Supplementary Material for: 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, Tarragona, Catalonia
{najeeb.jebreel, david.sanchez, josep.domingo}@urv.cat

Note. This document provides supplementary material referenced by the main manuscript. It is intended to be submitted as a separate PDF alongside the paper.

S1 Surveyed Attacks

This section provides the detailed survey referenced in the main manuscript. In particular, Table S1 summarizes the representative black-box membership inference attacks (MIAs) reviewed in the paper, including their core ideas, publication venues, and citation counts.

Table S1: Summary of the surveyed attacks. Citation counts according to Google Scholar, January 8, 2026.

Attack	Approach	Venue	# citations
[26]	ML membership classifier on predictions from shadow models	IEEE SP 2017	7,083
[32]	Loss global thresholding	IEEE CSF 2018	1,751
[25]	Confidence and entropy global thresholding	NDSS 2019	1,329
[24]	Per-sample loss calibration and thresholding	ICML 2019	515
[19]	Hypothesis testing on loss values of selected vulnerable records	Euro SP 2020	143
[13]	Perturbed input loss thresholding	PoPETs 2021	194
[27]	Class-specific modified entropy thresholding	USENIX Security 2021	574
[18]	ML membership classifier on the sample's loss of trajectory from distilled models	ACM CCS 2022	165
[30]	Per-sample loss calibration and thresholding	ICLR 2022	203
[31]	Hypothesis testing on loss values from reference/distilled models	ACM CCS 2022	429
[4]	Hypothesis testing based on likelihood ratio of scores from shadow models	IEEE SP 2022	1180
[2]	Quantile regression on confidence scores	NeurIPS 2024	81
[34]	Hypothesis testing based on likelihood ratio of scores from shadow models	ICML 2024	91

The first notable MIA against DNN models was presented in [26], 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.

[25] relax some assumptions made by [26] 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.

[32] 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 [32,25] 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.

[27] 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.

[13] 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.

[19] 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 differentiate between them. To address this issue, [18] 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.

[24] 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, [30] 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. [30] demonstrate that this approach can improve the inference performance of several existing attacks, including those based on loss, gradient norm, and confidence scores.

[31] 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.

[4] introduced the Likelihood Ratio Attack (LiRA), which improves over [24,31] 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 [24,31,30,4] in achieving high TPR at low FPR, their computational costs render them unscalable for practical privacy auditing [34].

Recent methods by [2] and [34] aim to address the scalability limitations of LiRA and related attacks. [34] demonstrated that, under practical computational budgets (e.g., two shadow models), LiRA's performance degrades to near-random guessing, while Attack-R [31] shows low true-positive rates (TPR) at low FPR.

[2] 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.

[34] 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.

S2 Background

S2.1 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 [8]. 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) [15] 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.

S2.2 Differential Privacy

Differential privacy (DP) [7] 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

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta, \tag{S1}$$

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 [6] and privacy-preserving ML [17]. 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 [32].

S2.3 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 [1]. Many works have enforced DP on DNN models to mitigate MIAs [32,33,5,10,12,13,16,22]. 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 [12,3].

Anti-overfitting techniques, on the other hand, provide empirical privacy protection (without formal guarantees) while better retaining (or even improving) model utility [3,4]. These defenses include methods such as weight regularization [14], weight dropout [28], data augmentation [29,35], learning rate tuning [11], and early stopping [21].

S2.4 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 [26], 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 [20,9]. 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 [24], 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 [32,13,4].

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:

$$\mathcal{A}(x, y) = \mathbf{1}[\mathcal{A}'(x, y) > \tau]. \tag{S2}$$

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 [23,4,30]. Similarly, optimizing for accuracy can inadvertently inflate FPR, thereby compromising the reliable detection of membership [30].

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 [13].

