Optimal Noise Injection on Training Data: A Defense Against Membership Inference Attacks

Radia Kassa[1,2], Kamel Adi[2] and Myria Bouhaddi[2]
[1]*Laboratoire LITAN, École supérieure en Sciences et Technologies de l'Informatique et du Numérique, RN 75, Amizour 06300, Bejaia, Algeria*
[2]*Computer Security Research Laboratory, University of Quebec in Outaouais, Gatineau, Quebec, Canada*

Keywords: Membership Inference Attacks, Data Privacy, Machine Learning, Defense Mechanism, Optimal Noise Injection, Prediction Entropy, Black-Box Defense, Optimized Noise, Shapley Values.

Abstract: Membership inference attacks (MIAs) present a serious risk to data privacy in machine learning (ML) models, as they allow attackers to determine whether a given data point was included in the training set. Although various defenses exist, they often struggle to effectively balance privacy and utility. To address this challenge, we propose in this paper a novel defense mechanism based on Optimal Noise Injection during the training phase. Our approach involves injecting a carefully designed and controlled noise vector into each training sample. This optimization maximizes prediction entropy to obscure membership signals while leveraging Shapley values to preserve data utility. Experiments on benchmark datasets show that our method reduces MIA success rates significantly without sacrificing accuracy, offering a strong privacy-utility trade-off for black-box scenarios.

1 INTRODUCTION

Deep learning has significantly transformed the field of artificial intelligence, achieving remarkable performance in tasks such as autonomous driving, natural language processing, and medical diagnostics. Its ability to automatically extract complex features from large datasets has made it essential for sensitive applications in areas such as edge computing, finance, and healthcare. However, this success relies on access to vast amounts of often sensitive or private data, exposing these models to considerable privacy risks. Indeed, many studies have shown that Machine Learning (ML) models can inadvertently memorize sensitive information from training data, primarily due to overfitting. This memorization makes models vulnerable to sophisticated attacks aimed at extracting confidential information. One of the most critical threats is Membership Inference Attacks (MIAs), which enable adversaries to infer whether a specific data point was included in a model's training set by analyzing subtle cues in its outputs, such as prediction vectors. MIAs exploit the fact that models often produce overconfident predictions for training samples compared to unseen data. This difference in confidence creates a vulnerability that attackers can use to distinguish between training set members and non-members. Authors in (Shokri et al., 2017) were the first to demonstrate the vulnerability of widely used Machine Learning as a Service (MLaaS) platforms, such as the Google Prediction API and Amazon ML, to membership inference attacks (MIAs). Their work highlighted how easily attackers could extract information about training data using only model outputs. Since then, more sophisticated variants of MIAs have been proposed to target different model architectures and data types. In response to these attacks, various defense mechanisms have been developed, generally falling into two categories: provable defenses and empirical defenses. Provable defenses are grounded in formal privacy guarantees, most notably through differential privacy (DP). Although such approaches provide strong theoretical protection, they often lead to significant accuracy degradation. Empirical defenses, on the other hand, aim to protect privacy while maintaining high model accuracy. These methods focus on obfuscating the signals that MIAs exploit, without providing formal privacy guarantees. Among the commonly used strategies are regularization techniques, masking of confidence scores, and knowledge distillation. However, despite their apparent effectiveness, these defenses have significant limitations

Kassa, R., Adi, K. and Bouhaddi, M.
Optimal Noise Injection on Training Data: A Defense Against Membership Inference Attacks.
DOI: 10.5220/0013639300003979
In *Proceedings of the 22nd International Conference on Security and Cryptography (SECRYPT 2025)*, pages 531-538
ISBN: 978-989-758-760-3; ISSN: 2184-7711

against sophisticated attacks and do not always ensure an effective trade-off between privacy and utility. This trade-off highlights the urgent need for new defense mechanisms capable of balancing privacy and utility effectively without degrading model performance.

To achieve this goal, we propose a novel defense mechanism based on optimal noise injection. The main idea is to inject a carefully designed and controlled noise vector into each training sample during the training phase. This noise is optimized through the following mechanisms:

- Maximizing prediction entropy to minimize the confidence gap between member and non-member samples.
- Leveraging Shapley values to guide feature-adaptive noise injection.
- Enhancing local robustness to mitigate misclassifications of inputs near the decision boundary.

