Article

Empirical Evaluation of Membership Inference Attacks Against Differentially Private Deep Learning Models

Yuzhou Zhang[1,*]

[1] Engineering University of PAP, Xi'an, China
* Correspondence: Yuzhou Zhang, Engineering University of PAP, Xi'an, China

Abstract: Membership inference evaluations (MIEs) determine whether a sample was used to train a machine-learning model and therefore provide an empirical measure of training-data leakage. This study evaluates black-box MIEs against non-private and differentially private convolutional neural networks trained on MNIST and CIFAR-10. Twenty-five shadow models were used to generate member and non-member prediction vectors, and class-specific LightGBM classifiers served as the inference models. Differentially private stochastic gradient descent (DP-SGD) was evaluated at privacy-budget settings of $\varepsilon = 1$, 2, 4, and 8. Across the evaluated configurations, DP-SGD reduced inference accuracy, precision, and recall relative to conventional SGD. On MNIST, inference accuracy decreased from 0.944 for the non-private model to 0.536 at $\varepsilon = 1$. On CIFAR-10, inference accuracy was 0.761 for the non-private model and 0.659–0.679 at $\varepsilon = 2$–8; the value of 0.370 at $\varepsilon = 1$ indicated substantial degradation of the inference classifier under the strongest privacy setting. These findings demonstrate the practical value of membership inference as an empirical privacy-testing tool and show that the effect of DP-SGD varies across datasets and privacy configurations.

Keywords: membership inference evaluation; differential privacy; deep learning; DP-SGD; privacy budget

1. Introduction

Deep learning models are often trained on records that contain identity, health, behavioral, or other sensitive information. Removing direct identifiers does not eliminate every privacy risk. An adversary may still infer whether a specific record was included in the training set by examining the model's outputs. Such membership information can reveal a person's association with a sensitive dataset, even when the underlying record is not reconstructed.

Shokri et al. formulated membership inference as a binary classification problem and introduced a practical black-box attack based on shadow models [1]. The attacker trains models that imitate the target model, collects their outputs for member and non-member samples, and learns an attack classifier from the resulting differences. Subsequent studies relaxed the assumptions required by the original attack, introduced white-box attacks, and developed likelihood-ratio methods that provide substantially stronger evaluation at low false-positive rates [2,3,4].

Differential privacy (DP) offers a formal framework for limiting the influence of a single training record. Differentially private stochastic gradient descent (DP-SGD) combines per-example gradient clipping with calibrated Gaussian noise and tracks cumulative privacy loss using a privacy accountant [5]. Formal guarantees, however, depend on the complete training configuration, including the clipping norm, noise multiplier, sampling rate, number of optimization steps, δ, and the accountant. Empirical attacks therefore complement, but do not replace, formal privacy accounting [6,7].

This study provides a compact empirical comparison of black-box membership inference across two image-classification tasks. A common shadow-model and LightGBM

Received: 15 August 2026
Revised: 09 October 2026
Accepted: 23 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

pipeline is applied to MNIST and CIFAR-10 under non-private training and four DP-SGD privacy-budget settings. The study focuses on whether the observed attack metrics decrease under DP-SGD and on how strong attack degradation should be interpreted before a model is deployed.

The contribution is threefold. First, the same attack construction is used across both datasets, enabling a consistent comparison of membership leakage under different privacy settings. Second, the study examines accuracy, precision, and recall jointly and relates their changes to the privacy budget and dataset characteristics. Third, the analysis provides a deployment-oriented interpretation framework for understanding strong attack degradation and cross-dataset variability. The paper is positioned as an applied empirical evaluation of membership inference against differentially private image classifiers.

2. Background and Related Work

2.1. Differential Privacy for Deep Learning

A randomized mechanism M satisfies (ε, δ)-differential privacy if, for every pair of adjacent datasets D and D′ that differ in one record and for every measurable output set S,

$$Pr[M(D) \in S] \leq e\hat{}\varepsilon \, Pr[M(D') \in S] + \delta \qquad (1)$$

where ε bounds the multiplicative change in output probabilities and δ permits a small probability of additional privacy loss [8]. Smaller ε generally indicates a stronger privacy constraint when δ and the remaining training conditions are fixed. In deep learning, DP-SGD enforces a bound on the contribution of each training example and adds random noise to the aggregate update [5]. A privacy accountant then accumulates the privacy cost over the training process.

