A Critical Review on the Effectiveness and Privacy Threats of Membership Inference Attacks

Najeeb Jebreel, David Sánchez, and Josep Domingo-Ferrer

Universitat Rovira i Virgili,
Department of Computer Engineering and Mathematics,
CYBERCAT-Center for Cybersecurity Research of Catalonia,
ComSCIAM-Center for Computational Science and Applied Mathematics
Av. Països Catalans 26, 43007 Tarragona, Catalonia
{najeeb.jebreel, david.sanchez, josep.domingo}@urv.cat

Abstract. Membership inference attacks (MIAs) aim to determine whether a data sample was included in a machine learning (ML) model's training set and have become the *de facto* standard for measuring privacy leakages in ML. We propose an evaluation framework that defines the conditions under which MIAs constitute a genuine privacy threat, and review representative MIAs against it. We find that, under the realistic conditions defined in our framework, MIAs represent weak privacy threats. Thus, relying on them as a privacy metric in ML can lead to an overestimation of risk and to unnecessary sacrifices in model utility as a consequence of employing too strong defenses.

Keywords: Machine learning · Data privacy · Membership inference attacks.

1 Introduction

Advancements in machine learning (ML) have significantly improved performance in many tasks [19,52]. However, these advances have also raised privacy concerns among individuals and regulatory authorities due to the use of potentially sensitive data to train ML models.

The most widespread concern is the vulnerability of ML models to membership inference attacks (MIAs) [65,77,63]. MIAs aim to determine whether a specific data point was included in a model's training data set by exploiting differences in model behavior (such as loss or confidence scores) when presented with member and non-member data points. Indeed, in some scenarios, successfully inferring an individual's membership in a training data set can compromise privacy by revealing sensitive information. For example, if an ML model is trained on medical records of patients with a common sensitive condition, such as HIV or cancer, confirming membership would expose the individual's sensitive health status. Recognizing this risk, organizations such as NIST (USA) and ICO (UK) [70,53] as well as researchers in the health domain [56] consider MIAs a possible violation of training data confidentiality.

Driven by these concerns, and to comply with privacy protection laws such as the European Union's General Data Protection Regulation (GDPR) [25], several defenses against MIAs have been proposed. Differential privacy (DP) [24] is the most widespread and natural defense, as it obscures the influence of any individual data point in the training data on the trained ML model by adding random noise. However, DP reduces the utility of the model due to the noise required to achieve meaningful privacy protection [6,75,10].

On the other hand, early work on MIAs by [77] established a connection between the vulnerability of a model to MIAs and the degree to which the model overfits the training data. Indeed, it seems reasonable that a model that is overfitted on its training data will reveal more about them than a model that has learned more general patterns. However, overfitting is undesirable in production ML because it reduces the ability of models to generalize. Moreover, as has been acknowledged in the privacy literature for decades [34], for a disclosure inference —such as membership— to involve a real privacy threat, that inference should be unequivocal, which is something very rare in MIAs.

A critical question arises: *do membership inference attacks pose a significant threat to privacy under realistic conditions, thereby justifying the adoption of defenses such as DP despite their utility downsides?*

Related work. Several studies have critically examined the effectiveness of MIAs. The literature has established a strong connection between model overfitting and vulnerability to MIAs [65,77,63]. Although MIAs can be effective against overfitted models, their performance deteriorates when applied to well-generalized models [20]. However, even with non-overfitted models, some samples may remain vulnerable to MIAs [49,12,79]. It is shown in [71] that MIAs work well under specific conditions, such as when the data set is complex, the model is sensitive to individual instances, and the attacker can generate accurate approximations of the target model. In [23,29] it is demonstrated that MIAs tend to barely outperform random guessing for large-scale models trained on big and comprehensive data. In [38], it is shown that the effectiveness of the strong LiRA MIA [12] collapses under a realistic evaluation that combines anti-overfitting training, shadow-based threshold calibration, and skewed membership priors, with poor reproducibility of inferred vulnerable samples across runs. In [5] it is theoretically shown that the effectiveness of MIAs is inherently constrained by the statistical properties of the training data. It is argued in [59,60] that MIAs are unreliable because they exhibit high false positives due to the frequent misclassification of semantically similar non-members as members. Adjusting the precision to account for realistic membership priors is proposed in [37]. Recently, a consensus has emerged to use the true positive rate (TPR) at an extremely low false positive rate (FPR) as a benchmark to ensure MIA reliability [12,79,29].

Our contributions. Although the related work mentioned above provides valuable insight into the effectiveness of MIAs, there is a gap to understand the conditions under which MIAs constitute a meaningful privacy threat and how

realistic these conditions are. In this work, we address this gap with the following contributions:

- We propose an evaluation framework that defines the necessary conditions for MIAs to be considered genuine privacy threats in predictive ML —the scenario for which MIAs were designed and on which the surveyed works focus. This framework assesses MIAs along five dimensions: disclosure potential of the target model, applicability to non-overfitted models, applicability to models that are competitive for real-world deployment, attack reliability, and computational feasibility.
- We review representative MIAs through the lens of this framework, evaluating their effectiveness and implications for real-world privacy.
- We discuss the role of MIAs in the assessment of privacy risk and whether their impact justifies the adoption of utility-hampering defenses such as DP.

Our findings reveal that, despite their apparent success in controlled settings, existing MIAs on realistic training data sets and ML models do not meet the conditions required for meaningful privacy threats in real-world scenarios. Thus, relying on MIAs as the primary metric for privacy risk in predictive ML may lead to overestimated threats and unnecessary sacrifices in model utility.

The remainder of this paper is organized as follows. Section 2 examines MIAs from the point of view of disclosure risk. Section 3 presents our framework for evaluating when MIAs pose a genuine privacy threat. Section 4 reviews representative MIAs against our evaluation framework. Section 5 discusses the wider implications of our findings for privacy and ML utility. Finally, Section 6 provides concluding remarks.

2 MIAs and Disclosure Risk

The concept of *disclosure risk* has long served as a foundation for understanding how sensitive information can be accidentally exposed in data sets released. Originally developed in the context of database privacy [34], its principles remain relevant in the machine learning domain.

There are two types of disclosure [34]: i) *identity disclosure (a.k.a. re-identification)*, which allows associating a released non-identified record with the subject to whom it corresponds; and ii) *attribute disclosure*, which enables determining the value of a subject's confidential attribute, such as income or diagnosis.

Attributes that enable direct re-identification are *personal identifiers* (such as passport numbers), whereas *quasi-identifiers* (*e.g.*, zip code, gender or age) are those that do not uniquely identify the subject, but whose combination may because they may be present in public identified databases like electoral rolls. Finally, *confidential* attributes are unknown information on subjects that might reveal their sensitive data (*e.g.*, salaries, diagnoses) when unequivocally associated with them.

Unlike re-identification attacks, MIAs are not designed to identify the subject's record directly. Instead, they focus on determining whether a given data

sample was included in an ML model training set. However, disclosure of training membership can indirectly disclose sensitive information in two ways:

1. *Revealing confidential attributes:* If a data sample explicitly contains confidential features or labels, confirming its presence in the training set may reveal private information about the individual. For example, if a data set contains medical records, confirming that a person's data is included may reveal a sensitive diagnosis.
2. *Exposing confidential contexts:* Membership in a training data set associated with a sensitive context (for example, mental health studies, addiction treatment programs) can itself constitute a violation of privacy, even if there are no explicit sensitive attributes present. In fact, this scenario can be viewed as a training data set with a single homogeneous confidential value that is shared by all members.

To pose a privacy threat, MIAs must accurately and unequivocally disclose *new* information beyond what is already known from population-level statistics; or, in other words, for an MIA to be effective against privacy, it should significantly shift the attacker's prior belief about the attributes of a data point or its association with a confidential context to a different posterior belief.

However, there is a clash between membership disclosure and attribute disclosure, for the following reasons:

- A necessary condition for unequivocal attribute disclosure is that training data must be an exhaustive representation of a population. Otherwise, the attacker cannot be sure that the targeted subject was truly a member of the training data, as their known information could be shared by several other individuals in the population (present or not in the training data). That is, in non-exhaustive population samples, there is *plausible deniability* of any disclosure inference.
- Conversely, if training data are an exhaustive representation of the population (for example, a country-level census), then membership disclosure is trivial: everyone is known to be a member.

In fact, *exhaustivity* is not the only necessary condition for unequivocal attribute inference through MIAs. *Uniqueness of confidential attribute values* is also needed, which means that there should *not* be two or more records in the data set that: i) match the target subject's attributes known to the attacker; ii) have different values for the confidential attribute values the attacker wishes to infer. Without uniqueness, no unequivocal inference of those attributes is possible to the attacker. Finally, another necessary condition is that the information assumed to be known to the attacker about the target should be plausible. Assuming knowledge of too specific or too much information on the subject makes disclosure attacks contrived and unrealistic.

