. An attacker infers an input record $\boldsymbol{x}$ as a member if it is correctly predicted by the target model, otherwise the attacker infers it as a non-member. The intuition is that the target model is trained to predict correctly on its training data, which may not generalize well on the test data. The attack $\mathcal{M}_{\text{corr}}(\cdot,\cdot)$ is defined as follows:

$$\mathcal{M}_{\text{corr}}(\hat{p}(y\,|\boldsymbol{x}), y) = \mathbb{1}(\text{argmax }\hat{p}(y\,|\boldsymbol{x}) = y), \tag{15}$$

where $\mathbb{1}(\cdot)$ is an indicator function as follows:

$$\mathbb{1}(A) = \begin{cases} 1 & \text{if event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Prediction Loss-based MIA [227]. An attacker infers an input record as a member if its prediction loss is smaller than the average loss of all training members, otherwise the attacker infers it as a non-member. The intuition is that the target model is trained on its training members by minimizing their prediction loss. Thus, the prediction loss of a training record should be smaller than the prediction loss of a test record. The attack $\mathcal{M}_{\text{loss}}(\cdot,\cdot)$ is defined as follows:

$$\mathcal{M}_{\text{loss}}(\hat{p}(y\,|\boldsymbol{x}), y) = \mathbb{1}(\mathcal{L}(\hat{p}(y\,|\boldsymbol{x}); y) \leq \tau), \tag{17}$$

where $\mathcal{L}(\cdot)$ is the cross-entropy loss function and τ is a preset threshold.

Prediction Confidence-based MIA [174]. An attacker infers an input record as a member if its maximum prediction confidence is larger than a preset threshold, otherwise the attacker infers it as a non-member. The intuition is that the target model is trained by minimizing prediction loss over its training data, which means the maximum confidence score of a training member's prediction vector should be close to 1. The attack $\mathcal{M}_{\text{conf}}(\cdot)$ is defined as follows:

$$\mathcal{M}_{\text{conf}}(\hat{p}(y\,|\boldsymbol{x})) = \mathbb{1}(\max \hat{p}(y\,|\boldsymbol{x}) \geq \tau). \tag{18}$$

Prediction Entropy-based MIA [174]. An attacker infers an input record as a member if its prediction entropy is smaller than a preset threshold, otherwise the attacker infers it as a non-member. The intuition is that the prediction entropy distributions between training and test data are very different. The target model usually has a larger prediction entropy on its test data than its training data. The entropy of a prediction vector $\hat{p}(y\,|\,\boldsymbol{x})$ is defined as follows:

$$H(\hat{p}(y\,|\boldsymbol{x})) = -\sum_{i} p_i \log(p_i), \tag{19}$$

where p_i is the confidence score in $\hat{p}(y\,|\,\boldsymbol{x})$. The attack $\mathcal{M}_{\text{entr}}(\cdot)$ is then defined as follows:

$$\mathcal{M}_{\text{entr}}(\hat{p}(y\,|\boldsymbol{x})) = \mathbb{1}(H(\hat{p}(y\,|\boldsymbol{x})) \leq \tau). \tag{20}$$

Modified Prediction Entropy-based MIA [185]. The authors in Reference [185] argue that the existing prediction entropy-based MIA does not consider any information about the ground truth label, which might misclassify members and non-members. For example, a totally wrong classification with a probability score of one leads to zero prediction entropy value for an input record. The existing prediction entropy-based MIA will classify the record as a member. However, the record with a totally wrong classification is highly likely a non-member. Thus, they propose a modified prediction entropy metric that can leverage the information of the ground truth label as follows:

$$MH(\hat{p}(y\,|\boldsymbol{x}), y) = -(1-p_y)\log(p_y) - \sum_{i \neq y} p_i \log(1-p_i), \tag{21}$$

where p_y is the confidence score of the ground truth label. Then, the attack $\mathcal{M}_{\text{Mentr}}(\cdot,\cdot)$ is defined as follows:

$$\mathcal{M}_{\text{Mentr}}(\hat{p}(y\,|\boldsymbol{x}), y) = \mathbb{1}(MH(\hat{p}(y\,|\boldsymbol{x}); y) \leq \tau). \tag{22}$$

3.4 Membership Inference Attacks on Different ML models

Since the first work [181] proposed MIAs on classification models, there has been an increasing number of studies investigating MIAs on classification models as well as other ML models (e.g., generative models). In this section, we select a few pieces of literature to introduce MIAs on specific ML models, including classification models, generative models, embedding models, and regression models. Each of the selected papers either proposes new MIAs or is the first to investigate the membership privacy risks on a specific ML model or under a unique adversarial knowledge setting.

MIAs on Classification Models. Currently, many of the MIAs focus on classification models. In this case, an attacker aims to infer whether a data instance was used to train a target classifier. Shokri et al. [181] conducted the pioneering work to propose the first MIA on classification models.

They invented a shadow training technique to train a binary classifier-based attack model in a black-box setting. Salem et al. [174] relax two main assumptions of the shadow training technique in Reference [181], i.e., multiple shadow models and knowledge of the training data distribution. They argue that the two assumptions are relatively strong, which heavily limit the applicable scenarios of MIAs against ML models. They show that even with one single shadow model, the attacker can achieve comparable attack performance compared to using multiple shadow models. They also propose a data transferring attack where a dataset used to train the shadow model is not required to have the same distribution as the target model's private training dataset. Also, the shadow model is not required to have the same structure as the target model. Besides extending existing binary classifier-based MIAs in Reference [181], they propose two metric-based attacks leveraging the highest confidence score and prediction entropy. Yeom et al. [227] also propose two metric-based MIAs, i.e., the prediction correctness-based MIA and the prediction loss-based MIA. Compared to binary classifier-based MIAs, metric-based MIAs are much simpler and have a smaller computation cost. Long et al. [123, 124] investigate MIAs on ML models that are not overfitted to their training data. They propose a generalized MIA that can identify the membership of particular vulnerable records. The intuition is that some records have unique influences on the target model, even when the model is well-generalized. An attacker can exploit the unique influences of particular data records as an indicator of their presence in the training dataset. They show that the vulnerable records can be inferred correctly on well-generalized models even if the gap in training and testing accuracy is smaller than 1%.

An attacker of the above MIAs is given full confidence scores of a target record to infer the membership status of the record. Li and Zhang [112] and Choquette et al. [30] study MIAs in a more restricted scenario where the target model only provides the predicted label to the attacker. Li and Zhang [112] propose two label-only MIAs, i.e., a transfer-based MIA and a perturbation-based MIA. The transfer-based MIA aims to construct a shadow model to mimic the target model. The intuition is that if the shadow model is similar enough to the target model, then the shadow model's confidence scores on an input record will indicate its membership. The perturbation-based attack aims to add crafted noise to the target record to turn it into an adversarial example. The intuition is that it is harder to perturb a member record to a different class than a non-member instance. Thus, the magnitude of the perturbation can be used to distinguish members from non-members. Choquette et al. [30] also propose two label-only MIAs, i.e., data augmentation-based MIA and decision boundary distance-based MIA. For a target record, a data augmentation-based attack creates additional data records via different data augmentation strategies. The additional data records are then used to query the target model and the attacker can collect all the predicted labels. The attack intuition is that many models use data augmentation during the training process. Thus, a member record's augmented versions are less likely to change their predicted label. The decision boundary-based attack estimates a record's distance to the model's boundary and decides it is a member if its distance is larger than a threshold. The intuition is similar to Li and Zhang's [112] perturbation-based attack. The success of label-only MIAs demonstrate that ML models can be more vulnerable to MIAs than we expect.

While the above MIAs focus on a black-box setting, Nasr et al. [147] first propose white-box MIAs, where an attacker knows internal parameters of the target model. Their white-box MIAs can be considered as an extension of the binary classifier-based MIAs in black-box settings. Compared to black-box MIAs, white-box MIAs try to improve the attack performance by leveraging an input record's intermediate computations through the target model. They use the gradient of an input's prediction loss with regard to the target model's parameters as additional features to infer the membership of the record. The intuition is that the gradients of a training member's loss over the model's parameters is distinguishable from non-members through the training of the SGD

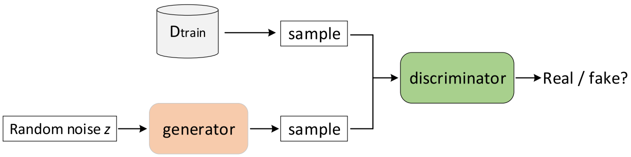


Fig. 6. Architecture of a generative adversarial network.

algorithm. However, Leino and Fredrikson [107] point out that the white-box setting in Nasr et al.'s paper [147] is too strong, which deviates from most settings of MIAs. Nasr et al. [147] assume that an attacker knows a significant portion of the target model's private training dataset, while the attacker is often assumed to only have a shadow dataset that is disjoint from the private training dataset. Thus, Leino and Fredrikson propose an effective white-box MIA that does not require any of the target model's training members. The attack intuition is that the membership information can be leaked through a target model's idiosyncratic use of features. Features distributed differently in the training data than in the true distribution can provide evidence for membership. They first build a Bayes-optimal attack assuming the target model is a simple linear softmax model. When the target is a DNN model, they approximate each layer as a local linear model, which is then applied to the Bayes-optimal attack. The attacks on different layers are then combined to compute the final membership decision.

MIAs on Generative Models. Besides classification models, MIAs are also investigated on generative models. Currently, MIAs on generative models focus on GANs, which is the most popular generative model. We first describe the architecture of a GAN and then introduce specific MIAs on GANs.

Figure 6 depicts the architecture of a GAN. A GAN consists of two competing neural network modules, a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$, which are trained to compete against each other. MIAs on generative models aim to identify whether a data record was used to train a generator or not, which is more challenging than MIAs on classification models. Unlike classification models, an attacker of an MIA on a generative model does not obtain confidence scores or prediction labels that are related to the target data record from the generative models. This means the attacker has few clues for implementing MIAs. Moreover, current GAN models often encounter model dropping and mode collapse, leading to a problem of underrepresenting certain data records, which poses additional attack difficulty to the attacker.

Hayes et al. [60] introduce the first MIA on generative models in both black-box and white-box settings. The attack intuition is that the discriminator of GANs is more confident to output a higher confidence value on training members, as it is trained to learn statistical differences between training data and generated data. In the white-box setting, an attacker puts all data records into the discriminator that will output confidence scores for each record corresponding to the probability of being a member. The attacker sorts these probability values in descending order and picks the first half records as members. In the black-box setting, the attacker collects generated records from the generator and uses them to train a local GAN to mimic the target GAN. After the local GAN has been trained, the attacker implements MIAs using the discriminator of the local GAN following the same attack approach as in the white-box setting.

Hilprecht et al. [66] propose two MIAs on generative models. One is Monte Carlo integration attack designed for GANs in the black-box setting, and the other is reconstruction attack designed for VAEs in the white-box setting. The Monte Carlo integration attack exploits generated records that are within a small distance of a target record to approximate the probability that this record is a member via Monte Carlo integration [167]. The attack intuition is that the generator of GANs should be able to produce synthetic records that are close to the training members if GANs overfit. The reconstruction attack directly makes use of the loss function of VAEs to calculate the reconstruction error of the target member, and the attack intuition is that training members should have smaller reconstruction errors compared to that of non-members. In addition to MIAs for a single record, Hilpreche et al. [66] introduce the concept of set membership inference where the attacker tries to identify whether a set of records belongs to the training dataset or not. Liu et al. [118] propose co-membership inference, which essentially is the same as the set membership inference proposed by Hilpreche et al. [66]. Liu et al.'s [118] proposed attack begins with attacking a single target record and then extends to a set of records. For a given record and a generator of the target GAN, the attacker first optimizes a neural network to reproduce the latent variable such that the generator can generate synthetic records nearly matching the target record. The attack intuition is that if a record belongs to the training dataset, then the attacker is able to reproduce similar synthetic records close to it. The attacker then measures the L2 distance between the synthetic record and the target record and infers the target record is a member if the distance is smaller than a threshold. This attack method is different from the attacks in Reference [66], because it requires retraining new neural networks for different input data, while the Monte Carlo integration attack and the reconstruction attack in Reference [66] only need fixed synthetic records of the generator.