In contrast to existing methods that apply noise uniformly, our defense strategy adaptively perturbs inputs according to feature influence, minimizing perturbations on salient features to preserve accuracy, while amplifying them on less critical features to introduce targeted uncertainty.

We evaluated our defense on the Purchase100 and Texas100 datasets, demonstrating a notable reduction in black-box MIA success rates and achieving a superior privacy-utility trade-off. These results confirm the effectiveness of optimal noise injection for robust and practical protection in privacy-preserving machine learning, while effectively addressing the limitations of existing approaches.

In summary, the key contributions of this paper are as follows.

- We propose a new defense mechanism based on optimal noise injection, which involves injecting a carefully designed and optimized noise vector into each training sample during the training phase.
- We show that our defense effectively mitigates black-box MIAs by confounding the attacker's inference classifier into a state of uncertainty, while still achieving a favorable trade-off between privacy and utility.

2 RELATED WORK

We provide a comprehensive overview of Membership Inference Attacks and the corresponding defense mechanisms.

2.1 Membership Inference Attacks

MIAs attacks can target all types of ML model deployments, including black-box scenarios, where the attacker has access only to the target model's prediction vectors, without any knowledge of its internal parameters. (Shokri et al., 2017) introduced one of the first black-box MIAs against ML models. Their approach involves building multiple shadow models to mimic the behavior of the target model, followed by training an attack model using the prediction vectors generated by these shadow models. The attack model is then used to infer whether a given input sample is a member of the target model's training set. (Salem et al., 2018) later relaxed many of the assumptions of (Shokri et al., 2017) and demonstrated that even a single shadow model can be sufficient to mount an effective MIA. Subsequently, a new class of metric-based membership inference attacks (MIAs) was proposed. These attacks determine membership by computing specific metrics (prediction correctness, prediction entropy, prediction confidence, or loss) on the model's output for a given sample and comparing the result to a predefined threshold. This approach was first introduced by (Yeom et al., 2018) and further explored by (Salem et al., 2018) and (Song et al., 2019). Later, (Song and Mittal, 2021) summarized and extended these works, and introduced a state-of-the-art metric-based attack called the privacy risk score. Other studies, such as (Choquette-Choo et al., 2021), investigated variants of MIAs that rely solely on predicted labels, without access to confidence scores. By leveraging input perturbation techniques, these attacks exploit the observation that the predicted class of a member instance is generally more resistant to changes than that of a non-member, as member predictions tend to be more stable. Furthermore, Carlini et al. (Carlini et al., 2022) introduced LiRA (Likelihood Ratio Attack), which employs a likelihood ratio test to compare a model's outputs when a sample is included in the training set versus when it is excluded. They also proposed evaluating attack effectiveness using the True Positive Rate (TPR) at very low False Positive Rates (FPR), a robust metric that has since become a standard benchmark in membership inference research.

2.2 Defenses Against MIAs

Several defense mechanisms have been proposed to mitigate the risks of MIAs by obscuring statistical differences between members and non-members. These include differential privacy, confidence score masking, regularization, and knowledge distillation, each

offering a distinct privacy-utility trade-off.

Differential Privacy: involves adding noise during training to prevent disclosure of member-specific information. Some approaches implement DP by adding noise to the model's objective function (Wang et al., 2017), while others apply it directly to gradients during optimization (Abadi et al., 2016; Pichapati et al., 2019). Despite offering strong privacy guarantees, it results in a significant degradation of model accuracy,

Confidence Score Masking: is another technique aimed at reducing the amount of information leaked through a model's predictions, thereby hindering membership inference. One variant limits outputs to the top-k predicted classes (Shokri et al., 2017) rather than the complete confidence vector. Another strategy perturbs the output probabilities by injecting noise to mislead attackers. For instance, (Bouhaddi and Adi, 2023; Jia et al., 2019), adds carefully crafted adversarial noise to the prediction vectors applied after the model's computation to preserve accuracy. However, (Song and Mittal, 2021) showed that even with such output perturbations, models can remain vulnerable to label-only attacks and metric-based attacks.