In summary, for an MIA to be considered a genuine privacy threat, it must not only infer membership with very high or, ideally, perfect precision, but the underlying data must also fulfill strict (and often improbable) conditions to

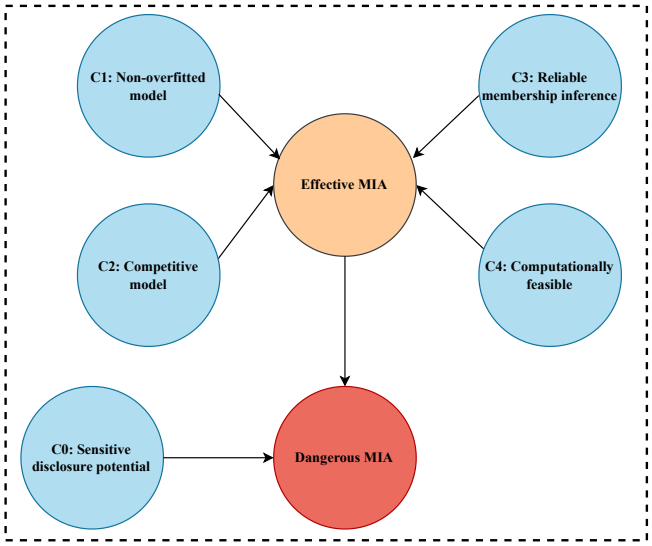


Fig. 1: Overview of the proposed evaluation framework for MIAs. Condition C0 relates to the MIA disclosure potential, which critically depends on the data set used to train the target model. Conditions C1–C4 characterize the effectiveness of the MIA itself, which should reliably attack non-overfitted and competitive models at a reasonable computational cost.

allow unequivocal disclosure of sensitive information. In particular, *unequivocal attribute disclosure occurs only when membership disclosure is trivial because the data set is exhaustive.*

3 Proposed Evaluation Framework

Building on the discussion so far, we propose an evaluation framework that defines five necessary conditions (C0–C4) that must hold simultaneously for an MIA to be a real threat to privacy in predictive ML. C0 is a data-level precondition independent of the attack design, while C1–C4 assess the attack itself. Figure 1 provides a schematic overview of the conditions and their implications.

Condition C0: Sensitive disclosure potential. This condition requires the attack to have the potential to bring about a meaningful disclosure of sensitive data. According to the discussion in Section 2, this depends on the data set on which the target model has been trained and requires i) *the training data to be an exhaustive sample of a population*, ii) *uniqueness of confidential attribute values*, and iii) *plausibility on information assumed to be known to the attacker about the target.* C0 can be viewed as a *precondition that is agnostic of the precise*

design of the MIA attack. If C0 does not hold, then an MIA cannot succeed, no matter how well it is crafted.

Condition C1: Non-overfitted model. An ML model is said to overfit when its average performance (*e.g.*, loss or accuracy) on the training data significantly exceeds its performance on unseen data from the same population, *i.e.*, there is a large generalization gap [64]. Overfitting occurs primarily when the model not only learns general patterns but also memorizes sample-specific details and noise, resulting in distinct behaviors for training data (*members*) versus unseen data (*non-members*). In such cases, MIAs can trivially distinguish between members and non-members [77,63]. However, for an MIA to be considered effective, it must succeed against non-overfitted models [12,79], which are the desirable ones in production settings.

To avoid any margin of doubt, we adopt quite a permissive threshold that considers a model as "non-overfitted" whenever the train-test accuracy gap (*i.e.*, $g = \text{Acc}_{\text{train}} - \text{Acc}_{\text{test}}$) is $g \leq 10\%$. Notice that $g > 10$ clearly indicates substantial overfitting or poor generalization.

Condition C2: Competitive model. For an MIA to be considered meaningful, it must target a model that could realistically be deployed in real-world applications and thus be accessible to potential attackers. This necessitates that the model's utility be competitive with state-of-the-art (SotA) benchmarks for its main task. Models that fail to meet this criterion are unlikely to be used in production and hence do not reflect practical privacy risks.

We consider a model to be *competitive* if its test accuracy is not more than 5% lower than that of its SotA —which is not much to ask—, *i.e.*, $\Delta = \text{Acc}_{\text{SotA}} - \text{Acc}_{\text{test}} \leq 0.05 \times \text{Acc}_{\text{SotA}}$. Models falling outside this range are unlikely to be selected for deployment when higher-performing alternatives are available.

Although the specific thresholds chosen for C1 and C2 can be debatable, our choices are deliberately permissive (*e.g.*, allowing up to 10% generalization gap for C1, or 5% deviation from SotA for C2). Because no one can reasonably argue that a model violating these thresholds is non-overfitted or competitive, stricter thresholds would only strengthen our conclusions.

Condition C3: Reliable membership inference. An effective MIA must reliably distinguish members from non-members under realistic conditions. Recent state-of-the-art works [76,12,8,79] focus on measuring the true positive rate (TPR) at an extremely low false positive rate (FPR) (FPR $\leq 0.1\%$) to ensure the reliability of positive inferences. However, this approach overlooks the typically low prior probability of membership, since the training set often represents only a small subset of the overall population. As noted in [37], standard precision does not account for this imbalance.

To address this issue, a weighted precision metric that incorporates the prior membership probability p was proposed in [37]:

$$\text{Prec} = \frac{p \times \text{TPR}}{p \times \text{TPR} + (1-p) \times \text{FPR}}, \tag{1}$$

where $p \ll 50\%$ in realistic settings.

We assess reliability using two criteria:

1. The attack must identify some true members with an FPR near 0%. This is crucial because non-members largely outnumber members in practice, and even small FPR values can lead to many false positives.
2. The weighted precision Prec (as defined above in Equation (1)) must be near perfect, indicating that the positive inferences are indeed true members, even when accounting for low realistic membership priors. Weighted precision also relates to plausible deniability [9], a privacy concept that ensures that an attacker cannot definitively determine whether an individual's data are included in a data set. In the context of MIAs, plausible deniability can be claimed if the attack is imperfect. The level of plausible deniability is inversely proportional to the attack's precision. We require a precision $\geq 95\%$ to counter plausible deniability claims. This ensures that at most 5% of flagged samples are false positives. Although this does not directly correspond to statistical significance ($\alpha = 5\%$), it aligns with a standard threshold to reduce erroneous identifications.

We evaluate the precision with a realistic membership prior $p = 10\%$, which is already conservative relative to many deployment scenarios. For some data sets (*e.g.*, [37]), even lower priors are considered.

Condition C4: Computational feasibility. Even if an attack is reliable against non-overfitted competitive models, it must be feasible with the computational resources reasonably available to potential attackers. This is in line with legal frameworks such as the GDPR (Recital 26) [25], which states that the evaluation of disclosure risk must account for the means reasonably likely to be used in an attack, including cost, time, and available technology.

An attack whose computational cost matches or exceeds that of training the target model is disproportionate, as the attacker's effort would rival or surpass the defender's. In such cases, the attack becomes largely unaffordable in realistic settings. We quantify the computational demands of MIAs through three resource factors, ordered by their contribution to overall cost.

The first and most significant factor is the number of additional models (M) —*e.g.*, shadow, distilled, or reference models— required by the attack: training those models represents a major computational overhead, which can even exceed the effort devoted to building the original target model. The cost is considered *low* when no additional models are required ($M = 0$), *moderate* when a single additional model is needed ($M = 1$) –with complexity on par with the target model–, and *high* when multiple additional models are required ($M \geq 2$).

The second factor is the cost of the inference model (I). This cost is categorized as *low* when a computationally trivial rule is used (*e.g.*, a simple threshold on loss values), *moderate* when a moderately complex model is employed (*e.g.*, a simple binary classifier), and *high* when a computationally expensive inference mechanism is required (*e.g.*, deep neural networks such as [8]).

The third factor is the number of model queries required per target sample (Q). Although not as dominant as M and I, the query count per sample still contributes to the overall cost. This requirement is considered *low* when a maximum of 100 queries per target sample is required, *moderate* when between 101–500 queries are needed, and *high* when a larger number is required. Notice that larger numbers also increase the risk of attack detection.

Aggregation into overall cost levels. The three independently assessed factors are finally aggregated into an overall cost level. A *low-cost* MIA does not require additional models ($M = 0$), uses a low-complexity inference rule, and requires a low number of queries per sample. An example is the simple threshold-based attack by [77] performed on a target model with a single query per sample.

A *moderate-cost* MIA requires a single additional model ($M = 1$), and/or an inference model of moderate complexity and/or a low-moderate number of queries per sample. An example is an attack that trains one shadow model and employs a shallow classifier for inference with 10-200 queries per target sample.

A *high-cost* MIA requires multiple additional models ($M \geq 2$), and/or a high-complexity inference model regardless of the query cost, and/or a very high number of queries per sample when coupled with even moderate levels of the other factors.

4 Review and Evaluation of Representative MIAs

This section evaluates, under the lens of our framework, key contributions on black-box MIAs against deep neural networks (DNNs) used for classification. The latter is by far the most common scenario considered by the surveyed works. We focus on the most influential and representative attacks in the literature, specifically, those presented in top-tier venues and that have attracted the highest number of citations (13,738 citations overall as of January 8, 2026). To gain a broader perspective and analyze their evolution, we include both pioneering (such as [65]) and recent state-of-the-art attacks (such as LiRA [12], [8], and [79]). Detailed descriptions of the surveyed attacks are provided in Appendix A.

Condition C0. In Section 2, we highlighted that for non-trivial membership disclosure to be possible, the training data set should not be an exhaustive sample of the population.