2.2. Membership Inference Attacks

A black-box attacker can query a target model and observe its predicted label, confidence scores, or probability vector, but cannot access model parameters or gradients. A white-box attacker may additionally exploit parameters, gradients, or intermediate representations. The original shadow-model attack trains auxiliary models on data drawn from a distribution similar to that of the target data and uses the auxiliary outputs to learn the distinction between training and non-training samples [1]. Later work demonstrated attacks with fewer assumptions [3], attacks that exploit white-box information [4], and likelihood-ratio attacks designed for evaluation at low false-positive rates [2]. MIAs have also been used to audit the implementation of differentially private learning algorithms [7].

2.3. Empirical Privacy Evaluation

Membership inference is closely related to the extent to which a model behaves differently on its training data and on unseen data. Yeom et al. connected membership risk with overfitting, while later systematic evaluations showed that the measured risk also depends strongly on the attack features, decision rule, reference data, and evaluation protocol [9,10]. Consequently, a high attack score may indicate exploitable memorization, but a low score only demonstrates that a particular attacker failed under a particular set of assumptions. It is not a proof that membership information is absent.

Average accuracy is intuitive but can hide operationally important behavior. In many privacy settings, even a small number of highly confident member predictions is harmful; true-positive rate at a low false-positive rate can therefore be more informative than average accuracy [2]. Label-only attacks further show that privacy risk may persist when confidence scores are unavailable [11]. Practical evaluation should thus state the adversary's observations, confirm the member base rate, report threshold-independent metrics when possible, and distinguish attack failure from a formal privacy guarantee.

Formal accounting and empirical auditing answer different questions. Accounting bounds worst-case privacy loss under explicit assumptions about sampling, clipping,

noise, and composition. Auditing tests whether an implementation exhibits leakage detectable by a chosen adversary and can reveal configuration or software errors [7,12,13]. A sound assessment uses both: the accountant provides the claimed guarantee, whereas attacks provide an implementation-level stress test.

3. Attack and Defense Framework

3.1. Threat Model

Let f_target denote a classifier trained on the private dataset D_target. The adversary submits a sample x to the model and receives the class-probability vector $\hat{y}$. The adversary does not read model parameters. Given $\hat{y}$ and the true class label y, the attack model f_attack($\hat{y}$, y) predicts whether x belongs to D_target. The attacker is assumed to know the target architecture and the data distribution and to have access to public data drawn from the same distribution (As shown in Figure 1).

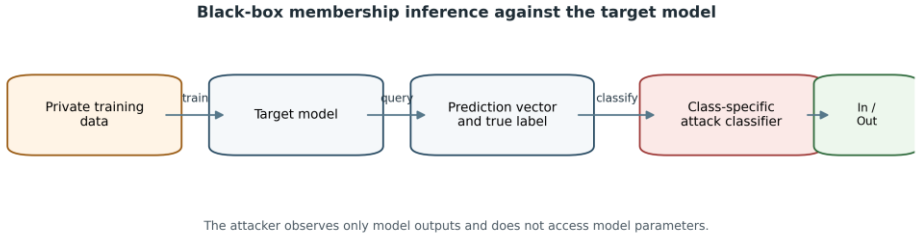


Figure 1. Black-box Membership Inference Framework, Conceptually Adapted from Shokri Et Al. [1].

3.2. Shadow Models and Attack Classifiers

The experiment used 25 shadow models with the same architecture as the target model. For each shadow model, outputs for its training samples were labeled in, whereas outputs for a disjoint test subset were labeled out. The confidence vectors were grouped by the true class, and a separate LightGBM binary classifier was trained for each class. During evaluation, the output of the target model was passed to the attack classifier corresponding to the sample's true class, which returned a member or non-member prediction (As shown in Figure 2).