Regularization Techniques: aim to mitigate model overfitting and improve generalization capabilities. Many works (Leino and Fredrikson, 2020; Salem et al., 2018; Shokri et al., 2017) have demonstrated that overfitting is a key factor contributing to the effectiveness of MIAs. In this context, (Shokri et al., 2017) highlighted that regularization of L1 and L2 can effectively reduce the success rate of MIA. Subsequently, various regularization methods have been explored to counter these attacks, including dropout (Srivastava et al., 2014), model stacking (Salem et al., 2018), and early stopping (Song and Mittal, 2021). (Nasr et al., 2018) introduced Adversarial Regularization, a min-max optimization approach that incorporates an adversarial regularizer into the loss function to jointly minimize prediction loss and maximize membership privacy. Furthermore, (Li et al., 2021) introduced a regularization method that combines Mixup with Maximum Mean Discrepancy (MMD). This technique interpolates between pairs of training samples and aligns the prediction output distributions of members and non-members using MMD.

Knowledge Distillation: is a technique that leverages the output of a teacher model to train a student model, which is then made public. The objective is to enable the student model to achieve accuracy close to that of the teacher model, without directly exposing sensitive data. (Shejwalkar and Houmansadr, 2021) proposed distillation for membership privacy, which uses public unlabeled datasets to train a student model with soft labels generated by a teacher model. However, its effectiveness is limited by the availability of suitable public datasets. To overcome this, complementary knowledge distillation and pseudo complementary knowledge distillation (Zheng et al., 2021), along with knowledge cross-distillation (Chourasia et al., 2021), perform knowledge distillation directly using private data. Additionally, two variants based on self-distillation have been introduced, namely SELENA (Tang et al., 2022) and SEDMA (Nakai et al., 2024), enable a single model to improve its performance by reusing its own predictions as pseudo-labels during training.

Despite the progress made, current defenses against MIAs still have limitations, and the privacy-utility trade-off remains a major challenge. It is therefore essential to develop more adaptive and lightweight defenses capable of ensuring an optimal privacy-utility trade-off.

3 PRELIMINARIES AND PROBLEM FORMULATION

In this section, we present the key concepts together with their respective notations used in this paper, and then describe the threat model adopted for our study.

3.1 Preliminaries and Notation

Supervised ML. In this paper, we study supervised ML for classification tasks. We consider a classification model, denoted by $f : \mathbb{R}^d \to \mathbb{R}^k$, where $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ represents an input feature vector and $f(x) \in \mathbb{R}^k$ corresponds to a predicted probability distribution over the k possible classes. The model, parameterized by θ, is trained on a dataset $\mathcal{D}_{\text{tr}} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{|\mathcal{D}_{\text{tr}}|}$, where $x^{(n)} \in \mathbb{R}^d$ denotes an input feature vector and $y^(n)$ is the corresponding ground truth label. The training objective is to minimize the average prediction loss over $\mathcal{D}_{\text{tr}}$:

$$\min_{\theta} \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{n=1}^{|\mathcal{D}_{\text{tr}}|} \mathcal{L}\left(f_\theta(x^{(n)}), y^{(n)}\right) \tag{1}$$

Where $|\mathcal{D}_{\text{tr}}|$ denotes the size of the training set and $\mathcal{L}$ is the prediction loss function (e.g., cross-entropy loss) measuring the discrepancy between the model's output and the ground truth labels. The model's output $f(x) \in \mathbb{R}^k$ satisfies $\sum_{j=1}^{k} f(x)_j = 1$, meaning it represents a valid probability distribution over the k possible classes. The predicted label is then obtained as $\hat{y} = \arg\max_j f(x)_j$.

Prediction Entropy: measures the uncertainty of a model with respect to its predictions. Low entropy indicates that the model is highly confident, while high entropy implies that the model is uncertain.

Let $f(x) = (y_1, \ldots, y_k)$ be the model prediction vector for a sample x. The prediction entropy $\mathcal{H}(f(x))$ is then calculated as:

$$\mathcal{H}(f(x)) = -\sum_{j=1}^{k} y_j \log(y_j) \tag{2}$$

Shapley Values: are used to assess the importance of each feature x_i in a model's prediction by quantifying its average marginal contribution across all possible subsets of features. The Shapley value $\varphi_i(x)$ of a feature x_i is defined as:

$$\varphi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(x_{S \cup \{i\}}) - f(x_S) \right] \tag{3}$$

Where N represents the set of all features in the model, defined as $N = \{1, 2, ..., d\}$. The subset S is a selection of features that exclude x_i, denoted $S \subseteq N \setminus \{i\}$. The function $f(x_S)$ corresponds to the prediction of the model using only the features of S, while $f(x_{S \cup \{i\}})$ represents the prediction after including the feature x_i. The difference $f(x_{S \cup \{i\}}) - f(x_S)$ quantifies the marginal contribution of x_i to the prediction of the model. Finally, the weighting factor $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ ensures a fair evaluation of the contribution of each feature in all possible subsets.