We assess condition C0 by analyzing the data sets used for the classification tasks considered by the MIAs surveyed. *Tabular data sets* include Adult [7], Purchase-100 [65], Texas-100 [65], UCI Credit [30], UCI Hepatitis [1], and UCI Cancer [74]. *Image data sets* are widely used to evaluate MIAs and they range from simple digit recognition (MNIST) to complex object detection (ImageNet). These are: MNIST [42], CIFAR-10 [40], CIFAR-100 [40], CINIC-10 [16], GTSRB (German Traffic Sign Recognition Benchmark) [68], ImageNet-1K [61] (a sample from ImageNet [17]), and LFW (Labeled Faces in the Wild) [32]. *Textual data*

sets include Newsgroups [1] and RCV1X (the Ms RCV1 data set) [43]. Finally, we consider the Locations *trajectory data set* [65].

None of these data sets is an exhaustive sample of the corresponding population, namely, American citizens for Adult, shoppers for Purchase-100, patients for Texas-100, bank customers for UCI Credit, hepatitis patients for UCI Hepatitis, cancerous cells for UCI Cancer, locations of Foursquare users for Locations, images of handwritten digits for MNIST, images in general for CIFAR-10, CIFAR-100, CINIC-10, and ImageNet-1K, German traffic sign images for GTSRB, face images for LFW, text documents for Newsgroups and RCV1X. Hence, *non-trivial membership disclosure is* a priori *possible.*

However, with respect to potential attribute disclosure:

- We explain in Section 2 that unequivocal attribute disclosure is not possible when training data are not exhaustive. Therefore, *none of the reviewed MIAs can result in unequivocal attribute disclosure.*
- Most of the data sets contain public non-sensitive data (MNIST, CIFAR-10, CIFAR-100, ImageNet-1k, CINIC-10, GTSRB, RCV1X, and Newsgroups), which is understandable for reproducibility and to enable comparative analyses. However, even if we consider that some of their attributes might be confidential, membership would only allow inferring them if those data sets were exhaustive and satisfied uniqueness of the allegedly confidential attribute values. That is, the MIA attacker should check that there are no two or more records in the data set that match the attribute values known to the attacker but have different values for the unknown (confidential) attributes. None of the attacks reviewed documents or even considers attribute uniqueness.

Regarding C0, we can conclude that i) the training data sets used are not exhaustive, which means that MIAs are non-trivial, but their results are not unequivocal, and ii) the lack of uniqueness of confidential attributes is a relatively outlying condition that should not significantly affect most data sets used.

Condition C1. The first columns of Table 1 report the evaluation of the model overfit on all the pairs of attack-data sets surveyed. Among the 61 pairs, 24 (39.34%) lack sufficient information to assess overfitting (NA). Among the remaining 37 pairs, 27 (72.97%) use overfitted target models with train-test gaps exceeding 10%, while only 10 (27.03%) satisfy C1.

This behavior is strongly data set-dependent. Simple benchmarks such as Adult and MNIST often meet C1, while complex vision benchmarks (CIFAR-10/100, ImageNet-1K) are frequently evaluated using models with large generalization gaps. Importantly, most violations of C1 are not marginal: many gaps substantially exceed 10%, indicating pronounced overfitting rather than borderline cases.

[1] https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

Condition C2. The evaluation of model competitiveness is reported in the last columns of Table 1. The required test accuracy is missing for 11 pairs (18.03%). Among the remaining 50 pairs, only 12 (24.00%) use target models within 5% of the SotA. The vast majority (38 pairs, 76.00%) evaluate MIAs on models that lag far behind SotA, often by large margins, particularly on CIFAR-10/100 and ImageNet-1K.

Taking into account only the 37 pairs for which C1 was assessable, just 7 pairs (18.92%) satisfy both C1 and C2 simultaneously. Most evaluations fail to meet at least one of the two conditions, due to model overfitting or non-competitive target models. As a result, the MIA effectiveness reported often reflects unduly favorable experimental conditions rather than realistic deployment scenarios.

All in all, the results in Table 1 indicate that a large fraction of the MIA literature evaluates attacks on target models that are not representative of well generalized and competitive systems. This inflates the apparent success of attacks and weakens their relevance for practical privacy risk assessment.

Condition C3. Table 2 reports the performance metrics of the surveyed attacks and their reliability. We derive missing metrics when possible using standard statistical relationships, such as computing FPR from prior, precision, and TPR, or inferring precision from prior, TPR, and FPR. These derived values are underlined in the table. When multiple attack variants are reported (such as in [73]), we consider the one with the highest precision. For attacks reporting precision under the balanced prior ($p = 50\%$), we recompute the weighted precision with $p = 10\%$. An attack is considered reliable only if it maintains precision $\geq 95\%$ with this more realistic prior, ensuring minimal false positives. When the precision with the balanced prior is 100% (indicating 0% FPR), we classify the attack as reliable, since perfect precision is maintained regardless of the prior.

Our analysis reveals that only 16 out of 61 attack-data set pairs (26.2%) demonstrate reliable membership inference by maintaining high precision with realistic priors. Recent approaches like [12] and [79] show particular promise. [12] achieve near-perfect precision ($\geq 99.95\%$) with FPR $\leq 0.001\%$ on CIFAR-10, CIFAR-100, and ImageNet-1K, therefore satisfying reliability even with $p = 10\%$. [79] report 100% precision and 0% FPR on all data sets, making it reliable regardless of the prior. [37] achieve 100% precision on CIFAR-100 and RCV1X, while [50] report 100% precision on MNIST, but only identify one true member. Similarly, [73] achieve 100% precision on UCI Hepatitis, CIFAR-10, and CIFAR-100, although the number of true members identified is limited: 1 for UCI Hepatitis, 2.8 ($\pm$2.4) for CIFAR-10, and 3.0 ($\pm$1.26) for CIFAR-100. This fairly small number of true positives with such high standard deviations adds another layer of uncertainty about the reliability of these attacks.

Earlier attacks such as [65], [77], and [63] generally fail to achieve reliable inference, with precision below our threshold due to the high FPR. This is because these attacks are typically optimized for high accuracy. Some attacks (*e.g.*, [62], [76]) lack critical metrics (precision/FPR), which prevents reliability assessment. It can also be seen that the attack in [48] exhibits substantial drops in precision

Table 1: Evaluation of model non-overfitting (C1) and competitiveness (C2)

Attack	Data set	Train acc. (%)	Test acc. (%)	g (%)	Non-overfitted?	SotA acc. (%)	Δ (%)	Competitive?
[65]	Adult	84.80	84.20	0.60	✓	88.16 [14]	3.96	✓
	Purchase-100	NA	72.90	NA	NA	90.00 [69]	17.10	✗
	Texas-100	NA	57.00	NA	NA	57.00 [65]	0.00	✓
	Locations	100.00	67.30	32.70	✗	72.00 [45]	4.70	✓
	MNIST	98.40	92.80	5.60	✓	99.87 [11]	7.07	✗
	CIFAR-10	NA	60.00	NA	NA	99.50 [22]	39.50	✗
	CIFAR-100	NA	20.00	NA	NA	96.08 [26]	76.08	✗
[77]	MNIST	NA	NA	NA	NA	99.87 [11]	NA	NA
	CIFAR-10	NA	NA	NA	NA	99.50 [22]	NA	NA
	CIFAR-100	NA	NA	NA	NA	96.08 [26]	NA	NA
[63]	Purchase-100	NA	~ 80.00	NA	NA	90.00 [69]	10.00	✗
	Locations	NA	~ 62.00	NA	NA	72.00 [45]	10.00	✗
	MNIST	NA	~ 98.00	NA	NA	99.87 [11]	1.87	✓
	CIFAR-10	NA	~ 60.00	NA	NA	99.50 [22]	39.50	✗
	CIFAR-100	NA	~ 22.00	NA	NA	96.08 [26]	74.08	✗
	LFW	NA	~ 68.00	NA	NA	99.83 [18]	31.83	✗
[62]	CIFAR-10	NA	NA	NA	NA	99.50 [22]	NA	NA
	ImageNet-1K	NA	NA	NA	NA	94.90 [55]	NA	NA
[50]	Adult	85.00	85.00	0.00	✓	88.16 [14]	3.16	✓
	UCI Cancer	95.00	94.00	1.00	✓	98.86 [74]	4.86	✓
	MNIST	99.00	99.00	0.00	✓	99.87 [11]	0.87	✓
[37]	Purchase-100X	100.00	71.00	29.00	✗	90.00 [69]	19.00	✗
	Texas-100	100.00	53.00	47.00	✗	57.00 [65]	4.00	✓
	CIFAR-100	48.00	18.00	30.00	✗	96.08 [26]	78.08	✗
	RCV1X	100.00	84.00	16.00	✗	92.85 [39]	8.85	✗
[66]	Purchase-100	99.80	80.90	18.90	✗	90.00 [69]	9.10	✗
	Texas-100	81.00	52.30	28.70	✗	57.00 [65]	4.70	✓
	Locations	100.00	60.70	39.30	✗	72.00 [45]	11.30	✗
	CIFAR-100	100.00	83.00	17.00	✗	96.08 [26]	13.08	✗
[48]	Purchase-100	NA	NA	NA	NA	90.00 [69]	NA	NA
	Locations	NA	NA	NA	NA	72.00 [45]	NA	NA
	CIFAR-10	100.00	82.60	17.40	✗	99.50 [22]	16.90	✗
	CINIC-10	99.90	65.80	34.10	✗	95.06 [3]	29.25	✗
	CIFAR-100	99.90	47.50	52.40	✗	96.08 [26]	48.58	✗
	GTSRB	100.00	88.10	11.90	✗	99.71 [4]	11.61	✗
	Newsgroups	NA	NA	NA	NA	89.50 [46]	NA	NA
[73]	Adult	90.40	84.20	6.20	✓	88.16 [14]	3.96	✓
	UCI Credit	92.10	75.10	17.00	✗	83.20 [30]	8.10	✗
	UCI Hepatitis	99.80	87.10	12.70	✗	97.44 [1]	10.34	✗
	MNIST	99.00	98.90	0.10	✓	99.87 [11]	0.97	✓
	CIFAR-10	94.60	74.00	20.60	✗	99.50 [22]	25.50	✗
	CIFAR-100	93.80	37.30	56.50	✗	96.08 [26]	58.78	✗
	ImageNet-1K	77.30	63.70	13.60	✗	94.90 [55]	31.20	✗
[76]	Purchase-100	100.00	75.50	24.50	✗	90.00 [69]	14.50	✗
	MNIST	98.60	97.10	1.50	✓	99.87 [11]	2.77	✓
	CIFAR-10	99.89	77.37	22.52	✗	99.50 [22]	22.13	✗
	CIFAR-100	97.90	20.40	77.50	✗	96.08 [26]	75.68	✗
[12]	Purchase-100	NA	NA	NA	NA	90.00 [69]	NA	NA
	Texas-100	NA	NA	NA	NA	57.00 [65]	NA	NA
	CIFAR-10	100.00	90.00	10.00	✓	99.50 [22]	9.50	✗
	CIFAR-100	100.00	60.00	40.00	✗	96.08 [26]	36.08	✗
	ImageNet-1K	100.00	65.00	35.00	✗	94.90 [55]	29.90	✗
[8]	CIFAR-10	NA	91.0	NA	NA	99.50 [22]	8.50	✗
	CINIC-10	NA	NA	NA	NA	95.06 [3]	NA	NA
	CIFAR-100	NA	68.60	NA	NA	96.08 [26]	27.48	✗
	ImageNet-1K	NA	67.50	NA	NA	94.90 [55]	27.40	✗
[79]	Purchase-100	100.00	83.40	16.60	✗	90.00 [69]	6.60	✗
	CIFAR-10	99.90	92.40	7.50	✓	99.50 [22]	7.10	✗
	CINIC-10	99.50	77.20	22.30	✗	95.06 [3]	17.86	✗
	CIFAR-100	99.90	67.50	32.40	✗	96.08 [26]	28.58	✗
	ImageNet-1K	90.20	58.60	31.60	✗	94.90 [55]	36.30	✗