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 [31,4,2,34] focus on measuring TPR at extremely low FPR ($\text{FPR} \leq 1\%$) 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 [13], 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:

$$\text{Prec} = \frac{p \times \text{TPR}}{p \times \text{TPR} + (1-p) \times \text{FPR}}, \tag{S3}$$

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

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016

ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)
2. Bertran, M., Tang, S., Roth, A., Kearns, M., Morgenstern, J.H., Wu, S.Z.: Scalable membership inference attacks via quantile regression. Advances in Neural Information Processing Systems **36** (2024)
3. 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)
4. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 1897–1914. IEEE (2022)
5. Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N.: Label-only membership inference attacks. In: International conference on machine learning. pp. 1964–1974. PMLR (2021)
6. Domingo-Ferrer, J., Sánchez, D., Soria-Comas, J.: Database anonymization: privacy models, data utility and microaggregation-based inter-model connections. Synthesis Lectures on Information Security and Privacy. Morgan & Claypool Publishers (2016)
7. Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep feedforward networks. Deep learning (1) (2016)
9. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR) **54**(11s), 1–37 (2022)
10. Humphries, T., Rafuse, M., Tulloch, L., Oya, S., Goldberg, I., Kerschbaum, F.: Differentially private learning does not bound membership inference. arXiv preprint arXiv:2010.12112 (2020)
11. Jacobs, R.A.: Increased rates of convergence through learning rate adaptation. Neural networks **1**(4), 295–307 (1988)
12. Jayaraman, B., Evans, D.: Evaluating differentially private machine learning in practice. In: 28th USENIX Security Symposium (USENIX Security 19). pp. 1895–1912 (2019)
13. Jayaraman, B., Wang, L., Knipmeyer, K., Gu, Q., Evans, D.: Revisiting membership inference under realistic assumptions. Proceedings on Privacy Enhancing Technologies **2021**(2) (2021)
14. Krogh, A., Hertz, J.: A simple weight decay can improve generalization. Advances in neural information processing systems **4** (1991)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
16. 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)
17. Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)
18. Liu, Y., Zhao, Z., Backes, M., Zhang, Y.: Membership inference attacks by exploiting loss trajectory. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. pp. 2085–2098 (2022)

19. 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: 2020 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 521–534. IEEE (2020)
20. 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: 2019 IEEE symposium on security and privacy (SP). pp. 739–753. IEEE (2019)
21. Prechelt, L.: Early stopping-but when? In: Neural Networks: Tricks of the trade, pp. 55–69. Springer (2002)
22. Rahman, M.A., Rahman, T., Laganière, R., Mohammed, N., Wang, Y.: Membership inference attack against differentially private deep learning model. Trans. Data Priv. **11**(1), 61–79 (2018)
23. Rezaei, S., Liu, X.: On the difficulty of membership inference attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7892–7900 (2021)
24. Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., Jégou, H.: White-box vs black-box: Bayes optimal strategies for membership inference. In: International Conference on Machine Learning. pp. 5558–5567. PMLR (2019)
25. Salem, A., Zhang, Y., Humbert, M., Fritz, M., Backes, M.: Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: Network and Distributed Systems Security Symposium 2019. Internet Society (2019)
26. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). pp. 3–18. IEEE (2017)
27. Song, L., Mittal, P.: Systematic evaluation of privacy risks of machine learning models. In: 30th USENIX Security Symposium (USENIX Security 21). pp. 2615–2632 (2021)
28. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
29. Van Dyk, D.A., Meng, X.L.: The art of data augmentation. Journal of Computational and Graphical Statistics **10**(1), 1–50 (2001)
30. Watson, L., Guo, C., Cormode, G., Sablayrolles, A.: On the importance of difficulty calibration in membership inference attacks. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=3eIrli0TwQ
31. Ye, J., Maddi, A., Murakonda, S.K., Bindschaedler, V., Shokri, R.: Enhanced membership inference attacks against machine learning models. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. pp. 3093–3106 (2022)
32. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy risk in machine learning: Analyzing the connection to overfitting. In: 2018 IEEE 31st computer security foundations symposium (CSF). pp. 268–282. IEEE (2018)
33. 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)
34. Zarifzadeh, S., Liu, P., Shokri, R.: Low-cost high-power membership inference attacks. In: Forty-first International Conference on Machine Learning (2024)

35. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020)