Chen et al. [21] propose a generic MIA on generative models that is applicable to all adversarial knowledge settings, from full black-box to full white-box settings. For a target record, the attacker tries to reconstruct a synthetic record that is closest to the target record. The attacker simply finds the synthetic record generated from the generator if possible. Otherwise, the attacker makes use of optimization algorithms to reconstruct the synthetic record. The distance between the reconstructed record and the target record is then used for calculating the probability that this target record is a member. The attack intuition is that the generator should be able to generate more similar samples for members than non-members. To make a more accurate probability estimation, they train a reference GAN with a relevant but disjoint dataset to calibrate the reconstruction error (i.e., the distance). The attacker decides the target record is a member when the calibrated reconstruction error is smaller than a threshold. The MIAs introduced above have been evaluated on state-of-the-art generative models, such as DCGAN [159], VAEGAN [102], PGGAN [91], WGANGP [56], and MEDGAN [29]. Notably, References [21, 66, 118] report that VAEs are more susceptible to MIAs compared to GANs, because VAEs are more prone to overfiting to their training data than GANs.

MIAs on Embedding Models. Embeddings are mathematical functions that map raw objects (such as words, sentences, and graphs) to real valued vectors with the aim of capturing and preserving important semantic information about the underlying objects. Embeddings have been successfully applied to various domains including natural language processing [90], social networks [55], movie feedback [63], and location [33]. Song and Raghunathan [182] introduced the first MIAs on word embedding and sentence embedding models. Unlike classification models whose training data consist of input feature vectors and class labels, text embedding models are trained on sequences of words or sentences. Thus, the goal of MIAs on text embedding models is to infer the membership of a sliding window of words or a pair of sentences. Song and Raghunathan [182] propose a metric-based MIA that leverages similarity scores of a sliding window of words or a pair of sentences to infer their membership status. The attack intuition is that words and sentences in the context used for training will be more similar to each other than those of non-members. Mahloujifar et al. [127] demonstrate that MIAs on embedding models can work even when the embedding layer of the embedding models is not exposed to the attacker. Duddu et al. [42] introduced the first

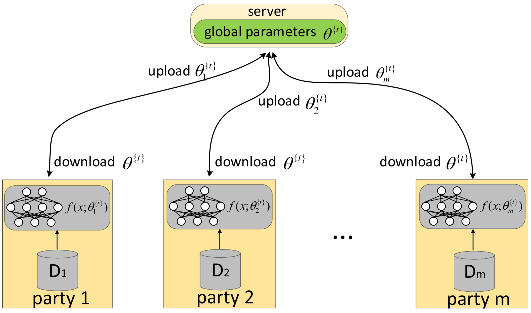

server
global parameters $\theta^{\{t\}}$
upload $\theta_1^{\{t\}}$
upload $\theta_2^{\{t\}}$
upload $\theta_m^{\{t\}}$
download $\theta^{\{t\}}$
download $\theta^{\{t\}}$
download $\theta^{\{t\}}$
$f(x;\theta_1^{\{t\}})$
$f(x;\theta_2^{\{t\}})$
$f(x;\theta_m^{\{t\}})$
D1
D2
Dm
party 1
party 2
...
party m

Fig. 7. Overview of the federated averaging algorithm.

MIA on graph embedding models. They propose a shadow model attack in a black-box setting where the embedding layer of the embedding model is used in **graph neural networks (GNN)** for node classification problems. The shadow model attack uses the shadow training technique and is essentially a binary classifier. They also propose a confidence score attack in a white-box setting where the attacker can directly access the graph embedding model. The attack intuition is that graph nodes with higher output confidence prediction are more likely to be members of the graph.

MIAs on Regression Models. Gupta et al. [57] introduced the first MIAs on deep regression models. They focus on age-prediction problems where regression models predict a person's age from their brain MRI scan. Because their work focuses on demonstrating the vulnerability of deep regression models to MIAs, they assume an attack is given white-box access to the target model and has access to some training members of the private training dataset. The attack model is a binary classifier that leverages features of gradients of parameters, activation, predictions, and labels of target records to infer their membership status.

3.5 Membership Inference Attacks against Federated Learning

Federated learning (FL) has recently emerged as an alternative to conventional centralized learning where all training data is pooled and an ML model is trained on this joint pool. FL allows multiple parties to jointly train an ML model in an interactive manner. It is an attractive framework for training ML models without direct access to diverse training data owned by different parties, especially for privacy-sensitive tasks [130, 131]. The above MIAs assume the target ML models are trained in a centralized manner, but FL provides interesting new avenues for MIAs. The success of MIAs against FL can shed light on how FL reveals sensitive information and provides an insight that FL may not always provide sufficient privacy guarantees. To better understand membership privacy risks in FL, we first introduce the **federated averaging (FedAvg)** [130] algorithm, which is the first and the most widely used FL algorithm. Existing MIAs against FL mainly focus on FedAvg.

Figure 7 shows the FedAvg algorithm [130]. During multiple rounds of communication between server and parties, a central model is trained. At each communication round, the server distributes the current central model to local parties. The local parties then perform local optimization using their own data. To minimize communication, parties might update the local model for several epochs during a single communication round. Next, the optimized local models are sent back to the server, who averages them to allocate a new central model. The performance of the new central model decides the training is either stopped or a new communication round starts. In FL, parties never share data, only their model weights or gradients.

In the context of FL, the attacker can be either the central server or a certain number of parties. Melis et al. [133] introduced the first MIA against FL. They focus on a text classification problem and the target models are recurrent neural networks with a word-embedding layer to transform inputs into a lower-dimensional vector representation via an embedding matrix. The embedding matrix is treated as a parameter of the global model and optimized collaboratively. During training, the embedding layer's gradient is sparse with respect to the input words. This means for a given batch of text, the embedding is updated only with the words that appear in the batch, and the gradients of the other words are zeros. The attacker can observe the non-zero gradients to infer which words occur in the training dataset. Truex et al. [200] introduce an MIA against heterogeneous FL where each party trains a local model and only shares confidence scores when making predictions for a new record. They assume different parties have very different datasets, which leads to sufficiently different decision boundaries for different parties. The decision boundary differences reveal the underlying training data, enabling an insider attacker to infer whether a data record is in the local datasets of the other parties.

The attacker of the above MIAs against FL passively follows the FL protocol to infer membership of a data record. However, Nasr et al. [147] argue that the attacker can also actively tamper with the FL training to achieve better attack performance. They propose an MIA called gradient ascent attack, which intentionally updates the local model parameters in the direction of increasing the loss on a target data record. The attack exploits the fact that SGD optimization updates model parameters in the opposite direction of the gradient of the loss. If the target record is a member, then applying the gradient ascent attack will trigger the target model to minimize the loss of this record by descending the model's gradient and nullify the effect of the attacker's ascent. However, for a non-member record, the target model will not change their gradient on it explicitly, as this record does not influence the training loss function.

Note that the attacker of an MIA in FL infers whether a data record was used to train the global model but does not infer whether the data record was used to train a particular local model. This is because MIAs on ML models aim to identify target models' members from non-members of the training dataset. In FL, each party contributes its local training data to jointly train an ML model, and thus the training dataset for the FL model consists of all the local data records. Recently, Hu et al. [74] propose source inference attacks that can determine which party owns a training record in FL. They argue that existing MIAs in FL ignore the source of a training member, i.e., the information of the party owning the training member. However, it is essential to explore source privacy in FL beyond membership privacy, because the leakage of such information can lead to further privacy issues. For instance, in the scenario where multiple hospitals jointly train an FL model for the COVID-19 diagnosis, MIAs can only reveal who have been tested for COVID-19, but the further identification of the source hospital where the people are from will make them more prone to discrimination, especially when the hospital is in a high-risk region or country [37]. They demonstrate that a malicious server in FedAvg [130] can implement source inference attacks effectively and non-intrusively. The intuition of their proposed source inference attacks is that the local model behaves differently on its local training data and the training data of other parties, which enables the malicious server to leverage the prediction loss of local models to steal non-trivial source information of the training members.

3.6 Taxonomies of Membership Inference Attacks

To give readers a general picture of MIAs and help readers find the most relevant papers easily, we create a taxonomy of MIAs on ML models in Figure 8. In this taxonomy, we categorize all released papers of MIAs based on different target models, adversarial knowledge, attack approaches, training paradigms, and domains. Specifically, for papers in the category of target model level, we