(for example, 98.65% to 89.02% for CIFAR-100) when evaluated with $p = 10\%$, which emphasizes the importance of realistic evaluation settings.

In summary, most attacks fail to satisfy condition C3 due to high FPR or precision degradation with realistic priors. Only attacks with 0% FPR (*e.g.*, [79] and [73] on CIFAR-10/100) or extremely low FPR ($\leq 0.001\%$) paired with $\geq 95.0\%$ precision (*e.g.*, [12]) satisfy C3.

Condition C4. Table 3 reports the resource requirements and characterizes the computational cost of the attacks surveyed, categorizing them into overall low, moderate, or high computational cost according to the criteria depicted in Section 3. We can see that 8 out of 13 attacks (about 62%) incur high computational costs. For example, the state-of-the-art attack LiRA [12], which demands 32 to 256 shadow models, is unaffordable in most realistic scenarios. RMIA [79] substantially reduces the number of models required (1-4) compared to LiRA, although at the expense of increased query volume. The approach of [8] eliminates shadow models, but requires training a deep quantile regression model and substantial hyperparameter tuning, resulting in significant computational overhead that limits its efficiency [8,79]. Three attacks fall into the moderate category and require only one additional model and a maximum of 100 queries per sample, while only simple global threshold-based methods [77,63] require no additional models and minimal queries, making their computational cost low.

In summary, although recent advances in MIAs have improved attack reliability, their computational feasibility remains a critical consideration.

Overall attack effectiveness. We assess the overall effectiveness of the surveyed attacks based on their simultaneous fulfillment of conditions C1, C2, C3, and C4, as demanded by our evaluation framework. We can see that even with the permissive thresholds we employed, *all attacks fail to simultaneously meet all four conditions, and hence no attack is deemed effective.* They either target overfitted models (violating C1), use models that significantly underperform w.r.t. the state of the art (violating C2), do not achieve high membership inference precision under realistic priors (violating C3), or have impractical computational costs (violating C4). Any of these violations makes the attack unrealistic in production settings. The only attack-data set pairs that meet three of the four conditions are [50] and [73] on MNIST. Specifically, [50] on MNIST meets C1, C2, and C3 but does not meet C4 because it requires 50 additional models to carry out the attack. On the other hand, [73] on MNIST fails to produce reliable inference (C3 not satisfied), but at least meets C1, C2 and C4. Although the state-of-the-art attacks LiRA [12] and RMIA [79] achieve reliable inferences in most data sets (fulfilling C3), they generally do not satisfy the remaining conditions.

5 Discussion

In the following, we discuss the main limitations and challenges that we observe in current MIAs.

Table 2: Evaluation of attack performance and membership reliability (C3)

Attack	Data set	Attack performance (%)			Prior p (%)	Reliable?
		Prec.	TPR	FPR		
[65]	Adult	50.30	NA	NA	50	✗
	Purchase-100	73.00	86.00	31.81	50	✗
	Locations	67.80	98.00	46.54	50	✗
	Texas-100	69.00	86.00	38.64	50	✗
	MNIST	51.70	NA	NA	50	✗
	CIFAR-10	71.00	99.00	40.44	50	✗
	CIFAR-100	97.00	99.00	3.06	50	✗(Prec 78.24% with $p = 10\%$)
[77]	MNIST	50.50	99.00	97.04	50	✗
	CIFAR-10	69.40	99.00	43.65	50	✗
	CIFAR-100	87.40	99.00	14.27	50	✗
[63]	Purchase-100	$\sim$ 55.00	$\sim$ 100.0	$\sim$ 81.82	50	✗
	Locations	$\sim$ 80.00	$\sim$ 95.00	$\sim$ 23.75	50	✗
	MNIST	$\sim$ 50.00	$\sim$ 99.00	$\sim$ 99.00	50	✗
	CIFAR-10	$\sim$ 60.00	$\sim$ 15.00	$\sim$ 10.00	50	✗
	CIFAR-100	$\sim$ 90.00	$\sim$ 100.00	$\sim$ 11.11	50	✗
	LFW	$\sim$ 70.00	$\sim$ 100.0	$\sim$ 42.86	50	✗
[62]	CIFAR-10	NA	NA	NA	50	NA
	ImageNet-1K	NA	NA	NA	50	NA
[50]	Adult	73.91	NA	NA	50	✗
	UCI Cancer	88.89	NA	NA	50	✗
	MNIST	100	NA	0.00	50	✓
[37]	Purchase-100X	98.00	NA	NA	50	NA
	Purchase-100X	97.50	NA	NA	9.0	✓
	Texas-100	95.70	NA	NA	50	NA
	Texas-100	97.40	NA	NA	33	NA
	CIFAR-100	100.0	NA	NA	50	✓
	CIFAR-100	100.0	NA	NA	33	✓
	RCV1X	100.0	NA	NA	50	✓
	RCV1X	93.00	NA	NA	9.0	✗
[66]	Purchase-100	63.40	7.80	4.50	50	✗
	Texas-100	85.40	21.20	3.62	50	✗
	Locations	NA	NA	NA	50	NA
	CIFAR-100	NA	NA	NA	50	NA
[48]	Purchase-100	97.14	3.40	0.10	50	✗(Prec 84.48% with $p = 10\%$)
	Locations	98.00	4.90	0.10	50	✗(Prec 84.00% with $p = 10\%$)
	CIFAR-10	93.75	1.50	0.10	50	✗(Prec 62.50% with $p = 10\%$)
	CINIC-10	98.21	5.50	0.10	50	✗(Prec 85.94% with $p = 10\%$)
	CIFAR-100	98.65	7.30	0.10	50	✗(Prec 89.02% with $p = 10\%$)
	GTSRB	98.65	7.30	0.10	50	✗(Prec 89.02% with $p = 10\%$)
	Newsgroups	96.15	2.50	0.10	50	✗(Prec 73.53% with $p = 10\%$)
[73]	Adult	80.00	NA	0.0	0.50	✗
	UCI Credit	90.00	NA	0.0	0.50	✗
	UCI Hepatitis	100.0	NA	0.0	0.50	✓
	MNIST	66.00	NA	NA	0.50	✗
	CIFAR-10	100.0	NA	0	0.50	✓
	CIFAR-100	100.0	NA	0	0.50	✓
	ImageNet-1K	NA	NA	NA	0.50	NA
[76]	Purchase-100	NA	NA	NA	0.50	NA
	MNIST	NA	NA	NA	0.50	NA
	CIFAR-10	NA	NA	NA	0.50	NA
	CIFAR-100	NA	NA	NA	0.50	NA
[12]	Purchase-100	97.62	4.10	0.10	0.50	✗(Prec 82.00% with $p = 10\%$)
	Texas-100	99.70	33.20	0.10	0.50	✓
	CIFAR-10	99.95	2.20	0.001	0.50	✓
	CIFAR-100	99.99	11.20	0.001	0.50	✓
	ImageNet-1K	99.88	0.85	0.001	0.50	✓
[8]	CIFAR-10	64.48	0.18	0.10	0.50	✗
	CINIC-10	85.46	0.59	0.10	0.50	✗
	CIFAR-100	85.41	0.59	0.10	0.50	✗
	ImageNet-1K	99.64	27.68	0.10	0.50	✓
[79]	Purchase-100	100	0.41	0.00	0.50	✓
	CIFAR-10	100	2.13	0.00	0.50	✓
	CINIC-10	100	1.39	0.00	0.50	✓
	CIFAR-100	100	3.50	0.00	0.50	✓
	ImageNet-1K	100	0.06	0.00	0.50	✓

Table 3: Evaluation of computational feasibility (C4). Cost factors: M: additional models required, I: inference model complexity, Q: queries required per target sample.

