# An Overview of Membership Invariant Privacy State-Of-The-Art

Amirhossein Jadidi[1*], Amirahmad Shafiee[2*]

*Abstract*— **Our work is focused on membership invariance privacy, an area closely linked with the exploration of membership inference (MI) attacks. In Member Inference (MIA) attacks, the adversary tries to determine whether an instance is used to train a machine learning (ML) model. MIA attacks are a major privacy concern when using private data to train ML models. Existing literature mainly revolves around exploiting the tendency of machine learning models to overfit training data and thus have a very low loss on training instances. Proposed defenses either rely on theoretical constraints like DP-SGD or use machine learning-based methods, which have shown promising results. In our study, we highlight a potent attack, a promising machine learning technique called MIST, and a defense strategy that differs from the conventional paradigms of DP and ML reliance.**

## I. INTRODUCTION

Machine learning is the foundation of popular Internet services such as image and speech recognition and natural lan- guage translation. Many companies also use machine learning internally, to improve marketing and advertising, recommend products and services to users, or better understand the data generated by their operations. In all of these scenarios, ac- tivities of individual users—their purchases and preferences, health data, online and offline transactions, photos they take, commands they speak into their mobile phones, locations they travel to—are used as the training data. Internet giants such as Google and Amazon are already offering "machine learning as a service." Any customer in possession of a dataset and a data classification task can upload this dataset to the service and pay it to construct a model. The service then makes the model available to the customer, typically as a black-box API. For example, a mobile-app maker can use such a service to analyze users' activities and query the resulting model inside the app to promote in-app purchases to users when they are most likely to respond. Some machine- learning services also let data owners expose their models to external users for querying or even sell them.

ML models are often trained on data with sensitive user information such as clinical records and personal photos. Hence, ML models trained using sensitive data can leak private information about their data owners. This has been demonstrated through various inference attacks ( [1], [2], [3]) , and most notably the member- ship inference attack (MIA) ( [4]) which is the focus of our work. An MIA adversary with a blackbox or whitebox access to a target model aims to determine if a given target sample belonged to the private training data of the target model or not. MIAs are able to distinguish the members from non-members by learning the behavior of the target model on member versus non-member inputs.

**Shokri's Attack (First MIA)**: At the first time, Shokri et al. [4] proposed a method to attacks machine learning models. The attacker first decides on the attack model, which can be either a black-box approach (where the attacker only has access to the model's predictions) or a white-box approach (where the attacker has more detailed knowledge about the model, including its architecture and parameters).

The attacker collects or generates a dataset that is similar to the target model's training data. This dataset is used to probe the target model to observe its behavior. In some cases, the attacker may have access to a portion of the actual training data or data from a similar distribution.

In black-box settings, the attacker often trains

[1]Amirhossein Jadidi, 402300461, Department of Electrical Engineering

[2]Amirahmad Shafiee, 99104027, Department of Mathematical Sciences

* Equal Contribution

one or more shadow models to mimic the behavior of the target model. These shadow models are trained on data that is known to the attacker, which is split into "in" (used for training the shadow model) and "out" (not used for training) subsets. The goal is to replicate the target model's decision boundaries closely, allowing the attacker to learn how the target model behaves on data it has seen versus data it hasn't seen.

Using the outputs from the shadow models (or directly from the target model in white-box attacks), the attacker trains an attack model. This model is designed to distinguish between the target model's predictions on data points that were in its training set versus those that were not. The training process involves creating training examples for the attack model, labeled as "member" or "non-member," based on the shadow models' predictions or the observed behavior of the target model.

With the attack model trained, the attacker then queries the target model with new data points and feeds the observed predictions (and possibly additional metadata, such as confidence scores) into the attack model. The attack model then predicts whether each new data point was part of the target model's training dataset.

The study identifies model overfitting, the complexity of the model, the uniqueness of the data points, and the amount of information provided by model predictions as critical factors that can increase the vulnerability of ML models to these attacks. This analysis not only advances our understanding of why ML models are susceptible to MIAs but also guides the development of strategies to mitigate these vulnerabilities.

The paper goes beyond merely identifying vulnerabilities by evaluating various defense mechanisms designed to protect ML models from MIAs. Through comparative analysis, it assesses the effectiveness of techniques such as model regularization, differential privacy, and data augmentation in reducing the risk of MIAs while considering the trade-offs between model privacy, accuracy, and utility. This evaluation is invaluable for practitioners and researchers seeking to implement ML models that balance performance with privacy considerations.