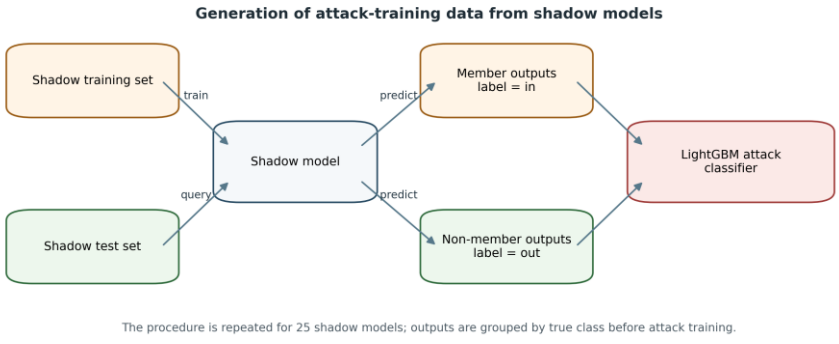


Figure 2. Generation of Attack-Training Data Using Shadow Models, Conceptually Adapted from [1].

3.3. Dp-sgd Defense

For each training sample x_i, DP-SGD computes the gradient g_t(x_i) at optimization step t and clips its l2 norm to the threshold C:

$$\bar{g}_t(x_i) = g_t(x_i) / \max(1, \|g_t(x_i)\|_2 / C) \quad (2)$$

The clipped gradients are aggregated, Gaussian noise is added, and the model parameters are updated according to

$$\tilde{g}_t = \frac{1}{L}\left[\sum_i \bar{g}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 I)\right], \quad \theta_{t+1} = \theta_t - \eta_t \tilde{g}_t \quad (3)$$

where L is the batch size, σ is the noise multiplier, I is the identity matrix, and η_t is the learning rate. The models were trained with Opacus at ε settings of 1, 2, 4, and 8, while a conventional SGD model served as the non-private control. These privacy-budget settings provide a consistent basis for comparing attack behavior as the strength of differential privacy changes.

3.4. Evaluation Logic

For a balanced member/non-member test set, random guessing has an expected accuracy of 0.5. Values above 0.5 indicate that the attack has learned a useful decision rule. A value near 0.5 indicates weak discrimination, whereas a value materially below 0.5 requires special care: reversing every prediction would produce an accuracy above 0.5. Such a result may be caused by an inverted in/out label convention, a shifted score direction, unbalanced classes, or unstable fitting. It should not be converted automatically into a privacy claim. Precision and recall provide complementary information, but both depend on the decision threshold and the prevalence of members in the evaluated sample.

The comparison therefore uses the non-private model as an empirical reference and asks whether the same attack becomes less successful under DP-SGD. Because model optimization and attack training are stochastic, small differences between neighboring ε settings are interpreted descriptively rather than as evidence of a universal monotonic relationship.

4. Experimental Setup

4.1. Datasets and Model Architectures

MNIST contains 28 × 28 grayscale images of handwritten digits, and CIFAR-10 contains 32 × 32 RGB natural images. Both datasets comprise ten classes. For the target model and for each shadow model, 2,500 training images and 1,000 test images were sampled. Within each model, the training and test subsets were disjoint (As shown in Table 1).

Table 1. Architectures Used for the Target and Shadow Models.

Dataset	Convolution 1	Convolution 2	Pooling	Fully connected layers
MNIST	5 × 5, 1 → 20	5 × 5, 20 → 50	2 × 2	Flatten → 128 → 10
CIFAR-10	5 × 5, 3 → 6	5 × 5, 6 → 16	2 × 2	Flatten → 128 → 10

Both architectures used two convolutional layers followed by 2 × 2 pooling and two fully connected layers. The exact flattened feature dimension is determined by the tensor shape produced by the implemented convolution and pooling operations.

4.2. Training Configuration and Evaluation Metrics

Table 2. Main Experimental Settings.

Parameter	Setting
Number of shadow models	25
Batch size	64
Learning rate	0.001
SGD momentum	0.5
Learning-rate decay	1×10^{-7}
Maximum epochs	100
Attack classifier	LightGBM

The target models were optimized using cross-entropy loss. The attack was evaluated as a binary classification task using accuracy, precision, and recall. Precision was calculated as TP/(TP + FP), and recall as TP/(TP + FN), where TP and FP denote true and false member predictions, respectively, and FN denotes members classified as non-members. Larger values indicate greater success in identifying training samples. These metrics should be interpreted together with the member-to-non-member ratio because accuracy is sensitive to class imbalance.