Attack	Resource requirements	Cost factors			Overall
		M	I	Q	cost
[65]	10-100 additional models; K ML attack models; 1 query	High	Moderate	Low	High
[77]	Simple thresholding rule on loss values; 1 query	Low	Low	Low	Low
[63]	Simple thresholding rule; 1K random data points per data set; 1 query	Low	Low	Low	Low
[62]	30 additional models; 1 query	High	Low	Low	High
[50]	50 additional models; 1 query	High	Low	Low	High
[37]	1 additional model; 100 queries	Moderate	Low	Low	Moderate
[66]	1 additional model; 1 query	Moderate	Low	Low	Moderate
[48]	2 additional models; 1 ML attack model; multiple queries	High	Moderate	Moderate	High
[73]	1 additional model; 1 query	Moderate	Low	Low	Moderate
[76]	15-999 additional models; 1 query	High	Low	Low	High
[12]	32-256 additional models; under 100 queries	High	Low	Low	High
[8]	Multiple quantile regression models; 1 query	Low	High	Low	High
[79]	1-4 additional models; as per paper, 10% of training set size is sufficient	High	Low	High	High

5.1 Fundamental Limitations of MIAs

The membership paradox. Our analysis reveals a fundamental paradox in membership inference: when training data are non-exhaustive (as most real-world training sets are), there exists plausible deniability and, therefore, intrinsic protection against any MIA-enabled attribute disclosure; conversely, when training data are exhaustive and attribute disclosure might be more certain, membership is already known and MIAs become irrelevant. This undermines the privacy threat posed by MIAs in a way that is agnostic of the attack design. Although non-unequivocal inferences with moderate precision under skewed priors may provide leads for further investigation, they still substantially weaken the certainty of individual membership claims and provide plausible deniability to data subjects.

Overfitting vs. practical deployment. We must distinguish between intentional overfitting —where models are deliberately overfitted to facilitate MIAs, as it was the case in several of the reviewed works—, and unavoidable overfitting —which arises naturally due to data set complexity—. Our framework focuses on privacy risks exploitable by realistic attackers, rather than artifacts of data or training processes. State-of-the-art MIAs often exaggerate their effectiveness by training target models on small subsets of available data and reserving large portions (*e.g.*, 50%) for shadow models. This practice induces overfitting and deviates significantly from real-world ML practices, where models are trained on large, diverse data sets with the ultimate goal of maximizing generalization.

Data diversity and size mitigate risks. Increasing the diversity and size of training data sets generally reduces the precision of MIAs [65]. Modern ML systems, particularly in high-stakes domains, typically utilize large and diverse training sets, which inherently provides protection against membership inference.

5.2 Methodological Issues in Current MIA Research

Reference data dependence. A critical limitation of state-of-the-art MIAs is their reliance on reference data sets that closely mirror the target model's training distribution. Attacks like LiRA [12] require shadow models trained on data similar to the target private data set. As demonstrated by [27], this dependence leads to high false positive rates when reference data differ, which is a common scenario when private or proprietary data sets are used to fine-tune modern ML models such as large language models (LLMs).

Unrealistic membership priors. The reliability of MIAs is further undermined by realistic membership priors, where non-members vastly outnumber members. As shown by [73], increasing the number of non-member samples improves attack accuracy by favoring non-membership inference, but this comes at the cost of degraded precision-recall trade-offs. In real-world scenarios with low membership priors (10% or less), achieving high precision often results in low recall, limiting the attack's ability to reliably identify members. This highlights the need to evaluate MIAs under realistic priors rather than the unrealistic balanced settings (50% of members) commonly used in the literature.

Inconsistent positive inference. For MIAs to be reliable, positive membership inferences must remain consistent across different random initializations and small variations in training data. However, [76] observe that the same target sample does not consistently produce a positive membership inference across different target models trained with varying initializations or slight changes in training data. Even when MIAs report high precision, their practical impact is limited by inconsistent results and low coverage. High standard deviations in multiple runs mean that *different* records are flagged as vulnerable depending on the initialization of the model or hyperparameters, thus reducing confidence in any single inference.

Reproducibility concerns. The previous issues are compounded by reproducibility challenges that further erode confidence in the reported performance of MIAs. For example, when reproducing the offline LiRA attack using the authors' own code, [79] observed significantly lower attack accuracy than reported in [12]. This inconsistency may be attributed to different training initializations or data splitting approaches, highlighting the instability of current MIA techniques.

5.3 Implications for Privacy-preserving ML

Unjustified utility loss. Given that MIAs pose a weak privacy threat (or none at all) under realistic conditions, the utility cost of defenses to protect against MIAs, such as DP or other noise addition techniques, appears disproportionate and unwarranted. In critical domains such as healthcare, even a 1% drop in performance can have serious consequences, particularly when diagnosing rare diseases. Therefore, practitioners should carefully weigh the actual privacy risks (if any) against utility losses before implementing systematic protection measures.

Targeted disclosure protection. Given that attribute disclosure risks come from training data features (condition C0), instead of applying general mechanisms to counter MIAs, a better balance between privacy and utility would be achieved by employing safeguards especially designed to break disclosure-enabling data features. Specifically, data sampling counteracts the exhaustiveness of the training data set [34], whereas formalisms such as l-diversity [51] break confidential attribute uniqueness by introducing a minimum variance in confidential attributes. Other measures may involve replacing outlying or otherwise vulnerable attribute values with synthetic imputed values (partial synthesis). Compared to DP, which systematically *distorts all data or model parameters with some probability*, the aforementioned methods *target specific data points or data features with certainty*.

Regulatory implications. Our findings suggest that current regulatory frameworks may overestimate the privacy risks associated with ML models. Unless MIAs are significantly improved or new privacy-meaningful attacks are devised, the alleged privacy risks of predictive ML seem mild enough to warrant reconsideration of stringent privacy requirements that might hinder innovation and utility.

5.4 Towards Improved Evaluation Standards

Standardized reporting practices. The lack of consistent and comprehensive reporting in MIA research hampers meaningful comparisons between studies. Critical metrics, such as training and test loss curves during training, accuracy/loss gaps after training, the number of true positives from member samples, and the stability of the results, are often omitted. We advocate for future studies to consistently report these metrics, along with runtime costs and standard deviations across multiple runs with realistic priors. Only this will enable a realistic evaluation of whether MIAs are effective or simply succeed under contrived conditions.

Alternative privacy metrics. Given the limitations of MIAs, rather than focusing exclusively on them as a proxy for privacy risk assessment, researchers should explore alternative metrics to assess meaningful privacy risks in ML, such as re-identifiability and actual attribute disclosure. These might include measures of information leakage that do not rely on binary membership inference, evaluations of model vulnerability to reconstruction attacks, or assessments of how model outputs correlate with sensitive attributes in the training data. This balanced assessment would help avoid unnecessary protection measures while ensuring appropriate safeguards where genuinely needed.

6 Conclusions, Recommendations, and Future Research

MIAs are the main method to assess privacy risks in ML and a building block for other privacy attacks such as data extraction or reconstruction. In this work, we have shown that MIAs are unlikely to disclose sensitive information. On the

one hand, the data used to train the target model must satisfy very stringent conditions such as exhaustivity and uniqueness of confidential attribute values; otherwise, any attribute disclosure is plausibly deniable. On the other hand, state-of-the-art MIAs fail to offer reliable performance on non-overfitted, competitive target models at a reasonable computational cost.

Thus, our analysis demonstrates that MIAs generally do not constitute significant privacy threats under realistic conditions. Hence, relying on MIAs as primary privacy risk metrics can lead to overestimation of privacy vulnerabilities and unjustified utility loss caused by unnecessary protection techniques. This finding has important implications for both research directions and practical privacy-preserving ML deployment.

A broader ramification is that, unless MIAs are improved or more effective privacy attacks are devised, the alleged privacy risks of predictive ML seem mild enough to consider a relaxation of regulatory frameworks.

Recommendations to practitioners that release trained models include:

- Check whether the data that will be used to train the model to be released satisfy the necessary conditions listed in Section 2 for the unequivocal disclosure of sensitive information.
- If they do not, then refrain from using privacy protection techniques that may result in unwarranted utility loss.
- If they do, resort to privacy-preserving mechanisms that target problematic data features (such as sampling or partial synthesis), and make sure that the trained model is not overfitted. To fix overfitting, standard anti-overfitting techniques should offer a better privacy-utility trade-off than DP [10].