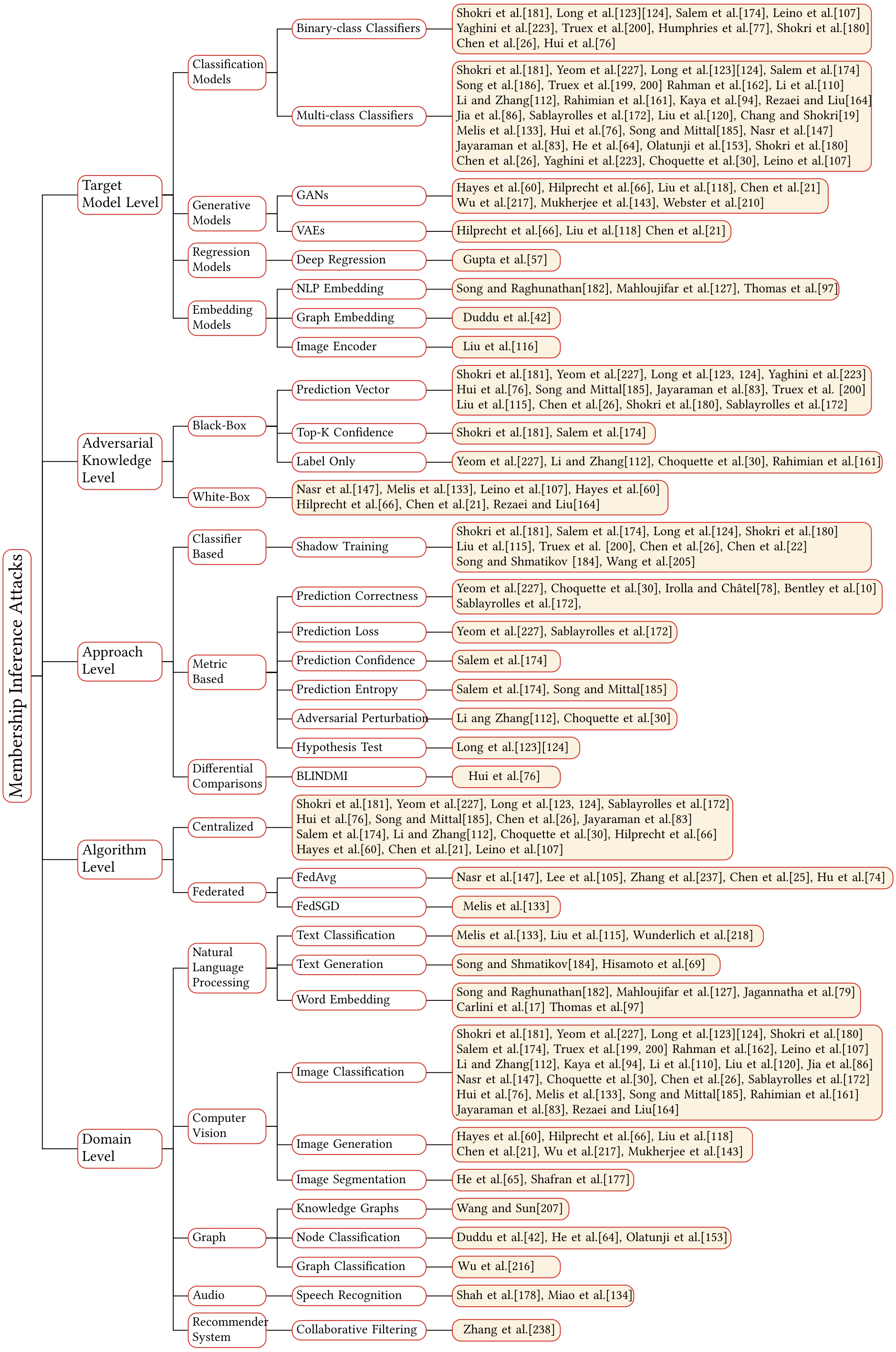

Membership Inference Attacks
Target Model Level
Classification Models
Binary-class Classifiers
Shokri et al.[181], Long et al.[123][124], Salem et al.[174], Leino et al.[107] Yaghini et al.[223], Truex et al.[200], Humphries et al.[77], Shokri et al.[180] Chen et al.[26], Hui et al.[76]
Multi-class Classifiers
Shokri et al.[181], Yeom et al.[227], Long et al.[123][124], Salem et al.[174] Song et al.[186], Truex et al.[199, 200] Rahman et al.[162], Li et al.[110] Li and Zhang[112], Rahimian et al.[161], Kaya et al.[94], Rezaei and Liu[164] Jia et al.[86], Sablayrolles et al.[172], Liu et al.[120], Chang and Shokri[19] Melis et al.[133], Hui et al.[76], Song and Mittal[185], Nasr et al.[147] Jayaraman et al.[83], He et al.[64], Olatunji et al.[153], Shokri et al.[180] Chen et al.[26], Yaghini et al.[223], Choquette et al.[30], Leino et al.[107]
Generative Models
GANs
Hayes et al.[60], Hilprecht et al.[66], Liu et al.[118], Chen et al.[21] Wu et al.[217], Mukherjee et al.[143], Webster et al.[210]
VAEs
Hilprecht et al.[66], Liu et al.[118] Chen et al.[21]
Regression Models
Deep Regression
Gupta et al.[57]
Embedding Models
NLP Embedding
Song and Raghunathan[182], Mahloujifar et al.[127], Thomas et al.[97]
Graph Embedding
Duddu et al.[42]
Image Encoder
Liu et al.[116]
Adversarial Knowledge Level
Black-Box
Prediction Vector
Shokri et al.[181], Yeom et al.[227], Long et al.[123, 124], Yaghini et al.[223] Hui et al.[76], Song and Mittal[185], Jayaraman et al.[83], Truex et al. [200] Liu et al.[115], Chen et al.[26], Shokri et al.[180], Sablayrolles et al.[172]
Top-K Confidence
Shokri et al.[181], Salem et al.[174]
Label Only
Yeom et al.[227], Li and Zhang[112], Choquette et al.[30], Rahimian et al.[161]
White-Box
Nasr et al.[147], Melis et al.[133], Leino et al.[107], Hayes et al.[60] Hilprecht et al.[66], Chen et al.[21], Rezaei and Liu[164]
Approach Level
Classifier Based
Shadow Training
Shokri et al.[181], Salem et al.[174], Long et al.[124], Shokri et al.[180] Liu et al.[115], Truex et al. [200], Chen et al.[26], Chen et al.[22] Song and Shmatikov [184], Wang et al.[205]
Metric Based
Prediction Correctness
Yeom et al.[227], Choquette et al.[30], Irolla and Châtel[78], Bentley et al.[10] Sablayrolles et al.[172],
Prediction Loss
Yeom et al.[227], Sablayrolles et al.[172]
Prediction Confidence
Salem et al.[174]
Prediction Entropy
Salem et al.[174], Song and Mittal[185]
Adversarial Perturbation
Li ang Zhang[112], Choquette et al.[30]
Hypothesis Test
Long et al.[123][124]
Differential Comparisons
BLINDMI
Hui et al.[76]
Algorithm Level
Centralized
Shokri et al.[181], Yeom et al.[227], Long et al.[123, 124], Sablayrolles et al.[172] Hui et al.[76], Song and Mittal[185], Chen et al.[26], Jayaraman et al.[83] Salem et al.[174], Li and Zhang[112], Choquette et al.[30], Hilprecht et al.[66] Hayes et al.[60], Chen et al.[21], Leino et al.[107]
Federated
FedAvg
Nasr et al.[147], Lee et al.[105], Zhang et al.[237], Chen et al.[25], Hu et al.[74]
FedSGD
Melis et al.[133]
Domain Level
Natural Language Processing
Text Classification
Melis et al.[133], Liu et al.[115], Wunderlich et al.[218]
Text Generation
Song and Shmatikov[184], Hisamoto et al.[69]
Word Embedding
Song and Raghunathan[182], Mahloujifar et al.[127], Jagannatha et al.[79] Carlini et al.[17] Thomas et al.[97]
Computer Vision
Image Classification
Shokri et al.[181], Yeom et al.[227], Long et al.[123][124], Shokri et al.[180] Salem et al.[174], Truex et al.[199, 200] Rahman et al.[162], Leino et al.[107] Li and Zhang[112], Kaya et al.[94], Li et al.[110], Liu et al.[120], Jia et al.[86] Nasr et al.[147], Choquette et al.[30], Chen et al.[26], Sablayrolles et al.[172] Hui et al.[76], Melis et al.[133], Song and Mittal[185], Rahimian et al.[161] Jayaraman et al.[83], Rezaei and Liu[164]
Image Generation
Hayes et al.[60], Hilprecht et al.[66], Liu et al.[118] Chen et al.[21], Wu et al.[217], Mukherjee et al.[143]
Image Segmentation
He et al.[65], Shafran et al.[177]
Graph
Knowledge Graphs
Wang and Sun[207]
Node Classification
Duddu et al.[42], He et al.[64], Olatunji et al.[153]
Graph Classification
Wu et al.[216]
Audio
Speech Recognition
Shah et al.[178], Miao et al.[134]
Recommender System
Collaborative Filtering
Zhang et al.[238]

Fig. 8. Taxonomy of membership inference attacks.

further categorize them based on specific types of the target ML models, i.e., classification models, generative models, regression models, and embedding models. For papers in the category of adversarial knowledge, we further divide them based on whether the MIAs is black-box or white-box attacks. For papers in the category of attack approach level, we further categorize the MIAs into classifier-based attacks, metric-based attacks, and differential comparisons-based attacks. For papers in the category of algorithm level, we further divide them based on whether the target models are trained in a centralized manner or a federated manner. Last, for papers in the category of domain level, we further categorize them based on the specific target domain that the MIAs involve, i.e., **natural language processing (NLP), computer vision (CV)**, graph, audio, and recommender system. Note that Figure 8 not only gives general taxonomies for MIAs according to the above criteria, but also provides detailed characteristics for specific categorized papers. For example, for papers under the category of metric-based attacks, readers can further find the specific metric-based attack approach proposed or involved in the relevant papers, e.g., the paper [174] of Salem et al. proposes the prediction confidence-based attack approach. In addition to categorizing all released papers of MIAs in Figure 8, we further select a few representative papers and list them with their characteristics in a table in the supplemental material due to the limited space of this article.

4 WHY MEMBERSHIP INFERENCE ATTACKS WORK

Conducting the theoretical analysis of why membership inference attacks can work is a very challenging task because of the high complexity existing in both training data and target models (especially for deep neural networks). There are some initial works [10, 45, 85, 227] formally formulating the problem of MIAs on ML models and providing some theoretical analysis, while the rigorous analysis of why MIAs can work is still in the infant stages. Because most existing literature provides explanations based on practical evaluations, in this section, we discuss why MIAs work from the perspective of the following three aspects mainly based on empirical reasoning.

Overfitting of Target Models. First, many papers [21, 107, 174, 181, 227] have pointed out that overfitting of the target ML models is the main factor contributing to the success of MIAs. An ML model is said to overfit to its training data when it performs much better on its training data than test data, i.e., the model cannot generalize well on its test data. The overfitting phenomenon of ML models is usually because of two reasons, i.e., the high model complexity and the limited size of the training dataset [12]. Deep learning models such as DNNs are often overparameterized with high complexity, which, on the one hand, enables them to learn effectively from big data, but, on the other hand, results in the fact that they may have unnecessarily high capacity of memorizing the noise or the details of a given training dataset [16, 144, 183, 234]. Moreover, ML models are trained using many (often tens to hundreds) epochs on the same instances repeatedly, rendering the training instances very prone to being memorized by the models. Also, a training dataset with a finite size often fails to represent the whole data distribution, which makes the ML model difficult to generalize to test data, behaving very differently on their training members and non-members. Because MIAs exploit the different behaviors of target ML models on their training data versus test data, ML models overfitted to their training data will be vulnerable to MIAs. Overfitting is sufficient to allow an attacker to perform non-trivial membership inference. For example, if a target classification model is overfitted to its training data that results in a difference (i.e., the generalization gap) between the test accuracy and the train accuracy of the classifier larger than 0, then the attacker can easily achieve an overall attack success rate larger than 50% (i.e., randomly guessing) by leveraging the prediction correctness-based attack approach as introduced in Section 3.3.2. Note that in Reference [10], Bentley et al. give a theorem

(Theorem 4.1) that also implies the overfitting of the target models can lead to the performance of an MIA better than randomly guessing (i.e., 50% **attack success rate (ASR)**). Theorem 4.1 is as follows, with the symbol denotations and the detailed proof available in the paper [10]:

Theorem 4.1 ([10]). *Given access to a model with generalization gap $g = p_0 - p_1 \geq 0$ (training accuracy minus testing accuracy) and the ratio of training dataset to input domain q, there exists a membership inference attack with expected attack success rate (ASR) at least:*

$$\begin{aligned} ASR &\geq \max\{q, 1-q, qp_0 + (1-q)(1-p_1)\}, \\ &\geq \max\{q, 1-q, \min\{q, 1-q\}(1+g)\}, \\ &\geq \tfrac{1}{2}. \end{aligned}$$

Types of Target Models. Second, the type of target model also plays an important role in the success of MIAs. In general, a target model whose decision boundary is unlikely to be drastically impacted by a particular data record will be more resilient to MIAs. For example, MIAs are evaluated on DNN models, logistic regression models, Naive Bayes models, k-nearest neighbor models, and decision tree models on seven datasets in [200]. The results show that the decision tree model has the highest attack precision for six datasets, and Naive Bayes models consistently show the lowest precision across all datasets. This is because a single training record only marginally affects the prediction decisions of a given class in Naive Bayes models. By contrast, a record that displays a unique feature can cause a decision tree to grow an entirely new branch, drastically changing the decision boundary. The decision tree models' sensitivity to single records makes MIAs more successful on them.

Diversity of Training Data. Last, if the training data is more representative, i.e., the training data can better represent the whole data distribution, then the target model will be less vulnerable to MIAs. This is because more representative training data can help the target ML model to generalize better on test data. For example, Reference [181] demonstrates a classification model has smaller and smaller attack precision when provided with more and more training records.

In conclusion, success of MIAs is directly related to three factors: (1) Overfitting of the target model; (2) type of the target model; (3) diversity of the target model's training data. Overfitting of the target model is the main reason why MIAs work. Moreover, different ML models remember different amounts of information about their training datasets due to their different structures and training dataset. This leads to different levels of vulnerability to MIAs, because different models have different levels of proneness to overfitting.

5 MEMBERSHIP INFERENCE DEFENSE ON MACHINE LEARNING MODELS

In this section, we introduce membership inference defenses on ML models. The existing defenses against MIAs fall into four main categories, i.e., confidence score masking, regularization, knowledge distillation, and differential privacy.