By uncovering the nuanced dynamics of MIAs against ML models and testing the effectiveness of different defense strategies, the paper sets a solid foundation for future research. It highlights the need for more robust privacy-preserving techniques in machine learning and opens up avenues for exploring MIAs in emerging ML paradigms, such as federated learning, and their intersection with other types of privacy and security threats.

In this report, we will analyze different type of MIA in section II and defences against MIAs in section III. Finally we evaluate defense effectiveness in section IV.

## II. Membership Inference Attacks

In the literature of Membership Inference Privacy (MIP), several techniques for Membership Inference (MIA) attacks have been proposed. The simplest ideas rely heavily on the discrimination inferred from the loss distribution of the attacked model, contrasting the training data with the validation data.

### *A. Using Target Instances*

**LOSS (Using loss with a global threshold):** Yeom et al. introduced an attack that predicts an instance $x$ as a member if the target model's loss on $x$ is below the average training loss.

**Class-NN (Training class-specific Neural Networks for MIA):** Shokri's attack [4] trains multiple neural network-based membership classifiers, one for each class. Training data are obtained using the shadow model technique, which is widely used in subsequent attacks.

Using the shadow model technique, one assumes that the adversary has access to a dataset $D_A$, which contains the target instance and is from the same distribution as the dataset used to train $F_T$. The adversary creates $k$ subsets $D_1, D_2, \ldots, D_k$ from $D_A$, and uses the same process used for training $F_T$ to train $k$ models, one from each $D_i$. These are called shadow models. For each instance $x$, some shadow models were trained using $x$, and others were trained without it. The predictions of these models on instances in $D_A$ provide training data for membership classifiers.

**Modified Entropy:** Song et al. proposed using a class-specific threshold on a modified entropy measure based on the model prediction to determine membership. This approach is a simplified

version of Shokri's attack [4] and is very similar to using a class-specific loss threshold to determine membership.

**LIRA:** Carlini et al. [3] proposed an instance-specific threshold attack. For each instance, distributions for losses from models trained with $x$ and another distribution from models trained without $x$ are obtained from the shadow models. A threshold can be chosen using the likelihood ratio. Carlini et al. suggest choosing the threshold that optimizes attack effectiveness at a very low false positive rate. This attack is referred to as the LIRA attack in subsequent sections.

### *B. Using Perturbations of Target Instances*

Some other methods, use other characteristics of overfitting and utilize perturbations of instances to further improve their efficiency towards violating models' privacy.

**Random perturbations:** Jayaraman et al. propose an attack that generates multiple perturbed instances by adding Gaussian noise to $x$, and then queries $F_T$ using these perturbed instances. The attack counts how many times the prediction loss of these instances is higher than that of $x$. The instance $x$ is predicted to be a member if the count is beyond a threshold. This attack is referred to as the random-perturbation attack in subsequent sections.

**CANARY:** The attack proposed by Wen et al. also uses shadow models. For each target instance $x$, shadow models are partitioned into two sets: those trained with $x$ and those trained without. A set of canaries is computed for each $x$, each generated via gradient descent starting from a slightly perturbed version of $x$, searching for an $x'$ such that the difference between the average losses of $x'$ from models in the two sets is as large as possible. These canaries are then used to query the target model and use the loss to predict membership. This attack is called the CANARY attack in subsequent sections. In the following comes the algorithm of Canary method. Notice that $x_{mal}$ is the canary, and $L_{out}$ is negative of the $CE_{loss}$ for the introduced group.

**Adversarial Perturbation (Label Only):** Choquette-Choo et al. [5] proposed two attacks for situations where only the predicted label is provided and identified the adversarial perturbation attack as more effective. In this attack, one applies the adversarial example generation technique from to generate $x'$ that is close to $x$ but with a different predicted label by $F_T$. The attack predicts membership if $||x' - x||_2$ is high.

**Algorithm 1** Canary Algorithm

**Require:** IN shadow models $S_{\text{in}} = \{\theta_{\text{in}1}, \ldots, \theta_{\text{in}n}\}$, OUT shadow models $S_{\text{out}} = \{\theta_{\text{out}1}, \ldots, \theta_{\text{out}m}\}$, target data point $(x^*, y^*)$, batch size $b$, optimization steps $T$, perturbation bound $\epsilon$, input domain $I$