As future work, we plan to analyze the potential of the privacy attacks proposed for generative models, including large language models and diffusion models, which have been shown to memorize and leak data from the training set [13]. For these models, it is less obvious how to formulate and check conditions on the training data, which do not consist of a single specific data set.

Acknowledgments. Thanks go to Alberto Blanco-Justicia for discussions on earlier versions of this paper. Partial support for this work has been received from the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez), MCIN/AEI under grant PID2024-157271NB-I00 "CLEARING-IT", and the European Commission under project HORIZON-101292277 "SoBigData IP".

This appendix provides supplementary material supporting the main text. It includes detailed descriptions of the surveyed attacks and additional background information.

A Surveyed Attacks

This section provides the detailed survey referenced in the main manuscript. In particular, Table 4 summarizes the representative black-box membership infer-

ence attacks (MIAs) reviewed in the paper, including their core ideas, publication venues, and citation counts.

Table 4: Summary of the surveyed attacks. Citation counts according to Google Scholar, January 8, 2026.

Attack	Approach	Venue	# citations
[65]	ML membership classifier on predictions from shadow models	IEEE SP 2017	7,083
[77]	Loss global thresholding	IEEE CSF 2018	1,751
[63]	Confidence and entropy global thresholding	NDSS 2019	1,329
[62]	Per-sample loss calibration and thresholding	ICML 2019	515
[50]	Hypothesis testing on loss values of selected vulnerable records	Euro SP 2020	143
[37]	Perturbed input loss thresholding	PoPETs 2021	194
[66]	Class-specific modified entropy thresholding	USENIX Security 2021	574
[48]	ML membership classifier on the sample's loss of trajectory from distilled models	ACM CCS 2022	165
[73]	Per-sample loss calibration and thresholding	ICLR 2022	203
[76]	Hypothesis testing on loss values from reference/distilled models	ACM CCS 2022	429
[12]	Hypothesis testing based on likelihood ratio of scores from shadow models	IEEE SP 2022	1180
[8]	Quantile regression on confidence scores	NeurIPS 2024	81
[79]	Hypothesis testing based on likelihood ratio of scores from shadow models	ICML 2024	91

The first notable MIA against DNN models was presented in [65], and involves training multiple "shadow" models on data sampled from a distribution similar to that of the target model to replicate its behavior. The prediction outputs of these shadow models on both their training and non-training data are labeled according to membership status. These labeled outputs are then used to train attack models that learn to distinguish members from non-members based on patterns in the model outputs.

[63] relax some assumptions made by [65] and demonstrate that MIAs could be conducted using one of the following threat models: 1) only a single shadow model and prior knowledge of training data distribution (Adversary 1); 2) a single shadow model and no prior knowledge of the target model's architecture or training data distribution (Adversary 2); or 3) no need to train shadow models at all (Adversary 3). For Adversary 3, the authors introduce label- and shadow model-free attacks based on the prediction entropy of the target model $f_\theta(x)$ or the maximum confidence score assigned by the target model for a given input x. In these cases, it is suggested that the membership status decision be made by using a global threshold τ, such as percentiles (*e.g.*, the top 10%), based on model output statistics derived from random or synthetic data points representing non-members.

[77] explore the privacy risks related to overfitting and theoretically show that higher generalization errors make models more vulnerable to MIAs, as members often exhibit higher prediction confidence or lower loss values compared to non-members. They propose several simple and low-cost MIAs based on the generalization gap of the target model. Among them, a method computes the loss of the target model f_θ on (x, y), and if the loss is below the expected training loss, the point is inferred to be a member.

Score-based attacks proposed by [77,63] exploit the training objective of minimizing prediction loss in training data, which often results in training samples

that achieve near-maximal confidence for their true labels, while test samples exhibit lower confidence. In such cases, a global threshold τ can be applied to infer membership: samples with confidence exceeding τ (or loss below τ) are classified as members.

[66] demonstrate that score-based approaches can be improved using class-specific thresholds τ_y (for a class label y) instead of a single global threshold for loss values. The intuition is that an unbalanced data set can cause the target model to exhibit varying confidence levels across different class labels. The class-dependent thresholds τ_y are learned by training a shadow model to mimic the behavior of the target model, collecting the shadow model's metric values (*e.g.*, prediction confidence) on both shadow training and test data, and selecting τ_y to maximize the accuracy of distinguishing between members and non-members of the shadow model for class y. Additionally, they propose using a modified prediction entropy of the sample as a metric, which incorporates information about the ground-truth label.

[37] leverage the observation that training samples are typically near a local minimum of the loss function of the model. Their attack perturbs an input x with fresh Gaussian noise, queries the model on perturbed inputs, and counts how often the perturbed inputs result in a higher loss than that of (x, y), where y is the class label. If this count exceeds a specified threshold τ, (x, y) is classified as a member. To define the threshold, a shadow model is trained on data sampled from the target model data distribution, loss values are collected on shadow training/test data, and a global threshold τ is selected that maximizes TPR while constraining FPR to a desired level (*e.g.*, $\alpha = 1\%$). Unlike prior attacks that assume an unrealistic balanced membership prior ($p = 0.5$), they evaluate the attack precision under a skewed prior $p \ll 0.5$, because in practical scenarios members are often a small subset of a broader population.

Other works focus on addressing a significant challenge shared by the above-mentioned MIAs: the high false-positive rate, which undermines the reliability of these attacks because they reduce confidence in the attack's predictions.

[50] aim to improve the precision of the attack by selectively targeting "vulnerable" samples that have a unique influence on the target model. To identify these samples, the authors compute the cosine similarity between the feature representations of the data points and select the top 10% with the greatest distances from their nearest neighbors. The intuition here is that samples with fewer neighbors are more likely to impose a unique influence on the model. After that, they train multiple reference models on data sets that exclude these vulnerable samples to estimate the loss distribution for non-members. Membership inference is then performed using a statistical hypothesis test. The attacker queries the target model with a record and computes a p-value under the null hypothesis that the record is a non-member, based on the estimated non-member loss distribution. If the p-value falls below a predefined threshold (*e.g.*, 0.01), the record is classified as a member.

In practical scenarios, well-generalized target models typically exhibit similar behavior on member and non-member samples, making it challenging to differ-

entiate between them. To address this issue, [48] exploit the differences in the training loss trajectories of the member and non-member samples. First, they use knowledge distillation to estimate the training trajectory of the target model. Then, they train a binary ML classifier using the concatenated loss values from the student model across training epochs, along with the loss from the target model, to capture membership patterns and distinguish between members and non-members.

[62] demonstrate that the optimal MIA relies solely on a model's loss function, with white-box access providing no additional information beyond the loss itself. The attack assigns a sample-specific score $\mathcal{A}'(x, y) = -\mathcal{L}(f_\theta(x), y) + \tau_{x,y}$, where the calibration term $\tau_{x,y}$ accounts for the inherent prediction difficulty of (x, y) when excluded from training. A low $\tau_{x,y}$ indicates that (x, y) is naturally easy to predict, which means that a low sample loss from the target model does not necessarily imply positive membership. To approximate $\tau_{x,y}$, the authors train multiple shadow models, half of them including (x, y) in their training set (IN models) and half excluding it (OUT models). They then determine the threshold that best separates the loss distributions of the IN and OUT models for each sample.

To reduce computational cost, [73] approximate the difficulty calibration by setting the term $\tau_{x,y}$ to the average score (based on metrics such as loss or confidence) calculated from a set of OUT shadow models that do not include (x, y) in their training data. The intuition is that some non-member samples may still exhibit high membership scores due to their being easy to predict, which results in high false positives. Hence, if a sample is easy to predict both for the target model and for the shadow models, its membership score should be reduced accordingly. [73] demonstrate that this approach can improve the inference performance of several existing attacks, including those based on loss, gradient norm, and confidence scores.

[76] designed sample-dependent (Attack R), and model-and-sample-dependent (Attack D) MIAs by training N shadow/reference/distilled models. In both attacks, the target sample (x, y) is excluded from their training data. Inference is performed by conducting an one-sided hypothesis test on the losses computed on (x, y) by the N models. To target a specific FPR α, their attack sets the decision threshold (depending on the attack, on the target model and/or sample) so that a fraction α of the measured losses lies below the threshold.

[12] introduced the Likelihood Ratio Attack (LiRA), which improves over [62,76] by modeling the confidence scores on a given sample (x, y) from multiple IN and OUT shadow models as Gaussian distributions using logit scaling trick. It performs a parametric likelihood-ratio test between the two distributions to infer membership, with thresholds set to achieve a desired low FPR. Two variants of the attack are presented: online and offline. The online version trains IN and OUT shadow models for each target sample, which allows for precise modeling but incurs significant computational cost. To mitigate this limitation, the offline version pretrains only the OUT shadow models and measures the probability of observing a score as high as the target model in the OUT scores distribution.

Despite the effectiveness of [62,76,73,12] in achieving high TPR at low FPR, their computational costs render them unscalable for practical privacy auditing [79].