5.1 Confidence Score Masking

Confidence score masking is mainly used to mitigate black-box MIAs on classification models. It aims to hide the true confidence scores returned by the target classifier and thus mitigates the effectiveness of MIAs. There are three methods belonging to this defense category. The first method is that the target classifier does not provide a complete prediction vector but provides top-k confidence scores to the attacker of an MIA. For example, in a classification problem of 10 classes, the target classifier only provides the largest three confidence scores when the attacker queries an input record. The second method is that the target classifier only provides the prediction label when the attacker queries an input record. This method provides the most limited knowledge to the attacker. The last method is to add crafted noise to the prediction vector to hide the true confidence scores. The three methods do not need to retrain target classifiers and are only implemented on the prediction vectors, thus they will not influence the target model's accuracy.

Shokri et al. [181] evaluate the first two defense methods on a fully connected neural network-based classifier on two datasets. They find that restricting the prediction vector to topthree classes does not reduce the attack accuracy of their proposed shadow training-based attack. This finding is not surprising, because a later paper [174] demonstrates that a black-box binary classifier-based attack leveraging partial confidence scores can achieve similar attack performance compared to using complete prediction vectors. Shokri et al. [181] indeed show that returning only the classifier's predicted label will reduce the attack accuracy. However, as long as the generalization gap exists, a simple prediction correctness-based attack will always achieve better attack performance than random guessing. The label-only attacks proposed by Li and Zhang [112] and Choquette et al. [30] further investigate the membership privacy risks when an attacker gets access only to the target classifiers' predicted labels. Unfortunately, the attacker with only prediction labels can still achieve strong attack performance. Jia et al. [86] observe that when the attack model is a DNN-based binary classifier, it is vulnerable to adversarial examples. Thus, they leverage an adversarial machine learning technique [100] and propose a defense method called MemGuard. MemGuard adds a carefully crafted noise vector to the prediction vector and turns it into an adversarial example of the attack model. MemGuard does not influence the target models' prediction accuracy while effectively mitigating the black-box DNN-based attack to a random guess level. However, Song and Mittal [185] re-evaluate the effectiveness of Memguard using metric-based attacks and find that the defended models are still susceptible to membership attacks.

The advantage of confidence score masking is the simplicity of implementation. It directly works on the trained models' prediction vector and thus does not need to retrain the target model. It is a natural mitigation mechanism against the attacker who uses the complete prediction vector of the target classifier to implement MIAs. However, as we discussed above, confidence score masking might not provide enough privacy guarantees because the label-only attacks still work well, and Memguard is vulnerable to metric-based attacks.

5.2 Regularization

Regularization aims to reduce the overfitting degree of target models to mitigate MIAs. Therefore, regularization methods that can reduce the overfitting of ML models can be leveraged to defend against MIAs. Existing regularization methods including L2-norm regularization, dropout [188], data argumentation, model stacking, early stopping, label smoothing [191], adversarial regularization [146], and Mixup + MMD (Maximum Mean Discrepancy) [110] have been proposed and investigated as defense methods in many papers [73, 76, 93, 110, 146, 174, 179, 181, 185]. Among them, L2-norm regularization, dropout, data argumentation, model stacking, early stopping, label smoothing are classical regularization methods proposed to improve the generalizability of a learned ML model. They are initially proposed to reduce the overfitting of ML models, but they are shown to be quite effective in mitigating MIAs. This is because they help the learned model generalize better to test data and reduce the difference of the model's behaviors on its training data and test data. The adversarial regularization [146] and Mixup + MMD [110] are specially designed regularization techniques that aim to mitigate MIAs. The two proposed methods add new regularization terms to a target classifier's objective function during the training phase and force the classifier to generate similar output distributions for training members and non-members. Adversarial regularization [146] adds membership inference gain of the attack model as a new regularization term to the objective function of the target model during the training process. The target ML model needs to simultaneously minimize its classification loss and the attack model's accuracy. The target model is trained in such a way as to preserve its prediction accuracy while mitigating the attacker's performance. Mixup + MMD [110] adds the distance between the output distributions of members and non-members computed by **Maximum Mean Discrepancy (MMD)** [46] as a new regularization term to the objective function of the target classifier. The new regularization term forces the classifier to generate similar output distributions for its training members and non-members. As MMD tends to reduce the prediction accuracy of the classifier, the authors in Reference [110] propose to combine MMD with mix-up training [236] to preserve the prediction utility. Note that regularization methods not only work for classification models, and some methods can be used to mitigate MIAs on generation models. For example, Hayes et al. [60] and Hilprecht et al. [66] demonstrate that dropout can be leveraged as an effective defense method against MIAs on GANs.

Unlike confidence score masking, regularization defends against MIAs no matter whether an attacker is in a black-box or white-box setting. This is because regularization methods change not only the target models' output distribution but also their internal parameters, while methods of confidence score masking only modify models' prediction vectors. Although regularization methods are effective and widely applicable, one drawback of them is that they might not be able to provide satisfactory membership privacy-utility tradeoffs. For example, Shokri et al. [181] show that L2-norm regularization can mitigate the accuracy of MIAs to random-guess level when setting the regularization factor to relatively large values. However, this results in a significant reduction of the target model's prediction accuracy.

5.3 Knowledge Distillation

Knowledge distillation [5, 67] uses the outputs of a large teacher model to train a smaller student model to transfer knowledge from the large model to the small one. It allows the smaller student model to have similar accuracy to their teacher models [32]. Based on knowledge distillation, Shejwalkar and Houmansadr [179] propose **Distillation For Membership Privacy (DMP)** defense method. DMP requires a private training dataset and an unlabeled reference dataset. DMP first trains an unprotected teacher model and uses it to label data records in the unlabeled reference dataset. Then, DMP selects data records from the labeled reference dataset that have low prediction entropy to train the target model. The intuition of the selection is that such records are easy to classify and will not be significantly affected by the members of the private training dataset. DMP finally trains a private model based on the selected labeled records. The intuition of DMP is to restrict the private classifier's direct access to the private training dataset, thus significantly reducing the membership information leakage. In contrast to the requirement of a public unlabeled reference dataset in DMP, Zheng et al. [244] propose **complementary knowledge distillation (CKD)** and **pseudo complementary knowledge distillation (PCKD)** where the transfer data of knowledge distillation all come from the private training set. CKD and PCKD eliminate the need for public data that may be hard to obtain in some applications, making knowledge distillation a more practical defense to mitigate MIAs on ML models.

5.4 Differential Privacy

Differential privacy (DP) [44] is a probabilistic privacy mechanism that provides an information-theoretical privacy guarantee. Many papers [21, 27, 30, 60, 76, 77, 82, 83, 86, 107, 110, 145, 162, 179, 199, 217, 227, 230] have applied DP to ML models to mitigate MIAs. When an ML model is trained in a differentially private manner, the learned model does not learn or remember any specific user's details if the privacy budget is sufficiently small. By definition, differenitially private models naturally limit the success probability of MIAs based solely on the model.

Shokri et al. [181] first discussed that differentially private models should be able to mitigate MIAs on ML models. Yeom et al. [227] theoretically connect DP to MIAs and prove that the membership advantage (refer to Section 6.1, "Metrics") of an attacker is limited by a function of the privacy budget ϵ. Rahman et al. [162] first empirically evaluate MIAs on differentially private DNN-based classifiers. They find that differentially private models provide privacy protection against strong attackers by only offering poor model utility. Jayaraman and Evans [82] further demonstrate that current mechanisms for differentially private ML rarely provide acceptable membership privacy-utility tradeoffs. They comprehensively evaluate MIAs on different variants of the DP mechanisms including differential privacy with advanced composition [43], zero concentrated DP [15], and Réiyi DP [138] for ML models. They find that membership privacy leakage is high when setting DP with limited classifiers' accuracy loss and setting DP to provide strong privacy guarantees, resulting in useless models. Truex et al. [199] evaluate how MIAs differ across classes and how DP affects models when they are trained on skewed data where the class distribution is imbalanced. They report that the minority groups are more vulnerable to MIAs. Moreover, as a mitigation technique, DP tends to decrease a model's utility on the minority groups. Training differentially private ML models is usually achieved by DP-SGD [1], which adds noise to the gradients of the model during training. Rahimian et al. [161] argue that DP-SGD might significantly hinder the model's prediction performance when the attacker is in the black-box setting. They propose DP-Logits that uses a Gaussian mechanism to only add noise to the logits of the input instance at prediction time and restrict the number of queries. They report that the privacy budget for the DP-Logits is generally lower than the DP-SGD method.

DP can also be used to defend against MIAs on generative models. Many papers [8, 27, 148, 198, 217, 220, 222, 240] have proposed various differentially private generative models to ensure the privacy of their training records. Hayes et al. [60] first evaluated how MIAs perform on a differentially private GAN proposed by Triastcyn and Faltings [198]. Hayes et al. [60] find that their proposed white-box MIA achieves great attack performance when ϵ of the differentially private GAN is relatively high. When ϵ is small, MIAs perform no better than random guessing. However, small ϵ also leads GANs to generate bad quality samples. Chen et al. [27] report similar findings that DP indeed reduces the effectiveness of MIAs on GANs even when ϵ exceeds practical values (i.e., $\epsilon > 10^{10}$). However, they also mention that DP heavily deteriorates the generation quality of GANs. Moreover, applying DP into training leads to a much higher computation cost where the training time is 10 times slower compared to training without DP. Wu et al. [217] theoretically prove that the generalization gap of GANs trained with differentially private learning algorithms can be bounded. This indicates DP limits the overfitting of GANs to a certain degree and explains why DP helps to mitigate MIAs.

DP provides a theoretical guarantee to protect the membership privacy of training records. DP can be leveraged to mitigate MIAs on both classification models and generative models, no matter whether an attacker is in a black-box or white-box setting. Although DP is widely applicable and effective, one drawback is that it rarely offers acceptable utility-privacy tradeoffs with guarantees for complex learning tasks. That is, it provides meaningless membership privacy guarantees at settings with limited model utility loss, and it results in useless models at settings with strong privacy guarantees [82]. However, one must be aware that DP cannot only be used to mitigate MIAs, but also other forms of privacy attacks such as attribute inference attacks [47, 48] and property inference attacks [4, 49]. Recent studies have also indicated DP has an interesting connection to model robustness against adversarial examples [104].