A high Shapley value for x_i indicates a strong influence on the prediction of the model, whereas a low value suggests a minor contribution.

3.2 Threat Model

Adversarial Capabilities. We consider an adversary with black-box access to the target model f, meaning they can query the model via a prediction API and receive the corresponding probability vectors. The adversary's objective is to carry out an MIA attack (Salem et al., 2018; Shokri et al., 2017) to determine whether a given sample x was part of the model's training dataset. By issuing multiple queries, the adversary gathers sufficient information to train a binary classifier, referred to as the "attack classifier" $\mathcal{A}$, which predicts the membership status of a sample using confidence scores. Formally, the attack classifier $\mathcal{A}$ is defined as follows:

$$\mathcal{A}(x, f(x)) \rightarrow [0, 1] \tag{4}$$

$\mathcal{A}$ takes as input a sample x and its prediction vector $f(x)$, then outputs a probability indicating whether x belongs to the training set. A value close to 0 suggests that x is a non-member, whereas a value close to 1 indicates a high likelihood of membership.

Adversarial Knowledge. We assume that the adversary does not have access to the internal details of the target model, such as its parameters, weights, or the full training dataset, except for a small subset of samples that can be leveraged to perform an MIA attack. However, it is assumed that the adversary is aware of the model architecture and the deployed defense mechanism. These assumptions allow us to consider powerful adversaries and evaluate the robustness of our approach against advanced threats.

4 PROPOSED DEFENSE MECHANISM

The proposed defense mechanism aims to mitigate MIA by introducing a carefully controlled noise perturbation η into training samples during the training phase. The goal is to obscure statistical patterns that an attacker might exploit while ensuring that the modified samples remain useful for classification. This requires balancing *security* and *utility* through well-defined constraints that guide the noise injection process. An overview of our defense mechanism is illustrated in Figure 1.

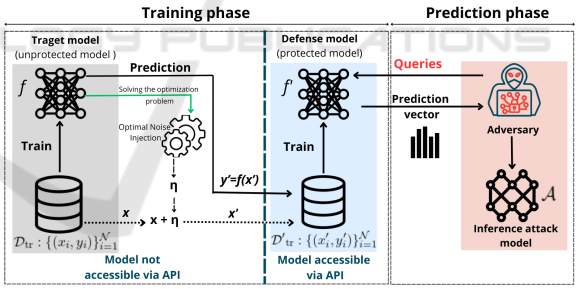

Figure 1: Overview of the Proposed Defense Mechanism Against Black-Box Membership Inference Attacks.

The fundamental idea of our approach is to introduce perturbations that increase the uncertainty in the model's predictions, thereby making it harder for an attacker to infer membership information. At the same time, the noise should be controlled so that it does not significantly degrade the model's classification performance. To achieve this, we define two key constraints: a *security constraint*, which ensures sufficient uncertainty, and a *utility constraint*, which preserves the correctness of the classification.

The *security constraint* ensures that noise increases the entropy of the prediction of the model, making it more difficult for an attacker to detect

the membership status. The intuition behind this is that low-entropy predictions provide clear confidence scores that can be exploited by an attacker, whereas higher entropy introduces uncertainty, masking the membership signal. We formalize this constraint as follows.

$$\mathcal{H}\big(f(x+\eta)\big) \geq \tau \tag{5}$$

where τ is a predefined entropy threshold that ensures a minimum level of uncertainty in the predictions of the model.

Next, to preserve *utility*, we impose a constraint that ensures that the perturbed sample remains classified in the same category as the original. The intuition here is that, while we want to introduce uncertainty for an attacker, we must ensure that the model still recognizes the sample correctly. This leads to the following constraint:

$$\arg\max_j f(x+\eta)_j = \arg\max_j f(x)_j \tag{6}$$

where *argmax* returns the argument that produces the maximum value of a function. This guarantees that the perturbation does not change the decision boundary in a way that causes misclassification.