4.3. Implementation Details

For each target or shadow model, the training subset contained 2,500 images and the disjoint test subset contained 1,000 images. Twenty-five shadow models were used to generate member and non-member outputs for the attack classifiers. The target and corresponding shadow models used the same network architecture and input format, and attack samples were grouped by their true class before LightGBM training. These choices maintain a consistent attack protocol across privacy settings and datasets (As shown in Table 2).

The implementation combines publicly available benchmark data, standard convolutional networks, LightGBM attack classifiers, and the Opacus differential privacy library [14]. This modular design allows the same evaluation workflow to be extended to other model architectures, datasets, and privacy configurations.

5. Results and Discussion

5.1. Membership Inference Performance

Table 3 reports accuracy, precision, and recall for the attack model. Accuracy represents the overall proportion of correctly classified member and non-member samples. Precision is the proportion of predicted members that are true members, whereas recall is the proportion of true members recovered by the attack. High values across these metrics indicate that the target model exposes output patterns that separate training samples from non-training samples.

Table 3. Membership Inference Performance under Different Privacy Settings.

Dataset and metric	ε = 1	ε = 2	ε = 4	ε = 8	Non-private
MNIST accuracy	0.536	0.629	0.656	0.740	0.944
CIFAR-10 accuracy	0.370	0.659	0.663	0.679	0.761
MNIST precision	0.482	0.523	0.546	0.675	0.937
CIFAR-10 precision	0.306	0.505	0.512	0.558	0.713
MNIST recall	0.479	0.520	0.537	0.653	0.923
CIFAR-10 recall	0.277	0.503	0.506	0.537	0.644

For MNIST, the non-private model produced an attack accuracy of 0.944, a precision of 0.937, and a recall of 0.923. These values indicate that the attack classifier could identify training samples with high reliability. DP-SGD reduced all three metrics, and the overall trend became stronger as ε decreased. At ε = 1, attack accuracy fell to 0.536, with precision and recall of 0.482 and 0.479. Under a balanced evaluation protocol, these values are close to random guessing, suggesting that the added noise substantially weakened the output differences exploited by the attack.

For CIFAR-10, attack accuracy against the non-private model was 0.761, lower than the corresponding MNIST value. This difference may reflect the combined effects of the dataset, architecture, and model-fitting behavior rather than dataset complexity alone. At ε = 8, 4, and 2, attack accuracy was 0.679, 0.663, and 0.659, respectively. The small differences among these three settings indicate broadly similar attack performance and do not support a strictly monotonic relationship between ε and attack success.

At ε = 1 on CIFAR-10, attack accuracy decreased to 0.370, while precision and recall were 0.306 and 0.277. The simultaneous decline of all three metrics indicates that the attack classifier was strongly degraded under this privacy setting and could no longer identify

positive member samples reliably. The below-chance accuracy also suggests that the learned decision boundary became unstable in the strong-noise regime. Accordingly, this value is interpreted as failure of the current attack classifier rather than as a directly quantifiable privacy-risk level.

Overall, the attack metrics under DP-SGD were lower than those for the non-private models in the reported runs. This pattern is consistent with the intended effect of gradient clipping and noise addition: the model becomes less sensitive to individual training samples, leaving fewer membership-specific patterns in its confidence scores. The magnitude of the reduction, however, depends on the task and on the evaluation protocol. A deployment decision should therefore combine formal privacy accounting, target-model utility, attack evaluation at relevant operating points, and repeated trials rather than rely on a single accuracy value.

5.2. Cross-Dataset Interpretation

The two datasets show different baseline leakage patterns. The non-private MNIST attack accuracy of 0.944 is substantially higher than the CIFAR-10 value of 0.761. A plausible explanation is that the MNIST target model separated its training and non-training examples more sharply in confidence space. This pattern may reflect differences in calibration, architecture capacity, optimization behavior, or the effective diversity of the selected training subsets. The observed difference is therefore best interpreted as a dataset-specific outcome of the complete learning and attack pipeline rather than as a consequence of dataset complexity alone.