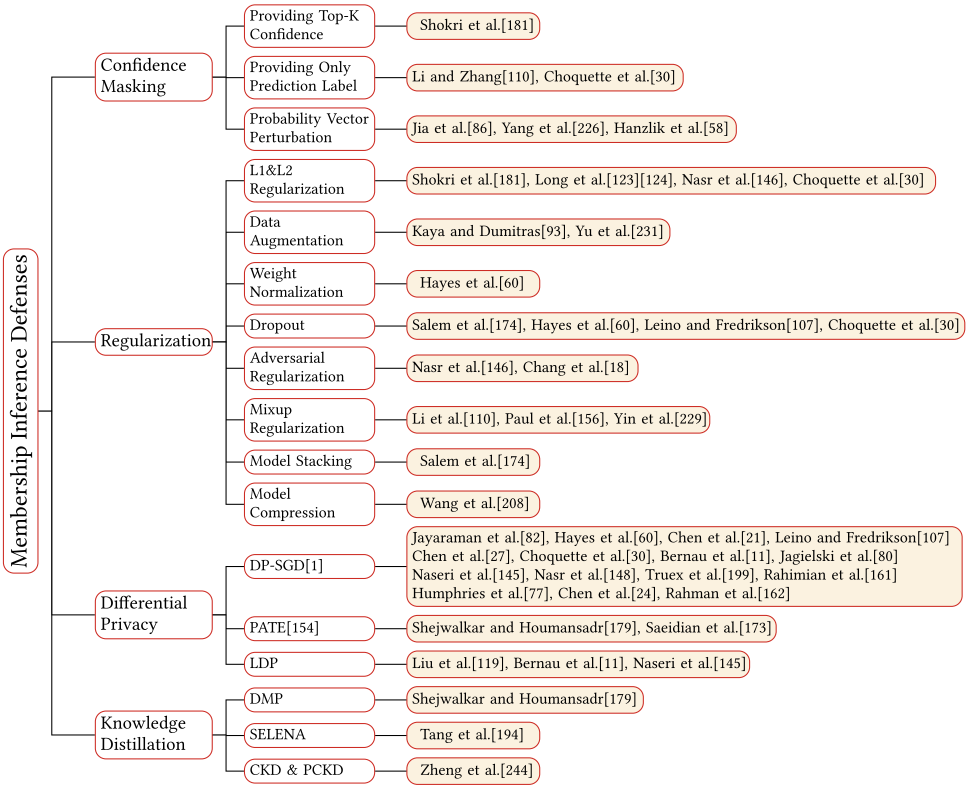

Membership Inference Defenses
Confidence Masking
Providing Top-K Confidence
Shokri et al.[181]
Providing Only Prediction Label
Li and Zhang[110], Choquette et al.[30]
Probability Vector Perturbation
Jia et al.[86], Yang et al.[226], Hanzlik et al.[58]
Regularization
L1&L2 Regularization
Shokri et al.[181], Long et al.[123][124], Nasr et al.[146], Choquette et al.[30]
Data Augmentation
Kaya and Dumitras[93], Yu et al.[231]
Weight Normalization
Hayes et al.[60]
Dropout
Salem et al.[174], Hayes et al.[60], Leino and Fredrikson[107], Choquette et al.[30]
Adversarial Regularization
Nasr et al.[146], Chang et al.[18]
Mixup Regularization
Li et al.[110], Paul et al.[156], Yin et al.[229]
Model Stacking
Salem et al.[174]
Model Compression
Wang et al.[208]
Differential Privacy
DP-SGD[1]
Jayaraman et al.[82], Hayes et al.[60], Chen et al.[21], Leino and Fredrikson[107]
Chen et al.[27], Choquette et al.[30], Bernau et al.[11], Jagielski et al.[80]
Naseri et al.[145], Nasr et al.[148], Truex et al.[199], Rahimian et al.[161]
Humphries et al.[77], Chen et al.[24], Rahman et al.[162]
PATE[154]
Shejwalkar and Houmansadr[179], Saeidian et al.[173]
LDP
Liu et al.[119], Bernau et al.[11], Naseri et al.[145]
Knowledge Distillation
DMP
Shejwalkar and Houmansadr[179]
SELENA
Tang et al.[194]
CKD & PCKD
Zheng et al.[244]

Fig. 9. Taxonomy of membership inference defenses.

5.5 Taxonomies of Membership Inference Defenses

Similar to the taxonomies of attacks, we also give readers a general picture of membership inference defenses to help readers find the most relevant papers easily. The taxonomy of membership inference defenses is illustrated in Figure 9. In this taxonomy, we categorize all released papers of membership inference defenses into four main categories, i.e., confidence masking-based defenses, regularization-based defenses, differential privacy-based defenses, and knowledge distillation-based defenses. For the papers under each of the categories, we further divide the papers based on the specific defense approach, enabling the readers to find the most relevant papers. In addition to Figure 9, we select a few representative papers and list them with their characteristics in a table in the supplemental material due to the limited space of this article.

6 METRICS, DATASETS, AND OPEN-SOURCE IMPLEMENTATIONS

In this section, we first summarize the metrics for evaluating attack and defense performance of membership inference. Then, we summarize common datasets used in membership inference attack and defense works on machine learning models. Last, we provide links to the open-source implementation of popular methods.

6.1 Metrics

In this subsection, we first briefly introduce the general evaluation metrics of target models. Then, we give a detailed introduction of particular evaluation metrics designed for attacks and defenses.

Table 2. Summary of Datasets Used for Evaluating Membership Inference Attacks and Defenses

Type	Task	Dataset	Source	# Records	# Features	# Classes	Paper
Binary Data	Classification	Adult	[40]	48,842	14	2	[26, 30, 45, 76, 77, 107, 122–124, 165, 174, 181, 195, 199, 205, 223, 233]
		Cancer	[40]	699	10	2	[107, 123, 124]
		Diabetes	[40]	768	8	2	[107]
		Hepatitis	[40]	155	19	2	[107]
		German credit	[40]	1,000	20	2	[107, 215]
		Hospital	[189]	101,766	127	2	[180]
		UTKFace	[241]	20,705	N.A.	106	[223]
		US-Accident	[142]	3,000,000	30	3	[26]
		Foursquare	[224]	528,878	446	30	[30, 41, 76, 86, 133, 161, 174, 181, 185, 242, 243]
		Purchase-100	[88]	197,324	600	100	[11, 30, 41, 76, 80, 82, 83, 110, 122, 145–148, 161, 165, 174, 179–181, 185, 195, 199, 200, 205, 242–244]
		Texas-100	[152]	67,330	6,170	100	[11, 23, 30, 76, 83, 86, 110, 146, 147, 161, 165, 179–181, 185, 195, 233]
Image Data	Classification	Colored-MNIST	[3]	70,000	28 × 28 × 1	2	[196]
		CH-MNIST	[92]	5,000	150 × 150 × 1	8	[76, 86, 161]
		SVHN	[149]	99,289	32 × 32 × 3	10	[85, 165, 186]
		Yale Face	[106]	2,414	168 × 192 × 1	38	[187]
		RCV1X	[108]	800,000	N.A.	103	[83]
		Birds-200	[211]	11,788	N.A.	200	[76]
		FaceScrub	[151]	100,000	N.A.	530	[226]
		ImageNet	[36]	1,281,167	N.A.	1,000	[7, 164, 165, 172, 231]
	Classification & Generation	IDC	[81]	277,524	50 × 50 × 3	2	[215, 217]
		EyePACS	[89]	88,702	N.A.	5	[60, 76, 156]
		MNIST	[103]	70,000	28 × 28 × 1	10	[22, 23, 25–27, 30, 66, 72, 78, 107, 110, 112, 118, 123, 124, 143, 148, 161, 162, 164, 165, 174, 181, 198–200, 205, 208, 227, 237]
		Fashion-MNIST	[219]	70,000	28 × 28 × 1	10	[41, 66, 78, 80, 93, 94, 143, 161, 165, 187]
		CIFAR-10	[98]	60,000	32 × 32 × 3	10	[7, 10, 22, 23, 25, 26, 30, 41, 60, 66, 78, 80, 85, 93, 94, 107, 110, 112, 116, 143, 148, 161, 162, 164, 165, 172, 174, 179–181, 186, 199, 200, 205, 208, 227, 229, 231, 239, 244]
		CIFAR-100	[98]	60,000	32 × 32 × 3	100	[7, 22, 30, 58, 72, 76, 82, 93, 94, 107, 110, 112, 145–147, 161, 164, 165, 172, 174, 179–181, 185, 199, 208, 227, 231, 242, 244, 246]
		LFW	[75]	13,233	62 × 47 × 3	5,749	[11, 60, 107, 112, 133, 143, 174, 199, 217]
	Generation	CelebA	[121]	202,599	218 × 178 × 3	10,177	[21, 118, 120, 177, 198]
		MIMIC-III	[87]	46,520	1,071	N.A.	[21]
		Insta-NY	[6]	34,336	4,048	N.A.	[21, 26]
		ChestX-ray8	[206]	108,948	1024 × 1024 × 1	32,717	[118]
	Segmentation	Cityscapes	[31]	20,000	N.A.	30	[65, 177]
		BDD100K	[232]	100,000	N.A.	N.A.	[65]
		Mapillary-Vistas	[150]	25,000	N.A.	37	[65]
Text Data	Classification	CSI	[203]	1,412	N.A.	2	[133]
		Review	[129]	364,038	N.A.	2	[20, 229]
		Tweet EmoInt	[140]	7,097	N.A.	4	[115]
		Yelp-health	[133]	17,938	N.A.	10	[133]
		News	[101]	20,000	N.A.	20	[174]
		Weibo	[235]	23,000	N.A.	N.A.	[115]
	Generation	Reddit comments	[163]	83,293	N.A.	N.A.	[184]
		Dialogs	[34]	220,579	N.A.	N.A.	[184]
		SATED	[135]	2,324	N.A.	N.A.	[184]
		WMT18	[13]	N.A.	N.A.	N.A.	[69]
	Embedding	Wikipedia	[128]	150,000	N.A.	N.A.	[182]
		BookCorpus	[245]	14,000	N.A.	N.A.	[182]
Graph Data	Classification	Pubmed	[175]	19,717	500	3	[42, 153]
		Citeseer	[175]	3,327	3,703	6	[42, 64, 153]
		Cora	[175]	2,708	1,433	7	[42, 64, 153]
		Lastfm	[169]	7,624	7,842	18	[64]

6.1.1 General Metric. According to Figure 8, many existing works tackle the binary or multi-class classification problem. The metric **Accuracy** for classification problems is used by existing works to reflect the classification performance of target models. Readers can refer to Reference [212] for a detailed explanation of **Accuracy**. Another metric **Generalization Error** [59] defined as absolute difference between the Train Accuracy and the Test Accuracy of the target model is used by existing works to reflect the overfitting level of target models. A larger Generalization Error indicates the target model is more overfitted to its training data, demonstrating the target model is associated with higher privacy risks of membership inference attacks [181].

6.1.2 Adversarial Metric. Besides the general metrics used for evaluating target models, a number of metrics that measure the attack and defense performance have been proposed or used by existing works. In this subsection, we introduce the detailed formulations and descriptions of widely used metrics. Each metric name follows the reference of the first paper that proposes or uses this metric, and the references inside the parentheses refer to other attack and defense papers using this metric.

- **Attack Success Rate (ASR) [181].** ([7, 10, 11, 19, 22–25, 30, 41, 42, 54, 60, 64, 66, 69, 72, 78, 85, 86, 94, 105, 115, 116, 134, 145–147, 164, 172, 174, 179, 180, 184–187, 194–196, 199, 200, 205, 208, 209, 215, 223, 226, 229, 231, 237, 239, 244, 246]). ASR is the most commonly used metric to measure the performance of a given attack approach:

$$\text{ASR} = \frac{\#\ \text{Successful attacks}}{\#\ \text{All attacks}}.$$

- **Attack Precision (AP) [181].** ([10, 20, 22, 30, 66, 83, 107, 115, 116, 123, 124, 133, 134, 153, 164, 174, 184–186, 200, 227, 237, 244, 246]). AP is the fraction of records classified as members that are indeed members of the training dataset:

$$\text{AP} = \frac{\#\ \text{Members correctly classified as members}}{\#\ \text{Records classified as members}}.$$

- **Attack Recall (AR) [181].** ([10, 22, 107, 116, 123, 133, 134, 148, 153, 164, 174, 184–186, 227, 244, 246]). AR is the fraction of the training dataset's members that are correctly classified as members:

$$\text{AR} = \frac{\#\ \text{Members correctly classified as members}}{\#\ \text{All members}}.$$

- **Attack False Positive Rate (FPR) [164].** ([148]). Attack FPR is the fraction of the testing dataset's records that are misclassified as members:

$$\text{FPR} = \frac{\#\ \text{Non-members classified as members}}{\#\ \text{All non-members}}.$$

- **Membership Advantage (MA) [227].** ([42, 45, 77, 82, 83, 93, 94, 107, 110, 182, 194, 218]). MA is the difference between the **Attack Recall (AR)** and the attack **False Positive Rate (FPR)**:

$$\text{MA} = \text{AR} - \text{FPR}.$$

- **Attack F_1-score [162].** ([65, 76, 134, 164, 215, 237, 239]). Attack F_1-score is the harmonic mean of Attack Precision and Attack Recall:

$$F_1\text{-score} = \frac{2 \cdot \text{AP} \cdot \text{AR}}{\text{AP} + \text{AR}}.$$

- **Attack AUC [184].** ([20, 21, 58, 65, 118, 153, 161, 165, 177, 215, 217, 218, 242, 243, 246]). **AUC is Area-under-the-ROC-curve**. Readers can refer to Reference [214] for a detailed explanation of AUC. Attack AUC is sensitive to the probability rank of members, which is larger when members are ranked higher than non-members according to the predicted probability of a membership inference binary classifier.