To refine this constraint, we incorporate *Shapley values* to control how noise is applied to different features. Since not all features contribute equally to the model prediction, perturbing critical features too much could lead to misclassification, while perturbing less important features can provide the necessary uncertainty without affecting the accuracy. Let $\eta = (\eta_1, \ldots, \eta_d)$ be the perturbation vector to be applied to $x = (x_1, \ldots, x_d)$, we formulate the constraint as follows:

$$|\eta_i| \leq \varepsilon \cdot \left(1 - \beta \cdot \frac{\varphi_i(x)}{\max(\varphi_i(x))}\right) \tag{7}$$

where ε is a noise budget, β is a weighting parameter, and $\varphi_i(x)$ represents the Shapley value of the feature i. This ensures that highly influential features receive minimal noise, preserving classification accuracy while still introducing controlled uncertainty.

However, for samples located near the decision boundary, small perturbations may still flip the predicted class. To prevent this, we introduce an additional utility-preserving constraint based on the local robustness of the model, which limits the overall magnitude of the perturbation according to the local sensitivity and confidence margin of the model. Thus, we formulate the constraint as follows:

$$\|\eta\|_2 \leq \frac{f_{\text{gap}}(x)}{\|\nabla f(x)\|_2} \tag{8}$$

where $f_{\text{gap}}(x)$ denotes the margin between the top two predicted class scores, and $\|\nabla f(x)\|_2$ is the norm of the input gradient, indicating the model's sensitivity to perturbations. This constraint ensures that if the model is highly sensitive or if the decision margin is small, the injected noise remains small to prevent altering the prediction. Critical features are minimally perturbed, and the total noise remains within a safe margin, preventing misclassification.

Bringing everything together, our final objective is to minimize the magnitude of the perturbation while satisfying both the security and utility constraints.

$$\begin{aligned} \min_{\eta} \quad & \|\eta\|_2 \\ \text{s.t.} \quad & \mathcal{H}\big(f(x+\eta)\big) \geq \tau, \\ & |\eta_i| \leq \varepsilon \cdot \left(1 - \beta \cdot \frac{\varphi_i(x)}{\max(\varphi_i(x))}\right) \quad \forall i = 1, \ldots, d, \\ & \|\eta\|_2 \leq \frac{f_{\text{gap}}(x)}{\|\nabla f(x)\|_2}. \end{aligned} \tag{9}$$

This formulation ensures that noise remains minimal while satisfying security and utility constraints, striking a trade-off between privacy protection and model performance.

5 EXPERIMENTATION

In this section, we present an experimental evaluation of our proposed defense mechanism. We first describe the experimental setup and the evaluation metrics used to assess the trade-off between privacy and utility. Then, we analyze the experimental results to demonstrate the effectiveness of our approach in mitigating MIAs.

5.1 Experimental Setup

Datasets. To evaluate the effectiveness of our proposed defense mechanism, we conduct experiments on two widely used benchmark datasets for MIAs: Purchase100 and Texas100.

Purchase100. Consists of 197,324 customer purchase records, each with 600 binary features indicating whether a specific item was purchased. The classification task is to predict customer shopping patterns in 100 classes.

Texas100. Contains 67,330 hospital discharge records, each with 6,170 binary features representing the presence or absence of specific symptoms. The goal is to predict the medical procedure assigned to the patient among 100 classes.

Target Model. We use a fully connected neural network as the target model for the Purchase100 and Texas100 datasets. The architecture includes four hidden layers with sizes [1024, 512, 256, 128]. Each hidden layer uses the ReLU activation function, while the output layer applies a softmax function to predict the probabilities of more than 100 classes. The models are trained using the Adam optimizer, with the cross-entropy loss function, a learning rate of 0.001, over 100 epochs. The datasets used in our experiments are summarized in Table 1.

Table 1: Dataset splits used in our experiments: **Train** is used to train the target model; **Test**, to evaluate its accuracy. **Known** denotes the subset of training data accessible to the adversary for building the attack model. **Target** is used to evaluate membership inference attacks and contains an equal number of member and non-member samples.

Dataset	Train	Test	Known	Target
Purchase100	20,000	20,000	10,000	10,000
Texas100	10,000	10,000	5,000	5,000

Inference Attack Model. In our evaluation, we adopt a black-box MIA setup where a shadow model is trained on part of the target model's training data and non-member samples from the same distribution. The purpose of this model is to generate outputs for training an attack model.