Recent methods by [8] and [79] aim to address the scalability limitations of LiRA and related attacks. [79] demonstrated that, under practical computational budgets (e.g., two shadow models), LiRA's performance degrades to near-random guessing, while Attack-R [76] shows low true-positive rates (TPR) at low FPR.

[8] proposed an attack that requires a single regression model and no knowledge of the architecture of the target model. Their approach involves querying the target model using a data set not included in its training to obtain confidence scores. Then a quantile regression model is trained on these confidence scores to predict the $1-\alpha$ quantile of the confidence score distribution, allowing input-specific thresholds for membership inference.

[79] introduced RMIA, which constructs a membership score by comparing the likelihood of the model's confidence scores when (x, y) is in the training set versus when a random data point z is used. The authors show that RMIA can achieve comparable membership detection with fewer reference models (between 1 and 4 models) than LiRA.

B Background

B.1 Machine Learning

A classification machine learning model $f_\theta : \mathcal{X} \to [0,1]^C$ maps an input $x \in \mathcal{X}$ to a probability distribution over C classes, where $f_\theta(x)_y$ denotes the predicted probability for class y [28]. Given a training set D_{train} drawn from an underlying distribution $\mathbb{D}$, the goal is to train f_θ so that it generalizes well to an unseen test set $D_{\text{test}} \sim \mathbb{D}$. Generalization is usually measured by the difference in performance (*e.g.*, accuracy or loss) between the training and test sets.

Our survey focuses on MIAs targeting classification deep neural networks (DNNs), as they are among the most commonly used in the literature.

A DNN model can be represented as $f(x) = \sigma(z(x))$, where $z : \mathcal{X} \to \mathbb{R}^C$ returns unnormalized logits, followed by the softmax function $\sigma(\cdot)$, which converts logits into a probability vector of length K. DNNs are commonly trained using stochastic gradient descent (SGD) [42] to minimize a chosen loss function $\mathcal{L}$. For classification tasks, the cross-entropy loss is widely used, that is,

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(p_i),$$

where C is the number of classes, y_i is the true probability for class i (usually 1 for the true class and 0 for all others), and p_i is the predicted probability for class i.

B.2 Differential Privacy

Differential privacy (DP) [24] is a privacy framework designed to protect individual contributions within a data set by adding calibrated noise to the outputs. Formally, a mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if, for any two neighboring data sets D and D' differing by a single entry, and for any subset $S \subseteq \text{Range}(\mathcal{M})$, it holds that

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

where the privacy budget ϵ controls the level of privacy, with smaller values indicating stronger privacy guarantees, and δ is the probability that the privacy guarantee may fail.

Originally proposed to protect released statistical data, DP has become the standard method to enhance privacy in a variety of applications, including data releases [21] and privacy-preserving ML [47]. In particular, DP is recognized as a rigorous defense against MIAs because, if meaningfully applied, it limits the influence of *any* single training data point on the resulting model, thus mitigating the risk that its membership can be inferred [77].

B.3 Privacy-preserving Model Training

To enhance the privacy of DNNs, two main training-time approaches have been employed: DP and anti-overfitting techniques. Incorporating DP in DNN training often involves modifying the SGD algorithm by clipping the gradients and adding calibrated noise to the clipped gradients [2]. Many works have enforced DP on DNN models to mitigate MIAs [77,78,15,33,36,37,44,58]. Although DP theoretically offers formal privacy guarantees, its practical implementation cannot afford meaningful values (*i.e.*, small enough) of its privacy budget ϵ without significantly degrading the utility of trained models [36,10].

Anti-overfitting techniques, on the other hand, provide empirical privacy protection (without formal guarantees) while better retaining (or even improving) model utility [10,12]. These defenses include methods such as weight regularization [41], weight dropout [67], data augmentation [72,80], learning rate tuning [35], and early stopping [57].

B.4 Membership Inference Attacks

Membership inference attacks (MIAs) aim to determine whether a given sample (x, y) was part of the training data of a target model f_θ [65], where x denotes the input feature and y the corresponding class label. Black-box MIAs only require the ability to query the model and observe its outputs, whereas white-box attacks need full access to the internal parameters of the model [54,31]. Due to this stringent requirement, approaches based on white-box MIAs are often unfeasible in real-world scenarios where attackers usually only have black-box access to the model. On the other hand, as shown in [62], the optimal attack

strategy is primarily based on the loss function of the model, making black-box attacks nearly as effective as their white-box counterparts. Due to these reasons, in this work we focus on black-box MIAs, which can be formalized as a security game between a challenger $\mathcal{C}$ and an attacker $\mathcal{A}$ following priorworks [77,37,12].

Definition 1 (Membership Inference Security Game).

1. *$\mathcal{C}$ samples a dataset $D_{train} \leftarrow \mathbb{D}$ and trains a model f_θ on D_{train}.*
2. *$\mathcal{C}$ randomly samples $b \in \{0,1\}$, where $b = 1$ with probability p (most MIAs assume $p = 0.5$).*
3. *If $b = 1$, $\mathcal{C}$ samples (x, y) from D_{train}; otherwise, (x, y) is sampled from $\mathbb{D} \setminus D_{train}$.*
4. *$\mathcal{C}$ gives $\mathcal{A}$ access to (x, y) and query access to f_θ, as well as potential access to the data distribution $\mathbb{D}$.*
5. *The attacker outputs $\hat{b} = \mathcal{A}(x, y)$.*
6. *The game outputs 1 if $\hat{b} = b$, and 0 otherwise.*

Instead of directly outputting a binary decision, $\mathcal{A}$ often computes a score $\mathcal{A}'(x, y)$ based on metrics such as the model's loss or prediction confidence, which is thresholded at τ to produce the final membership prediction:

$$\mathcal{A}(x, y) = \mathbf{1}[\mathcal{A}'(x, y) > \tau]. \tag{3}$$

The attack outcomes are true positive (TP) when $b = 1$ and $\hat{b} = 1$, false positive (FP) when $b = 0$ and $\hat{b} = 1$, true negative (TN) when $b = 0$ and $\hat{b} = 0$, and false negative (FN) when $b = 1$ and $\hat{b} = 0$.

Evaluation Metrics. Evaluating MIAs' performance involves a variety of metrics commonly used in binary classification. The true positive rate *TPR*, also called *recall*, is computed as $TP/(TP + FN)$, and is the fraction of real members correctly identified. The false positive rate *FPR* is computed as $FP/(FP + TN)$, and is the fraction of non-members incorrectly classified as members (false alarms).

Accuracy (Acc), computed as $(TP+TN)/(TP+TN+FP+FN)$, measures overall correct predictions, while the area under the ROC curve (AUC) measures the ranking quality of members over non-members across various thresholds, with $AUC = 1$ being perfect classification and $AUC = 0$ being random classification.

Although *Acc* and AUC provide measures of average-case performance, they can overlook the reliability of positive predictions. In particular, AUC aggregates performance for all *FPRs*, including regions —such as $FPR > 10\%$— that are practically irrelevant, since the *TPR* in these regions does little to capture the efficacy of real-world attacks [59,12,73]. Similarly, optimizing for accuracy can inadvertently inflate *FPR*, thereby compromising the reliable detection of membership [73].

Precision (Prec), defined as $TP/(TP + FP)$, indicates the reliability of positive predictions, though a high precision value may co-occur with an extremely low *TPR* if many members are missed. The *F1 score* balances precision and recall, yet it can mask the individual trade-offs between the two. *Membership*

advantage (MA) is the difference between *TPR* and *FPR* and quantifies the improvement over random guessing. However, high *MA* could occur with non-negligible *FPR*, and it also can overestimate privacy risks under imbalanced priors [37].

Given their widespread use in privacy auditing, MIAs must be evaluated based on their ability to reliably detect membership. Recent state-of-the-art works [76,12,8,79] focus on measuring *TPR* at extremely low *FPR* (FPR $\leq 1\%$) to ensure the reliability of positive detections. However, this approach overlooks the typically low prior probability of membership, since the training set often represents a small subset of the overall population. As noted in [37], traditional precision does not account for this realistic imbalance. For example, during an epidemic, the training set may consist of hospitalized patients with symptoms, while the non-member population comprises the broader city. To address this issue, they proposed a weighted precision metric that incorporates the prior membership probability, p, as follows:

$$\text{Prec} = \frac{p \times \text{TPR}}{p \times \text{TPR} + (1-p) \times \text{FPR}}, \tag{4}$$

where $p \ll 50\%$ in real-world scenarios.