6.2 Evaluation Datasets and Open-source Implementations

Table 2 summarizes all datasets used in membership inference attack and defense works on ML models. We categorize all datasets based on different data types. Among the binary datasets, Adult, Foursquare, Purchase-100, and Texas-100 have been widely used as classification benchmarks in many papers. Among the image datasets, MNIST, Fashion-MNIST, CIFAR-10, CIFAR-100, and LFW are widely used as both classification and generation benchmarks. Because MIAs are relatively unexplored on NLP, different works use different text datasets to evaluate membership privacy risks on different models. Among graph datasets, Citeseer and Cora are widely used node classification benchmarks. Due to the limited space of this article, we provide links to the open-source implementation of popular methods in our GitHub repository. We hope our work can facilitate the community to move towards the construction of benchmarks, similar to other areas [36, 204].

7 FUTURE DIRECTIONS

In this section, we discuss several main challenges and potential research opportunities on membership inference attacks and defenses to inspire interested readers to explore this field more.

7.1 Membership Inference Attacks

(1) The assumption that the target ML models are heavily overfitted to their training data both lacks the practicability of MIAs and limits their applicability, while it underpins the success of most existing works of MIAs. The attacker cannot guarantee that the target models are always heavily overfitted, because many regularization techniques have been used to prevent the overfitting of the ML models. Since MIAs on non-overfitted ML models are pretty challenging, they have not been explored in depth and thus inspire a practically interesting direction for MIAs. The feasibility of MIAs on such non-overfitted models still remains unknown in existing literature. It is even really difficult for an attacker to tell if a target ML model is overfitted or not, as she has quite limited knowledge about the training process.

(2) Recently, self-supervised learning models such as BERT [38], T5 [160], and MoCo [61] are popular and have achieved promising results for many complex downstream tasks such as computer vision and natural language processing, but MIAs on such emerging models have not been explored yet. It is urgent and crucial to investigate the membership privacy risks on self-supervised learning models, because their training datasets consist of large unlabeled data such as image, text, and audio without labelling, which can still be highly private and unauthorized. It can be intriguing to adopt the principles and designs in existing MIAs on supervised and unsupervised learning schemes for the increasingly important self-supervised learning models.

(3) Adversarial machine learning aims to fool or misguide a model with malicious input with typical examples such as data poisoning attacks and model evasion attacks and plays an increasingly important role in applications such as auto-driving safety and spam filtering. While adversarial ML and MIAs as two separate research areas have developed in parallel, it is interesting and challenging to understand their relationships in terms of their theoretical foundations, algorithmic designs, building blocks, and so on. For instance, how to explain the phenomenon that the behavior difference between non-members and members in MIAs is similar with that between adversarial and benign examples in adversarial ML will bridge the two areas to achieve private and secure ML. There have been a couple of initial works [30, 112] exploiting the fact that the training data is more robust against adversarial attacks than the test data to launch label-only MIAs in the black-box context. One specific research question can be how the attacker creates MIAs by leveraging the techniques of

white-box adversarial examples, e.g., **Fast Gradient Sign Method (FGSM)** [53] and **Projected Gradient Descent (PGD)** [126].

(4) There are more avenues where MIAs have not been explored but intensive research efforts are demanded due to their high importance, e.g., contrastive learning models and meta-learning models. Contrastive learning aims to learn similar/dissimilar representations from data that are organized into similar/dissimilar pairs [28]. Meta-learning, also known as learning to learn, refers to the process of improving a learning algorithm over multiple learning episodes [71]. The particular training paradigms of contrastive learning and meta-learning are pretty different from the conventional supervised learning scheme and therefore impose unique challenges on MIAs. For example, data augmentation is a core building block for contrastive learning for generalizable embedding features by enriching positive training examples with perturbation. An interesting question naturally arises: How does the data augmentation in contrastive learning affect MIAs? In meta-learning, the training dataset consists of a certain number of source tasks, and each source task has both training data and validation data. How do MIAs behave on the data of different source tasks? Moreover, for each source task, do MIAs behave differently on the training data and validation data?

(5) As discussed in Section 3.5, federated learning has emerged as a promising privacy-aware paradigm, and some initial works [105, 133, 147, 237] have demonstrated the feasibility of MIAs on federated learning. However, the applicability of existing MIAs is limited to homogeneous federated learning where each local party is assumed to have the same model architecture, while such an assumption is too strong for real applications, because the computation and communication capabilities of each party can vary significantly and dynamically [111]. On the contrary, heterogeneous federated learning schemes such as FedMD [109] and HeteroFL [39] have recently been proposed to handle the system heterogeneity [111] without requiring local models to share the same architecture. While heterogeneous federated learning is more practically realistic, little research has been done in depth for exploring the membership privacy risks in this learning paradigm. Thus, extensive efforts are required to investigate the feasibility and efficacy of MIAs on heterogeneous federated learning schemes to shed light on building more private federated learning systems. Another interesting question is how we can exploit the system heterogeneity information to perform inference attacks on a specific party, since a recent pioneer work [74] has shown the success of source inference attacks in the homogeneous context.

(6) It is of practical interest to investigate new applications by exploiting the information gained from MIAs, which we believe is still in its infancy. The following are a few recently emerging examples we listed to inspire more applications: Based on the membership information from MIAs, the recently proposed source inference attacks in federated learning can further identify the party (i.e., the source) owning a given training member [74]. Because training data are more prone to evasion attacks in the context of adversarial learning, the membership information from MIAs can also be leveraged to improve the design of adversarial examples. Another interesting application of membership inference is to audit if a data record has contributed to the training of an ML model. This is an essential step for data owners to achieve the full control of their data, as described by many recently issued privacy regulations and laws such as GDPR [213], because unauthorized use of their data for training an ML model can be detected with membership inference on the trained model. Further actions like machine unlearning [14] can be taken if the data owners wish to recall the contribution of their data to the trained model.

7.2 Membership Inference Defenses

(1) As revealed in existing literature, the overfitting of ML models is the main factor contributing to the success of MIAs, and the level of overfitting can be leveraged to measure the effectiveness of a membership inference defense method, but it is still a challenge to capture the overfitting phenomenon, especially for unsupervised learning. Most existing effective defenses can mitigate MIAs, because they are designed to reduce the overfitting level of the target ML models. With labelled validation data, we can estimate generalization error [141] to monitor the overfitting level for the supervised learning models and can further determine if a defense method is effective or not. However, it is still a challenge for unsupervised learning models to perceive and handle overfitting due to the lack of data labels, and this considerably limits the potential of designing and evaluating the defenses against MIAs on such models. For instance, no defense has been proposed to mitigate MIAs on word embedding models, to the best of our knowledge. Thus, it is a promising research direction to explore the defenses on unsupervised learning models from the perspective of overfitting, given that the unsupervised learning models such as GANs [52] and VAEs [95] have become increasingly popular and important in many applications where abundant unlabelled data are available.

(2) With the rapid development of generative models such as GANs and VAEs, the generated examples from these models can be highly similar to the original training data, and it is very intriguing to explore the possibility of leveraging the generated examples as surrogate data for model training to mitigate MIAs. The surrogate datasets can help decouple the direct relationship between the original training data and the target model output, while the population features can still be retained to train effective models. Besides, data augmentation techniques can also be used to create surrogate datasets by perturbing original training examples. Extensive efforts are required to explore the theoretical foundation for this category of membership inference defenses, and more defense mechanisms are expected based on a wide range of generative models and data augmentation techniques. It is worth noting that many emerging machine learning schemes can not only be the targets of MIAs, but they can also be exploited to defend against MIAs.

(3) While utility is the ultimate goal in many data analytic applications, it is very challenging to design defense solutions with an acceptable tradeoff between membership privacy and model utility. The defense methods offering strong privacy guarantees often come at the cost of high utility loss of the target model. In particular, the deployed query interface of the target models has been ignored in most existing works, while this factor plays an important role in the utility-privacy tradeoff, as it determines if a white-box or black-box attack can be launched. Taking the classification model as an example, existing differential privacy-based defenses often add a large amount of noise to the gradients of the target classifier in a white-box manner during the training process, and thus can heavily lower its prediction accuracy. However, if the classification model is deployed by only providing the query interface to users in a black-box manner, then can we achieve little prediction performance sacrifice while adding noise only to the output of the target classifier while guaranteeing the differential privacy requirements?

(4) Even though federated learning emerges as a privacy-aware learning paradigm, it faces an increasing number of privacy attacks [125], and it is urgent for the community to develop the corresponding defense techniques. Differential privacy is well known to offer strong privacy guarantees and can be integrated in federated learning, but the model utility can be retained only when the number of parties is very large [50, 132]. It is still a challenge to

achieve acceptable privacy-utility tradeoff when applying differential privacy to business-to-business federated learning [225] where the number of parties is usually small. Based on MIAs, source inference attacks [74] have been specifically designed against federated learning, but no specific defense techniques have been proposed, which offers many appealing opportunities for interested researchers in this field.

8 CONCLUSION

In this work, we have covered most, if not all, the released papers about membership inference attack and defense on ML models. We first give the definition of MIAs on ML models and introduce existing attack approaches. We give a taxonomy to categorize all the papers of MIAs. Next, we discuss why MIAs can work on ML models. Then, we introduce the existing defense approaches used to mitigate MIAs and give a taxonomy to categorize the papers of membership inference defense. We have summarized most existing evaluation metrics, datasets, and open-source implementations of popular approaches. Finally, for both membership inference attack and defense, we discuss the challenges and point out the potential research opportunities for future studies. Through this comprehensive survey, we hope to prepare a solid foundation for future research in this field.

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 *CCS*. ACM, 308–318.

[2] Mohanad Abukmeil, Stefano Ferrari, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. 2021. A survey of unsupervised generative models for exploratory data analysis and representation learning. *ACM Comput. Surv.* 54, 5 (2021), 1–40.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[4] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Netw. Secur.* 10, 3 (2015), 137–150.

[5] Lei Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *NeurIPS*. 2654–2662.

[6] Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. 2017. walk2friends: Inferring social links from mobility profiles. In *CCS*. ACM, 1943–1957.

[7] Aadesh Mahavir Bagmar, Shishira Maiya, Shruti Bidwalkar, and Amol Deshpande. 2021. Membership inference attacks on lottery ticket networks. In *ICML Workshop*. PMLR.

[8] Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. 2019. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* 12, 7 (2019).

[9] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. 2021. Deep learning for AI. *Commun. ACM* 64 (2021), 58–65.

[10] Jason W. Bentley, Daniel Gibney, Gary Hoppenworth, and Sumit Kumar Jha. 2020. Quantifying membership inference vulnerability via generalization gap and other model metrics. *arXiv preprint arXiv:2009.05669* (2020).

[11] Daniel Bernau, Jonas Robl, Philip W. Grassal, Steffen Schneider, and Florian Kerschbaum. 2021. Comparing local and central differential privacy using membership inference attacks. In *DBSec*. Springer, 22–42.

[12] Christopher M. Bishop. 2006. Pattern recognition. *Mach. Learn.* 128, 9 (2006).

[13] Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation *(WMT'18).* In *WMT*. ACL, 272–303.

[14] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *S&P*. IEEE, 141–159.

[15] Mark Bun and Thomas Steinke. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *TCC*. Springer, 635–658.

[16] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*. USENIX Association, 267–284.

[17] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805* (2020).

[18] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. 2019. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279* (2019).

[19] Hongyan Chang and Reza Shokri. 2021. On the privacy risks of algorithmic fairness. In *EuroS&P*. IEEE.

[20] Cen Chen, Bingzhe Wu, Minghui Qiu, Li Wang, and Jun Zhou. 2020. A comprehensive analysis of information leakage in deep transfer learning. *arXiv preprint arXiv:2009.01989* (2020).

[21] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. GAN-leaks: A taxonomy of membership inference attacks against generative models. In *CCS*. ACM, 343–362.