The attack model consists of three fully connected subnets, each operating on the prediction vector, the one-hot encoded label, and their concatenation. Each subnetwork uses a ReLU activation function, with weights initialized from a normal distribution $\mathcal{N}(0, 0.01)$ and biases initialized to zero. The model is trained using the Adam optimizer, learning rate of 0.001 for 100 epochs, with the cross-entropy loss function. The final output is a membership probability that indicates the likelihood that a given sample belongs to the target model's training data.

Defense Model. Our defense model adopts the same architecture as the target model. However, rather of being trained directly on the original data, it is trained on a noisy dataset, generated by applying our proposed feature-adaptive noise injection mechanism to each input sample. Specifically, for each training instance x, a noise vector η is computed under security and utility constraints, and added to generate a perturbed input $x' = x + \eta$. Each perturbed input x' is associated with a soft label $y' = f(x')$, representing the probability distribution output by the target model when evaluated on the noisy sample. Simultaneously, the original hard label y is retained to ensure that classification performance is preserved.

To train the defense model, we use a combined loss function, which includes two components: the cross-entropy loss (CE) and the Kullback–Leibler (KL) divergence. The loss function is defined as follows:

$$\mathcal{L}(x') = \alpha \cdot \mathrm{KL}\left(f'(x') \parallel y'\right) + (1-\alpha) \cdot \mathrm{CE}\left(f'(x'), y\right)$$

where $\alpha \in [0,1]$ is a hyperparameter that balances the trade-off between privacy and utility. The defense model is trained using the Adam optimizer, with a learning rate of 0.001 for 100 epochs.

Evaluation Metrics. To evaluate our defense, we use four key metrics: Inference Accuracy and attack AUC (Area Under the ROC Curve) to evaluate privacy protection, where values near 0.5 indicate a strong defense against MIAs, and Test Accuracy and Generalization Gap to measure model utility and generalization. These metrics together provide a comprehensive understanding of the privacy-utility trade-off.

5.2 Experimental Results

In this section, we evaluate the effectiveness of our proposed mechanism against black-box MIAs. To this end, we conducted a comparative study involving three models under the same training and attack configurations: an undefended model, trained without any privacy mechanism and serving as a baseline; a uniform noise defense model, in which a fixed perturbation is applied equally to all training samples regardless of feature importance; and our adaptive noise defense, which injects feature-wise optimized noise guided by utility and security constraints.

To assess the utility of each model, we track their classification performance on unseen data using test accuracy over training epochs. As illustrated in Figure 2, the adaptive noise defense achieves test accuracy close to that of the undefended model and consistently outperforms the uniform noise defense. This indicates that our method preserves classification performance by selectively perturbing features in a way that minimizes the impact on critical decision components.

Next, we examine the privacy protection offered by each model using Attack AUC, which measures the effectiveness of the MIA across all possible decision thresholds. As shown in Figure 3, the undefended model yields high AUC values, which confirms its vulnerability. The uniform noise defense offers limited mitigation, whereas our adaptive noise defense significantly reduces the AUC, approaching the ideal baseline value of 0.5, which corresponds to random guessing and thus provides strong privacy protection.

Finally, to understand the privacy-utility trade-off, we analyze the relationship between inference attack

accuracy and the generalization gap. As presented in Figure 4, the undefended model exhibits both a high generalization gap and high inference accuracy, indicating strong overfitting and exposure to MIAs. In contrast, our adaptive noise defense reduces both the gap and the success of the attack, suggesting that lowering the overfitting improves privacy while retaining generalization. The uniform noise defense falls between the two, with moderate performance on both fronts.

These results confirm that our adaptive noise injection method provides a balanced defense by safeguarding sensitive membership information while maintaining the classification utility of the model.

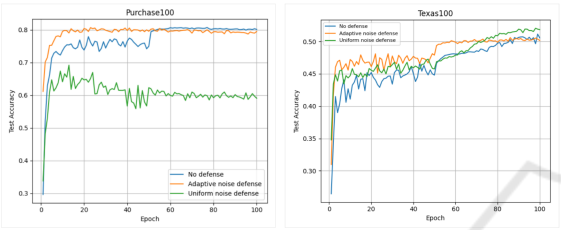

Figure 2: Test Accuracy over Epochs for Different Defense Strategies.