References

1. Hepatitis. UCI Machine Learning Repository (1983). https://doi.org/10.24432/C5Q59J
2. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318 (2016)
3. Antonio, B., Moroni, D., Martinelli, M.: Efficient adaptive ensembling for image classification. Expert Systems **42**(1), e13424 (2025)
4. Arcos-García, Á., Álvarez-García, J.A., Soria-Morillo, L.M.: Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods. Neural Networks **99**, 158–165 (2018)
5. Aubinais, E., Gassiat, E., Piantanida, P.: Fundamental limits of membership inference attacks on machine learning models. Journal of Machine Learning Research **26**(263), 1–54 (2025)
6. Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential privacy has disparate impact on model accuracy. In: Advances in Neural Information Processing Systems 32 (NeurIPS) (2019)
7. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996). https://doi.org/10.24432/C5XW20
8. Bertran, M., Tang, S., Kearns, M., Morgenstern, J., Roth, A., Wu, Z.S.: Scalable membership inference attacks via quantile regression. In: Advances in Neural Information Processing Systems (2023)
9. Bindschaedler, V., Shokri, R., Gunter, C.A.: Plausible deniability for privacy-preserving data synthesis. Proceedings of the VLDB Endowment **10**(5), 481–492 (2017)

10. Blanco-Justicia, A., Sánchez, D., Domingo-Ferrer, J., Muralidhar, K.: A critical review on the use (and misuse) of differential privacy in machine learning. ACM Computing Surveys **55**(8), 1–16 (2022)
11. Byerly, A., Kalganova, T., Dear, I.: No routing needed between capsules. Neurocomputing **463**, 545–553 (2021)
12. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F.: Membership inference attacks from first principles. In: IEEE Symposium on Security and Privacy (SP). pp. 1897–1914 (2022)
13. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C.: Extracting training data from large language models. In: 30th USENIX Security Symposium. pp. 2633–2650 (2021)
14. Chakrabarty, N., Biswas, S.: A statistical approach to adult census income level prediction. In: International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). pp. 207–212 (2018)
15. Choquette-Choo, C.A., Tramèr, F., Carlini, N., Papernot, N.: Label-only membership inference attacks. In: Proceedings of the 38th International Conference on Machine Learning. pp. 1964–1974 (2021)
16. Darlow, L.N., Crowley, E.J., Antoniou, A., Storkey, A.J.: CINIC-10 is not ImageNet or CIFAR-10. arXiv:1810.03505 (2018)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
18. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694 (2019)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
20. Dionysiou, A., Athanasopoulos, E.: SoK: Membership inference is harder than previously thought. Proceedings on Privacy Enhancing Technologies **2023**(3), 286–306 (2023)
21. Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J.: Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections. Morgan & Claypool Publishers (2016)
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=YicbFdNTTy
23. Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., Hajishirzi, H.: Do membership inference attacks work on large language models? In: Proceedings of the First Conference on Language Modeling (COLM) (2024), https://openreview.net/forum?id=av0D19pSkU
24. Dwork, C.: Differential privacy. In: Automata, Languages and Programming (ICALP). Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer (2006)
25. European Parliament, Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection

of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://data.europa.eu/eli/reg/2016/679/oj (2016)
26. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: Proceedings of the 9th International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=6Tm1mposlrM
27. Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., Jiang, T.: Membership inference attacks against fine-tuned large language models via self-prompt calibration. In: Advances in Neural Information Processing Systems 37 (NeurIPS) (2024)
28. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
29. He, Y., Li, B., Wang, Y., Yang, M., Wang, J., Hu, H., Zhao, X.: Is difficulty calibration all we need? towards more practical membership inference attacks. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. pp. 1226–1240 (2024)
30. Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994). https://doi.org/10.24432/C5NC77
31. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys **54**(11s) (2022)
32. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (2007)
33. Humphries, T., Oya, S., Tulloch, L., Rafuse, M., Goldberg, I., Hengartner, U., Kerschbaum, F.: Investigating membership inference attacks under data dependencies. In: Proceedings of the 36th IEEE Computer Security Foundations Symposium (CSF). pp. 473–488 (2023)
34. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P.: Statistical Disclosure Control. Wiley (2012)
35. Jacobs, R.A.: Increased rates of convergence through learning rate adaptation. Neural Networks **1**(4), 295–307 (1988)
36. Jayaraman, B., Evans, D.: Evaluating differentially private machine learning in practice. In: 28th USENIX Security Symposium. pp. 1895–1912 (2019)
37. Jayaraman, B., Wang, L., Knipmeyer, K., Gu, Q., Evans, D.: Revisiting membership inference under realistic assumptions. Proceedings on Privacy Enhancing Technologies **2021**(2), 348–368 (2021)
38. Jebreel, N., Khalil, M., Sánchez, D., Domingo-Ferrer, J.: Revisiting the LiRA membership inference attack under realistic assumptions. arXiv:2603.07567 (2026)
39. Johnson, R., Zhang, T.: Supervised and semi-supervised text categorization using LSTM for region embeddings. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). pp. 526–534 (2016)
40. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
41. Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. In: Advances in Neural Information Processing Systems 4 (NIPS). pp. 950–957 (1991)
42. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
43. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5**, 361–397 (2004)

44. Li, J., Li, N., Ribeiro, B.: Membership inference attacks and defenses in classification models. In: Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy. pp. 5–16 (2021)
45. Li, J., Li, N., Ribeiro, B.: MIST: Defending against membership inference attacks through Membership-Invariant subspace training. In: 33rd USENIX Security Symposium. pp. 2387–2404 (2024)
46. Lin, Y., Meng, Y., Sun, X., Han, Q., Kuang, K., Li, J., Wu, F.: BertGCN: Transductive text classification by combining GNN and BERT. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP. pp. 1456–1462 (2021)
47. Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: A survey and outlook. ACM Computing Surveys **54**(2), 31:1–31:36 (2022)
48. Liu, Y., Zhao, Z., Backes, M., Zhang, Y.: Membership inference attacks by exploiting loss trajectory. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. pp. 2085–2098 (2022)
49. Long, Y., Bindschaedler, V., Wang, L., Bu, D., Wang, X., Tang, H., Gunter, C.A., Chen, K.: Understanding membership inferences on well-generalized learning models. arXiv:1802.04889 (2018)
50. Long, Y., Wang, L., Bu, D., Bindschaedler, V., Wang, X., Tang, H., Gunter, C.A., Chen, K.: A pragmatic approach to membership inferences on machine learning models. In: IEEE European Symposium on Security and Privacy (EuroS&P). pp. 521–534 (2020)
51. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: ℓ-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data **1**(1), 3–es (2007)
52. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics **19**(6), 1236–1246 (2018)
53. Murakonda, S.K., Shokri, R.: ML privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. In: Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs) (2020)
54. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: IEEE Symposium on Security and Privacy (SP). pp. 739–753 (2019)
55. Nikzad, N., Liao, Y., Gao, Y., Zhou, J.: SATA: Spatial autocorrelation token analysis for enhancing the robustness of vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9730–9739 (2025)
56. Pilgram, L., et al.: A consensus privacy metrics framework for synthetic data. Patterns **6**(10), 101320 (2025)
57. Prechelt, L.: Early stopping — but when? In: Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol. 1524, pp. 55–69. Springer (1998)
58. Rahman, M.A., Rahman, T., Laganière, R., Mohammed, N., Wang, Y.: Membership inference attack against differentially private deep learning model. Transactions on Data Privacy **11**(1), 61–79 (2018)
59. Rezaei, S., Liu, X.: On the difficulty of membership inference attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7892–7900 (2021)
60. Rezaei, S., Liu, X.: On the discredibility of membership inference attacks. arXiv:2212.02701 (2022)

61. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
62. Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., Jegou, H.: White-box vs black-box: Bayes optimal strategies for membership inference. In: Proceedings of the 36th International Conference on Machine Learning. pp. 5558–5567 (2019)
63. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: 26th Annual Network and Distributed System Security Symposium (NDSS) (2019)
64. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press (2014)
65. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy (SP). pp. 3–18 (2017)
66. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: 30th USENIX Security Symposium. pp. 2615–2632 (2021)
67. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1), 1929–1958 (2014)
68. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The German traffic sign recognition benchmark: A multi-class classification competition. In: International Joint Conference on Neural Networks (IJCNN). pp. 1453–1460 (2011)
69. Suri, A., Zhang, X., Evans, D.: Do parameters reveal more than loss for membership inference? Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=fmKJfbGKFC
70. Tabassi, E., Burns, K.J., Hadjimichael, M., Molina-Markham, A.D., Sexton, J.T.: A taxonomy and terminology of adversarial machine learning. Tech. Rep. NISTIR 8269, National Institute of Standards and Technology (2019)
71. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing **14**(6), 2073–2089 (2021)
72. van Dyk, D.A., Meng, X.L.: The art of data augmentation. Journal of Computational and Graphical Statistics **10**(1), 1–50 (2001)
73. Watson, L., Guo, C., Cormode, G., Sablayrolles, A.: On the importance of difficulty calibration in membership inference attacks. In: International Conference on Learning Representations (ICLR) (2022), https://openreview.net/forum?id=3eIrli0TwQ
74. Wolberg, W.: Breast Cancer Wisconsin (Original). UCI Machine Learning Repository (1992). https://doi.org/10.24432/C5HP4Z
75. Xu, R., Baracaldo, N., Joshi, J.: Privacy-preserving machine learning: Methods, challenges and directions. arXiv:2108.04417 (2021)
76. Ye, J., Maddi, A., Murakonda, S.K., Bindschaedler, V., Shokri, R.: Enhanced membership inference attacks against machine learning models. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. pp. 3093–3106 (2022)
77. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: Analyzing the connection to overfitting. In: IEEE 31st Computer Security Foundations Symposium (CSF). pp. 268–282 (2018)

78. Ying, Z., Zhang, Y., Liu, X.: Privacy-preserving in defending against membership inference attacks. In: Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice. pp. 61–63 (2020)
79. Zarifzadeh, S., Liu, P., Shokri, R.: Low-cost high-power membership inference attacks. In: Proceedings of the 41st International Conference on Machine Learning. pp. 58244–58282 (2024)
80. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. pp. 13001–13008 (2020)