[22] Hanxiao Chen, Hongwei Li, Guishan Dong, Meng Hao, Guowen Xu, Xiaoming Huang, and Zhe Liu. 2022. Practical membership inference attack against collaborative inference in industrial IoT. *IEEE Trans. Industr. Inform.* 18, 1 (2022), 477–487.

[23] 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 *KDD*. ACM, 127–137.

[24] Junjie Chen, Wendy Hui Wang, and Xinghua Shi. 2020. Differential privacy protection against membership inference attack on machine learning for genomic data. In *Biocomputing*. World Scientific, 26–37.

[25] Jiale Chen, Jiale Zhang, Yanchao Zhao, Hao Han, Kun Zhu, and Bing Chen. 2020. Beyond model-level membership privacy leakage: An adversarial approach in federated learning. In *ICCCN*. IEEE, 1–9.

[26] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. 2020. When machine unlearning jeopardizes privacy. *arXiv preprint arXiv:2005.02205* (2020).

[27] 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).

[28] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).

[29] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *MLHC*. PMLR, 286–305.

[30] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *ICML*. PMLR, 1964–1974.

[31] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*. IEEE.

[32] Elliot J. Crowley, Gavin Gray, and Amos Storkey. 2018. Moonshine: Distilling with cheap convolutions. In *NeurIPS*. Curran Associates Inc., 2893–2903.

[33] Amine Dadoun, Raphaël Troncy, Olivier Ratier, and Riccardo Petitti. 2019. Location embeddings for next trip recommendation. In *WWW*. ACM, 896–903.

[34] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *CMCL*. ACL, 76–87.

[35] Emiliano De Cristofaro. 2020. An overview of privacy in machine learning. *arXiv preprint arXiv:2005.08679* (2020).

[36] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.

[37] Delan Devakumar, Geordan Shannon, Sunil S. Bhopal, and Ibrahim Abubakar. 2020. Racism and discrimination in COVID-19 responses. *Lancet* 395 (2020), 1194.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[39] Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *ICLR*. Retrieved from OpenReview.net.

[40] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml.

[41] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. 2020. GECKO: Reconciling privacy, accuracy and efficiency in embedded deep learning. In *NuerIPS Workshop*.

[42] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. 2020. Quantifying privacy leakage in graph embedding. In *EAI MobiQuitous*. 1–11.

[43] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *TAMC*. Springer, 1–19.

[44] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *TCC*. Springer, 265–284.

[45] Farhad Farokhi and Mohamed Ali Kaafar. 2020. Modelling and quantifying membership information leakage in machine learning. *arXiv preprint arXiv:2001.10648* (2020).

[46] Robert Fortet and Edith Mourier. 1953. Convergence de la répartition empirique vers la répartition théorique. In *Annales scientifiques de l'École Normale Supérieure*, Vol. 70. 267–285.

[47] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *CCS*. ACM, 1322–1333.

[48] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security*. USENIX Association.

[49] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. Property inference attacks on fully connected neural networks using permutation invariant representations. In *CCS*. ACM, 619–633.

[50] Robin C. Geyer, Tassilo Klein, and Moin Nabi. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557* (2017).

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. The MIT Press.

[52] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

[53] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[54] Kathrin Grosse, Michael T. Smith, and Michael Backes. 2021. Killing four birds with one Gaussian process: The relation between different test-time attacks. In *ICPR*. IEEE, 4696–4703.

[55] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*. ACM, 855–864.

[56] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of Wasserstein GANs. In *NeurIPS*. 5769–5779.

[57] Umang Gupta, Dimitris Stripelis, Pradeep K. Lam, Paul Thompson, Jose Luis Ambite, and Greg Ver Steeg. 2021. Membership inference attacks on deep regression models for neuroimaging. In *MIDL*, Vol. 143. PMLR, 228–251.

[58] Lucjan Hanzlik, Yang Zhang, Kathrin Grosse, Ahmed Salem, Maximilian Augustin, Michael Backes, and Mario Fritz. 2021. MLcapsule: Guarded offline deployment of machine learning as a service. In *CVPR*. IEEE, 3300–3309.

[59] Moritz Hardt, Ben Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*. PMLR, 1225–1234.

[60] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. Logan: Membership inference attacks against generative models. *Proc. Priv. Enhanc. Technol.* 2019, 1 (2019), 133–152.

[61] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*. IEEE, 9729–9738.

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. IEEE, 770–778.

[63] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. ACM, 173–182.

[64] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. 2021. Node-level membership inference attacks against graph neural networks. *arXiv preprint arXiv:2102.05429* (2021).

[65] Yang He, Shadi Rahimian, Bernt Schiele, and Mario Fritz. 2020. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. In *ECCV*. Springer, 519–535.

[66] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte Carlo and reconstruction membership inference attacks against generative models. *Proc. Priv. Enhanc. Technol.* 2019, 4 (2019), 232–249.

[67] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. Retrieved from https://arxiv.org/abs/1503.02531.

[68] Geoffrey E. Hinton, Terrence Joseph Sejnowski, et al. 1999. *Unsupervised Learning: Foundations of Neural Computation*. The MIT Press.

[69] Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Trans. Assoc. Comput.* 8 (2020), 49–63.

[70] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 4 (2008), 1–9.

[71] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439* (2020).

[72] Jiahui Hou, Jianwei Qian, Yu Wang, Xiang-Yang Li, Haohua Du, and Linlin Chen. 2019. ML defense: Against prediction API threats in cloud-based machine learning service. In *IWQoS*. ACM, 1–10.

[73] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, Yi Chen, and Xuyun Zhang. 2021. EAR: An enhanced adversarial regularization approach against membership inference attacks. In *IJCNN*. IEEE, 1–8.

[74] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. 2021. Source inference attacks in federated learning. *arXiv preprint arXiv:2109.05659* (2021).

[75] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in "Real-Life" Images: Detection, Alignment, and Recognition.*

[76] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. 2021. Practical blind membership inference attack via differential comparisons. In *NDSS*. Internet Society.

[77] 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).

[78] Paul Irolla and Grégory Châtel. 2019. Demystifying the membership inference attack. In *CMI*. IEEE, 1–7.

[79] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305* (2021).

[80] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing differentially private machine learning: How private is private SGD? In *NeurIPS*. 22205–22216.

[81] Andrew Janowczyk and Anant Madabhushi. 2016. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inf.* 7 (2016), 29–29.

[82] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *USENIX Security*. USENIX Association, 1895–1912.

[83] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. 2020. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881* (2020).

[84] Malhar S. Jere, Tyler Farnan, and Farinaz Koushanfar. 2020. A taxonomy of attacks on federated learning. *IEEE Secur. Priv.* 19, 2 (2020), 20–28.

[85] Sumit Kumar Jha, Susmit Jha, Rickard Ewetz, Sunny Raj, Alvaro Velasquez, Laura L. Pullum, and Ananthram Swami. 2020. An extension of Fano's inequality for characterizing model susceptibility to membership inference attacks. *arXiv preprint arXiv:2009.08097* (2020).

[86] 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 *CCS*. ACM, 259–274.

[87] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, H. Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data* 3 (2016), 1–9.

[88] Kaggle. 2014. Acquire Valued Shoppers Challenge. Retrieved from https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data.

[89] Kaggle. 2015. Diabetic Retinopathy Detection. Retrieved from https://www.kaggle.com/c/diabetic-retinopathy-detection#references.

[90] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *KDD*. ACM.

[91] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[92] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M. Melchers, Lothar R. Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. 2016. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.* 6 (2016), 1–11.

[93] Yigitcan Kaya and Tudor Dumitras. 2021. When does data augmentation help with membership inference attacks? In *ICML*. PMLR, 5345–5355.

[94] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2020. On the effectiveness of regularization against membership inference attacks. *arXiv preprint arXiv:2006.05336* (2020).

[95] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).

[96] Thomas N. Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[97] Dietrich Klakow. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *TSD*. Springer, 273–281.

[98] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. Citeseer 1.

[99] Solomon Kullback. 1997. *Information Theory and Statistics*. Courier Corporation.

[100] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).

[101] Ken Lang. 1995. NewsWeeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*. Elsevier, 331–339.

[102] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *ICML*. PMLR, 1558–1566.

[103] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86 (1998), 2278–2324.

[104] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *S&P*. IEEE, 656–672.

[105] 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.

[106] Kuang-Chih Lee, Jeffrey Ho, and David J. Kriegman. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (2005), 684–698.

[107] Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *USENIX Security*. USENIX Association, 1605–1622.

[108] David D. Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5 (2004), 361–397.

[109] Daliang Li and Junpu Wang. 2019. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581* (2019).

[110] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership inference attacks and defenses in classification models. In *CODASPY*. ACM, 5–16.

[111] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Sig. Process. Mag.* 37, 3 (2020), 50–60.

[112] Zheng Li and Yang Zhang. 2021. Membership leakage in label-only exposures. In *CCS*. ACM.

[113] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Med. Image Anal.* 42 (2017), 60–88.

[114] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv.* 54, 2 (2021), 1–36.

[115] Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng. 2019. SocInf: Membership inference attacks on social media health data with machine learning. *IEEE Trans. Computat. Soc. Syst.* 6, 5 (2019), 907–921.

[116] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. 2021. EncoderMI: Membership inference against pre-trained encoders in contrastive learning. In *CCS*. ACM.

[117] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Anil K. Jain, and Jiliang Tang. 2021. Trustworthy AI: A computational perspective. *arXiv preprint arXiv:2107.06641* (2021).

[118] Kin Sum Liu, Chaowei Xiao, Bo Li, and Jie Gao. 2019. Performing co-membership attacks against deep generative models. In *ICDM*. IEEE, 459–467.

[119] 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 (2020), 24–31.

[120] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. 2021. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. *arXiv preprint arXiv:2102.02551* (2021).

[121] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. IEEE, 3730–3738.

[122] Yunhui Long, Vincent Bindschaedler, and Carl A. Gunter. 2017. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136* (2017).

[123] 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).

[124] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. 2020. A pragmatic approach to membership inferences on machine learning models. In *EuroS&P*. IEEE, 521–534.

[125] Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133* (2020).

[126] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[127] Saeed Mahloujifar, Huseyin A. Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa. 2021. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384* (2021).

[128] Matt Mahoney. 2011. Large text compression benchmark. Retrieved from https://cs.fit.edu/~mmahoney/compression/text.html. (2011).

[129] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *RecSys*. ACM, 165–172.

[130] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*. PMLR, 1273–1282.

[131] Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. In *ICLR*. Retrieved from OpenReview.net.

[132] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963* (2017).

[133] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *S&P*. IEEE, 691–706.

[134] Yuantian Miao, Xue Minhui, Chao Chen, Lei Pan, Jun Zhang, Benjamin Zi Hao Zhao, Dali Kaafar, and Yang Xiang. 2021. The audio auditor: User-level membership inference in Internet of Things voice services. *Proc. Priv. Enhanc. Technol.* 2021, 1 (2021), 209–228.

[135] Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817* (2018).

[136] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. 2018. Deep learning for healthcare: Review, opportunities and challenges. *Brief. Bioinf.* 19, 6 (2018), 1236–1246.

[137] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. 2020. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254* (2020).

[138] Ilya Mironov. 2017. Rényi differential privacy. In *CSF*. IEEE, 263–275.

[139] Tom M. Mitchell, et al. 1997. Machine learning. *Burr Ridge, IL: McGraw Hill* 45, 37 (1997), 870–877.

[140] Saif M. Mohmmad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *WASSA*. ACL.

[141] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of Machine Learning*. The MIT Press.

[142] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. 2019. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *SIGSPATIAL*. ACM.

[143] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, Nabajyoti Patowary, and Juan L. Ferres. 2021. privGAN: Protecting GANs from membership inference attacks at low cost to utility. *Proc. Priv. Enhanc. Technol.* 2021, 3 (2021), 142–163.

[144] Sasi Kumar Murakonda and Reza Shokri. 2020. ML privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339* (2020).