The MNIST values also exhibit a clearer ordering across the reported ε settings than the CIFAR-10 values. A target ε summarizes the intended privacy constraint, whereas the observed attack score is additionally affected by optimization, sampling, calibration, and attack-model behavior. The small differences among the CIFAR-10 results at $\varepsilon = 2$, 4, and 8 indicate broadly similar attack performance across these configurations. The overall comparative pattern remains clear: every evaluated DP-SGD configuration reduced the selected attack metrics relative to its non-private counterpart, although the relationship between ε and empirical leakage was not strictly linear.

5.3. Implications for Pre-Deployment Testing

The results suggest a practical four-stage assessment process. First, the evaluator should validate the attack dataset: member and non-member labels, class proportions, split independence, and preprocessing must be checked before interpreting the score. Second, the DP implementation should be documented with δ, clipping norm, noise multiplier, sampling probability, number of steps, accountant, and achieved ε. Third, several attack views should be compared, including accuracy, precision, recall, ROC-AUC, and, where the application is sensitive to rare false alarms, true-positive rate at a low false-positive rate. Finally, attack results should be read together with target-model utility and uncertainty across random seeds.

This process prevents two common errors. The first is to equate a failed attack with proof of privacy: an attacker may fail because its assumptions or features are weak. The second is to select the smallest ε without considering whether the resulting model remains useful. In deployment, privacy is a constrained design problem. A candidate configuration is acceptable when its formal accounting, empirical leakage, predictive utility, and operational requirements are jointly understood. Together, these components convert membership inference from a standalone metric into a structured deployment assessment.

5.4. Scope and Future Work

The scope of this study is an empirical comparison using two benchmark datasets, compact convolutional networks, and a classical shadow-model attack. The reported comparison reflects one experimental realization and focuses on attack accuracy, precision, and recall. Future work will extend the same framework to larger architectures and application datasets, incorporate repeated trials and target-model utility, and report

complete privacy-accounting records. ROC-AUC and true-positive rates at low false-positive rates will also provide useful complementary views of attack performance [2].

6. Conclusion

This study examined whether DP-SGD weakens a classical black-box membership inference attack based on shadow models and class-specific LightGBM classifiers. Under the evaluated configurations, the attack achieved its highest accuracy, precision, and recall against the non-private MNIST and CIFAR-10 models. DP-SGD consistently reduced the separability of member and non-member outputs. The reduction was especially clear for MNIST, while CIFAR-10 showed similar attack performance at $\varepsilon = 2$, 4, and 8 and pronounced attack degradation at $\varepsilon = 1$.

These results demonstrate that membership inference can serve as a practical pre-deployment stress test for differentially private models. They also show that the empirical effect of a privacy setting depends on the complete interaction among the dataset, target model, and attack classifier rather than on ε alone. The proposed workflow provides an applied baseline for comparing privacy configurations and identifying models that expose distinguishable training-sample behavior. Future work will extend this evaluation to broader architectures, application datasets, and additional operating-point metrics.

Data Availability: MNIST and CIFAR-10 are publicly available benchmark datasets. No new dataset was created for this study. The numerical results analyzed in this article are reported in Table 3.

References

1. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy*, 2017, pp. 3–18.
2. N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy*, 2022, pp. 1897–1914.
3. A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, 2019.
4. M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE Symposium on Security and Privacy*, 2019, pp. 739–753.
5. M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.
6. M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model," *Trans. Data Privacy*, vol. 11, no. 1, pp. 61–79, 2018.
7. M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private SGD?" in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 22205–22216.
8. C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science, vol. 4978, 2008, pp. 1–19.
9. S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium*, 2018, pp. 268–282.
10. L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *30th USENIX Security Symposium*, 2021, pp. 2615–2632.
11. C. A. Choquette-Choo, F. Tramèr, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 1964–1974.
12. B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *28th USENIX Security Symposium*, 2019, pp. 1895–1912.
13. M. Nasr, J. Hayes, T. Steinke, B. Balle, F. Tramèr, M. Jagielski, and N. Carlini, "Tight auditing of differentially private machine learning," in *32nd USENIX Security Symposium*, 2023, pp. 1637–1654.
14. A. Yousefpour et al., "Opacus: User-friendly differential privacy library in PyTorch," arXiv:2109.12298, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.