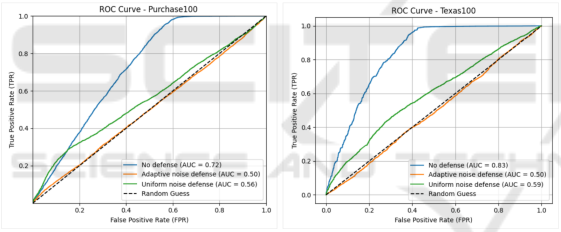

Figure 3: Attack AUC Under Different Defenses.

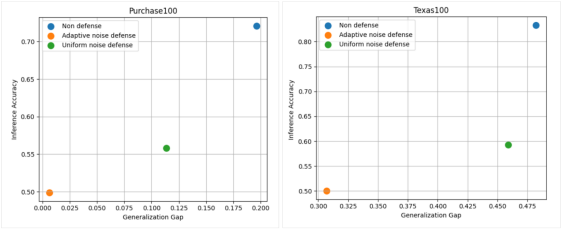

Figure 4: Privacy-Utility Trade-off: Inference Accuracy Vs. Generalization Gap.

6 CONCLUSION

In this work, we introduced a novel defensive approach against black-box MIAs, based on an optimized and feature-adaptive noise injection mechanism. The originality of our approach lies in its ability to adjust the injected perturbation for each feature according to its influence on the decision of the model, as measured by the Shapley values.

Our method is built upon two key constraints. The first is a utility constraint, guided by Shapley values and local robustness, which aims to preserve the most influential features and prevent misclassification. The second is a security constraint that increases the prediction entropy to introduce controlled uncertainty. Together, these constraints strike a balance between maintaining model accuracy and mitigating the risk of membership inference, thereby enhancing privacy protection.

Our experimental results demonstrate that the proposed method achieves a favorable trade-off between privacy and utility, effectively reducing the risk of membership inference without substantially degrading classification performance. These findings highlight the potential of combining utility-aware and entropy-based constraints to enhance privacy in machine learning models.

Future work will focus on refining these constraints, particularly through the development of more adaptive noise injection strategies. We also aim to extend our approach to other threat models, including white-box attacks, and to evaluate its effectiveness on more complex architectures and across diverse application domains.

REFERENCES

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

Bouhaddi, M. and Adi, K. (2023). Mitigating membership inference attacks in machine learning as a service. In *IEEE International Conference on Cyber Security and Resilience, CSR 2023, Venice, Italy, July 31 - Aug. 2, 2023*, pages 262–268. IEEE.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. (2022). Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.

Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. (2021). Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1964–1974. PMLR.

Chourasia, R., Enkhtaivan, B., Ito, K., Mori, J., Teranishi, I., and Tsuchida, H. (2021). Knowledge cross-distillation for membership privacy. arXiv preprint, available at https://arxiv.org/abs/2111.01363.

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. (2019). Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Con-*

ference on Computer and Communications Security, pages 259–274.

Leino, K. and Fredrikson, M. (2020). Stolen memories: Leveraging model memorization for calibrated White-Box membership inference. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security 20)*, pages 1605–1622.

Li, J., Li, N., and Ribeiro, B. (2021). Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 5–16.

Nakai, T., Wang, Y., Yoshida, K., and Fujino, T. (2024). Sedma: Self-distillation with model aggregation for membership privacy. *Proceedings on Privacy Enhancing Technologies*.

Nasr, M., Shokri, R., and Houmansadr, A. (2018). Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 634–646.

Pichapati, V., Suresh, A. T., Yu, F. X., Reddi, S. J., and Kumar, S. (2019). Adaclip: Adaptive clipping for private sgd. arXiv preprint, available at https://arxiv.org/abs/1908.07643.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. (2018). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv preprint, available at https://arxiv.org/abs/1806.01246.

Shejwalkar, V. and Houmansadr, A. (2021). Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9549–9557.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Song, L. and Mittal, P. (2021). Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632. USENIX.

Song, L., Shokri, R., and Mittal, P. (2019). Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 241–257. ACM.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Tang, X., Mahloujifar, S., Song, L., Shejwalkar, V., Nasr, M., and Houmansadr, A. (2022). Mitigating membership inference attacks by Self-Distillation through a novel ensemble architecture. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security 22)*, pages 1433–1450.

Wang, D., Ye, M., and Xu, J. (2017). Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, volume 30.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. (2018). Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE.

Zheng, J., Cao, Y., and Wang, H. (2021). Resisting membership inference attacks through knowledge distillation. *Neurocomputing*, 452:114–126.