[145] Mohammad Naseri, Jamie Hayes, and Emiliano De Cristofaro. 2020. Toward robustness and privacy in federated learning: Experimenting with local and central differential privacy. *arXiv preprint arXiv:2009.03561* (2020).

[146] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *CCS*. ACM, 634–646.

[147] 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 *S&P*. IEEE, 739–753.

[148] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. 2021. Adversary instantiation: Lower bounds for differentially private machine learning. In *S&P*. IEEE.

[149] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. 2011. Reading digits in natural images with unsupervised feature learning. In *NuerIPS Workshop*.

[150] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The Mapillary Vistas dataset for semantic understanding of street scenes. In *ICCV*. IEEE, 4990–4999.

[151] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *ICIP*. IEEE, 343–347.

[152] Texas Department of State Health Services. 2006. Texas Hospital Inpatient Discharge Public Use Data File. Retrieved from https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm.

[153] Iyiola E. Olatunji, Wolfgang Nejdl, and Megha Khosla. 2021. Membership inference attack on graph neural networks. *arXiv preprint arXiv:2101.06570* (2021).

[154] Nicolas Papernot, Martín Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. 2017. Semi-supervised knowledge transfer for deep learning from private training data. In *ICLR*. Retrieved from OpenReview.net.

[155] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814* (2016).

[156] William Paul, Yinzhi Cao, Miaomiao Zhang, and Phil Burlina. 2021. Defending medical image diagnostics against privacy attacks using generative methods. In *MICCAI*. Springer.

[157] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2018. Knock knock, who's there? Membership inference on aggregate location data. In *NDSS*. The Internet Society.

[158] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2020. Measuring membership privacy on aggregate location time-series. *Proc. ACM Meas. Anal. Comput. Syst.* 4 (2020), 1–28.

[159] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[160] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21 (2020), 1–67.

[161] Shadi Rahimian, Tribhuvanesh Orekondy, and Mario Fritz. 2020. Sampling attacks: Amplification of membership inference attacks by repeated queries. In *NuerIPS Workshop*.

[162] 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.

[163] Reddit. 2017. Reddit comments dataset. Retrieved from https://bigquery.cloud.google.com/dataset/fh-bigquery:redditcomments.

[164] Shahbaz Rezaei and Xin Liu. 2021. On the difficulty of membership inference attacks. In *CVPR*. IEEE, 7892–7900.

[165] Shahbaz Rezaei, Zubair Shafiq, and Xin Liu. 2021. Accuracy-privacy trade-off in deep ensemble. *arXiv preprint arXiv:2105.05381* (2021).

[166] Maria Rigaki and Sebastian Garcia. 2020. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646* (2020).

[167] Christian Robert and George Casella. 2013. *Monte Carlo Statistical Methods*. Springer Science & Business Media.

[168] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. 2021. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Comput. Surv.* 54, 5 (2021), 1–36.

[169] Benedek Rozemberczki and Rik Sarkar. 2020. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *CIKM*. ACM, 1325–1334.

[170] Stuart Russell and Peter Norvig. 2002. Artificial intelligence: A modern approach. Pearson Education, Inc.

[171] David Saad. 1998. Online algorithms and stochastic approximations. *Online Learn.* 5 (1998), 6–3.

[172] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs. black-box: Bayes optimal strategies for membership inference. In *ICML*. PMLR, 5558–5567.

[173] Sara Saeidian, Giulia Cervia, Tobias J. Oechtering, and Mikael Skoglund. 2021. Quantifying membership privacy via information leakage. *IEEE Trans. Inf. Forens. Secur.* 16 (2021), 3096–3108.

[174] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *NDSS*. Internet Society.

[175] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI Mag.* 29 (2008), 93–93.

[176] Alex Serban, Erik Poll, and Joost Visser. 2020. Adversarial examples on object recognition: A comprehensive survey. *ACM Comput. Surv.* 53, 3 (2020), 1–38.

[177] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. 2021. Reconstruction-based membership inference attacks are easier on difficult problems. In *ICCV*. IEEE.

[178] Muhammad A. Shah, Joseph Szurley, Markus Mueller, Athanasios Mouchtaris, and Jasha Droppo. 2021. Evaluating the vulnerability of end-to-end automatic speech recognition models to membership inference attacks. *Proc. Interspeech 2021* (2021), 891–895.

[179] Virat Shejwalkar and Amir Houmansadr. 2021. Membership privacy for machine learning models through knowledge transfer. In *AAAI*. AAAI Press, 9549–9557.

[180] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the privacy risks of model explanations. In *AIES*. ACM.

[181] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *S&P*. IEEE, 3–18.

[182] Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *CCS*. ACM, 377–390.

[183] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *CCS*. ACM, 587–601.

[184] Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *KDD*. ACM.

[185] Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security*. USENIX Association, 2615–2632.

[186] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Membership inference attacks against adversarially robust deep learning models. In *S&P Workshops*. IEEE, 50–56.

[187] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy risks of securing machine learning models against adversarial examples. In *CCS*. ACM, 241–257.

[188] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (2014), 1929–1958.

[189] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. 2014. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Res. Int.* 2014 (2014).

[190] Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S. Yu, Lifang He, and Bo Li. 2018. Adversarial attack and defense on graph data: A survey. *arXiv preprint arXiv:1812.10528* (2018).

[191] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*. IEEE, 2818–2826.

[192] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[193] Elham Tabassi, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and Julian Sexton. 2019. A taxonomy and terminology of adversarial machine learning. *J. Res. Natl. Inst. Stand. Technol* (2019), 1–29.

[194] 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).

[195] 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).

[196] Shruti Tople, Amit Sharma, and Aditya Nori. 2020. Alleviating privacy attacks via causal learning. In *ICML*. PMLR.

[197] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *USENIX Security*. USENIX Association, 601–618.

[198] Aleksei Triastcyn and Boi Faltings. 2019. Generating artificial data for private deep learning. In *AAAI-SSS*. CEUR Workshop Proceedings, 33–40.

[199] 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 *TPS-ISA*. IEEE, 82–91.

[200] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying membership inference attacks in machine learning as a service. *IEEE Trans. Serv. Comput.* 01 (2019), 1–1.

[201] Vladimir Vapnik. 1992. Principles of risk minimization for learning theory. In *NeurIPS*. 831–838.

[202] Michael Veale, Reuben Binns, and Lilian Edwards. 2018. Algorithms that remember: Model inversion attacks and data protection law. *Philos. Trans. Roy. Soc. A* 376, 2133 (2018), 20180083.

[203] Ben Verhoeven and Walter Daelemans. 2014. CLiPS stylometry investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC*. 3081–3085.

[204] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[205] Kehao Wang, Zhixin Hu, Qingsong Ai, Quan Liu, Mozi Chen, Kezhong Liu, and Yirui Cong. 2021. Membership inference attack with multi-grade service models in edge intelligence. *IEEE Netw.* 35, 1 (2021), 184–189.

[206] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*. IEEE, 2097–2106.

[207] Yu Wang and Lichao Sun. 2021. Membership inference attacks on knowledge graphs. *arXiv preprint arXiv:2104.08273* (2021).

[208] Yijue Wang, Chenghong Wang, Zigeng Wang, Shanglin Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. 2021. Against membership inference attack: Pruning is all you need. In *IJCAI*. Retrieved from ijcai.org.

[209] Ryan Webster, Julien Rabin, Loïc Simon, and Frédéric Jurie. 2021. Generating private data surrogates for vision related tasks. In *ICPR*. IEEE, 263–269.

[210] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. 2021. This person (probably) exists. Identity membership attacks against GAN generated faces. *arXiv preprint arXiv:2107.06018* (2021).

[211] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. California Institute of Technology. CNS-TR-2010-001. 2010.

[212] Wikipedia. 2021. Confusion Matrix. Retrieved from https://bit.ly/2wHUpcf.

[213] Wikipedia. 2021. General Data Protection Regulation. Retrieved from https://en.wikipedia.org/wiki/General_Data_Protection_Regulation.

[214] Wikipedia. 2021. ROC. Retrieved from https://bit.ly/341yHfa.

[215] 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 *AAAI*. AAAI Press.

[216] Bang Wu, Xiangwen Yang, Shirui Pan, and Xingliang Yuan. 2021. Adapting membership inference attacks to GNN for graph classification: Approaches and implications. *arXiv preprint arXiv:2110.08760* (2021).

[217] 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. In *NeurIPS*. 307–317.

[218] Dominik Wunderlich, Daniel Bernau, Francesco Aldà, Javier Parra-Arnau, and Thorsten Strufe. 2021. On the privacy-utility trade-off in differentially private hierarchical text classification. *arXiv preprint arXiv:2103.02895* (2021).

[219] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[220] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* (2018).

[221] Hui Y. Xiong, Babak Alipanahi, Leo J. Lee, Hannes Bretschneider, Daniele Merico, Ryan K. C. Yuen, Yimin Hua, Serge Gueroussov, Hamed S. Najafabadi, Timothy R. Hughes, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347, 6218 (2015).

[222] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. 2019. GANobfuscator: Mitigating information leakage under GAN via differential privacy. *IEEE Trans. Inf. Forens. Secur.* 14 (2019), 2358–2371.

[223] Mohammad Yaghini, Bogdan Kulynych, Giovanni Cherubin, and Carmela Troncoso. 2019. Disparate vulnerability: On the unfairness of privacy attacks against machine learning. *arXiv preprint arXiv:1906.00389* (2019).

[224] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Trans. Intell. Syst. Technol.* 7 (2016), 1–23.

[225] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. 2019. Federated learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 13 (2019), 1–207.

[226] 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).

[227] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*. IEEE, 268–282.

[228] Xuefei Yin, Yanming Zhu, and Jiankun Hu. 2021. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Comput. Surv.* 54, 6 (2021), 1–36.

[229] Yu Yin, Ke Chen, Lidan Shou, and Gang Chen. 2021. Defending privacy against more knowledgeable membership inference attackers. In *KDD*. ACM, 2026–2036.

[230] Zuobin Ying, Yun Zhang, and Ximeng Liu. 2020. Privacy-preserving in defending against membership inference attacks. In *PPMLP*. ACM, 61–63.

[231] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. 2021. How does data augmentation affect privacy in machine learning? In *AAAI*, Vol. 35. AAAI Press, 10746–10753.

[232] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. 2018. BDD100K: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687* (2018).

[233] 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).

[234] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* 64 (2021), 107–115.

[235] Huaping Zhang. 2017. Weibo content corpus. In Proceedings of the http://www.nlpir.org/wordpress/download/weibo_content_corpus.rar.

[236] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *ICLR*. Retrieved from OpenReview.net.

[237] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. 2020. GAN enhanced membership inference: A passive local attack in federated learning. In *ICC*. IEEE, 1–6.

[238] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhumin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership inference attacks against recommender systems. *arXiv preprint arXiv:2109.08045* (2021).

[239] Tianwei Zhang, Zecheng He, and Ruby B. Lee. 2018. Privacy-preserving machine learning through data obfuscation. *arXiv preprint arXiv:1807.01860* (2018).

[240] Xinyang Zhang, Shouling Ji, and Ting Wang. 2018. Differentially private releasing via deep generative model. *arXiv preprint arXiv:1801.01594* (2018).

[241] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *CVPR*. IEEE, 5810–5818.

[242] Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson. 2021. On the (In) feasibility of attribute inference attacks on machine learning models. In *EuroS&P*. IEEE.

[243] Benjamin Zi Hao Zhao, Hassan Jameel Asghar, Raghav Bhaskar, and Mohamed Ali Kaafar. 2019. On inferring training data attributes in machine learning models. In *CCS Workshop*. ACM.

[244] Junxiang Zheng, Yongzhi Cao, and Hanpin Wang. 2021. Resisting membership inference attacks through knowledge distillation. *Neurocomputing* 452 (2021), 114–126.

[245] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*. IEEE, 19–27.

[246] Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. 2020. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. *arXiv preprint arXiv:2009.04872* (2020).

Received April 2021; revised December 2021; accepted December 2022