$\Delta_{\text{mal}} = 0$
**for** $t = 1, \ldots, T$ **do**
  Shuffle the index for $S_{\text{in}}$ and $S_{\text{out}}$
  Calculate loss on OUT models:
  $g_{\Delta_{\text{mal}}} = \nabla_{\Delta_{\text{mal}}} \frac{1}{b} \sum_{i=1}^{b} L_{\text{out}}(x^* + \Delta_{\text{mal}}, y^*, \theta_{\text{out}i})$
  Calculate loss on IN models (removed when offline):
  $g_{\Delta_{\text{mal}}} += \nabla_{\Delta_{\text{mal}}} \frac{1}{b} \sum_{i=1}^{b} L(x^* + \Delta_{\text{mal}}, y^*, \theta_{\text{in}i})$
  Update $\Delta_{\text{mal}}$ based on $g_{\Delta_{\text{mal}}}$
  Project $\Delta_{\text{mal}}$ onto $\|\Delta_{\text{mal}}\|_\infty \leq \epsilon$ and $(x^* + \Delta_{\text{mal}}) \in I$
  $x_{\text{mal}} = x^* + \Delta_{\text{mal}}$
**end for**
**return** $x_{\text{mal}}$

## III. Defences against MIAs

### *A. Differential Privacy Approaches*

**Differential privacy (DP).** is a widely used privacy-preserving technique. DP based defense techniques, such as DP-SGD, add noise to the training process. This provides a theoretical upper-bound on the effectiveness of any MIA against any instance. Unfortunately, achieving a meaningful theoretical guarantee (e.g., with a resulting $\varepsilon < 5$) requires the usage of very large noises. However, model trainers could use much smaller noises in DP-SGD. While doing this fails to provide a meaningful theoretical guarantee (the $\varepsilon$ value would be too large), this can nonetheless provide empirical defense against MIA. In, extensive experiments have shown that several other defenses can provide a better empirical privacy-utility tradeoff than DP-SGD.

*B. Machine Learning-based Approaches*

**Adversarial Regularization (Adv-reg).** Nasr et al. Proposed a defense that uses similar ideas as GAN. The classifier is trained in conjunction with an MIA model. The optimization objective of the target classifier is to reduce the prediction loss while minimizing the MIA accuracy.

**Mixup+MMD.** Li et al. proposed a defense that combines mixup data augmentation and MMD (Maximum Mean Discrepancy) based regularization. Instead of training with original instances, mixup data augmentation uses linear combinations of two original instances to train the model. It was shown in that this can improve target model's generalization. Li et al. found that they also help to defend against MIAs. Li et al. also proposes to add a regularizer that is the MMD between the loss distribution of members and the loss distribution of a validation set not used in training. This helps make the loss distribution of members to be more similar to the loss distribution on non-members.

**SELENA.** Tang et al. proposed a framework named SELENA. One first generates multiple (overlapping) subsets from the training data, then trains one model from each subset. One then generates a new label for each training instance, using the average of predictions generated by models trained without using that instance. Finally, one trains a model using the training dataset using these new labels.

**HAMP.** Chen et al. proposed a defense combining several ideas. First, labels for training instances are made smoother, by changing 1 to $\lambda$ and each 0 to $\frac{1-\lambda}{k-1}$, where $\lambda$ is a hyperparameter and $k$ is the number of classes. Second, an entropy-based regularizer is added in the optimization objective. This step is somewhat redundant given that the first step already increases the entropy of the labels. Third, the model does not directly return its output on a queried instance $x$. Instead, one randomly generates another instance, reshuffles the prediction vector of the randomly generated instance based on the order of the probabilities of $x$, and returns the reshuffled prediction vector. In essence, this last defense means returning only the order of the classes in the prediction vector but not the actual values.

**Mem-guard.** Jia et al. proposed the Mem-guard defense. In this defense, one trains an MIA model in addition to the target classifier. When the target classifier is queried with an instance, the resulting prediction vector is not directly returned. Instead, one tries to find a perturbed version of the vector such that the perturbation is minimal and does not change the predicted label, and the MI attack model output $(0.5, 0.5)$ as its prediction vector.

**MIST.** Li et al. [6] Proposed a novel defense, that capitalizes on the combined strengths of two distinct recent advancements in representation learning: Counterfactually-invariant representations and subspace learning methods. The main idea behind MIST was to focus on the instances that are "Distinctive" meaning that the difference between being in the training data and not being in the training data is high for them. Thus, we need not apply privacy methods to non-distinctive data. The idea is to train counterfactually-invariant models subject to interventions of single instances. However, this main idea comes with a huge computational cost, Therefore, the concept of subspace learning comes to the rescue. The formula of having a model, invariant to instances comes as follows:

We start by defining a regularization that we call cross-difference loss as follows:

$$\text{xdiff}(\theta_D, \theta_{D\setminus\{(x_M,y_M)\}}) =$$

$$\sup_{(x,y)\in\Omega} \left\| F(x;\theta_D)y - F(x;\theta_{D\setminus\{(x_M,y_M)\}})y \right\|_1,$$

where $\Omega$ is the support of $p(x, y)$ and $\|\cdot\|_1$ is the $L1$ norm. Equation (1) pushes the classifier $F(\cdot;\theta_D)$ trained with $(x_M, y_M)$ to have the same output as the classifier $F(\cdot;\theta_{D\setminus\{(x_M,y_M)\}})$ trained without $(x_M, y_M)$. Then, any classifier $F$ that satisfies

$$\sum_{M\in\{1,\ldots,n\}} \text{xdiff}(\theta_D, \theta_{D\setminus\{(x_M,y_M)\}}) = 0,$$

is membership-invariant. This is easy to verify since the condition in Equation, implies $F(x;\theta_{D\setminus\{(x_M,y_M)\}}) = F(x;\theta_D) = F(x;\theta_{D\setminus\{(x_{M'},y_{M'})\}})$, for all $M, M' \in \{1,\ldots,n\}$, that is, the classifier output is invariant to the environment in which it was trained.

To address the computational challenge, the idea of subspace learning is further proposed:

$$\text{xdiff}_c({w_t}^t) = \sum_{(x,y)\in D^c} \|F(x;{w_t}^c)y - \frac{\sum_{i\neq c} F(x;{\theta_e}^i)y}{C-1}\|_1.$$

In which, Data $D$ is divided into $C$ disjoint Subsets. ${\theta_e}^i$ is the parameters of the $i$th model after some iteration, and ${w_t}^c$ is the parameter to update, related to the $c$th model.

Whenever faced with distinctive data, minimizing the proposed loss over that data causes a huge impact over parameter ${w_t}^c$. However, if the datapoint is not distinctive, the model is already close to the average of all other models, hence the gradient won't be significant.

The algorithm consists of 3 phases. In the first phase, each model is trained locally. In the second phase, to each model, the proposed loss is applied for some $T_2$ steps. In the third phase, the parameters are being summed up. In the following comes a scheme of the algorithm:

**Algorithm 2** MIST: Membership-Invariant Subspace Training

**Require:** $C$: number of local models; $D$: training dataset; $E$: number of epochs. Hyperparameters $T1$, $T2$, and $\lambda$.

Initialize $\theta_0$ ▷ Model initialization
**for** $e = 1$ to $E$ **do**
  Partition $D$ into $D^c$, $1 \leq c \leq C$
  **for** $c = 1$ to $C$ **do**
    $\theta_e^c = \text{local\_training}(\theta_{e-1}, D^c, T_1)$
  **end for**
  **for** $c = 1$ to $C$ **do**
    $\theta_e^c =$ xdifference_update$(c, \theta^1, \ldots, \theta^C, D^c, \lambda, T_2)$
  **end for**
  $\theta_e = \frac{1}{C}\sum_{c=1}^{C} \theta_e^c$
**end for**
**Output:** $\theta_E$

### C. Other Approaches

In addition to the methods introduced in sections III-A and III-B, in this section we examine methods that are not examined in the form of DP and Machine Learning and have special methods.

**Algorithm 3** Local Training

**Require:** $\theta$: current model parameters; $D^c$: local training dataset; $T_1$: number of optimization steps

$w_0^c = \theta$
**for** $t = 1$ to $T1$ **do**
  Update $w_t^c$ with gradient descent to minimize $\frac{1}{|D^c|}L(w_{t-1}^c, D^c)$
**end for**
**return** $w_{T_1}^c$

**Algorithm 4** xdifference Update

**Require:** $c$: local model index; $\theta_1, \ldots, \theta^C$: current model parameters for all local models; $D^c$: local training dataset; $\lambda$: weight for cross-difference loss; $T_2$: number of optimization steps

$w_0^c = \theta^c$
**for** $t = 1$ to $T_2$ **do**
  Update $w_t^c$ with gradient descent to minimize $\lambda\frac{1}{|D^c|}\text{xdiff}_c(w_{t-1}^c)$
**end for**
**return** $w_{T_2}^c$

**DMP:** Shejwalkar et al. [7] proposed Distillation for Membership Privacy (DMP), a knowledge distillation based defense against membership inference attacks that significantly improves the membership privacy-model utility trade-offs compared to state-of-the-art defenses. It provided a novel criterion to generate/select reference data in DMP and achieve the desired trade-offs. It uses lower entropy of labels of the data generate/select reference data and build protected model against MIAs.

## IV. Defense Effectiveness Evaluation

In 1 a comparison of all defense mechanisms, can be seen. MIST has exceptionally outperformed all other mechanisms. In the figure, x-axis represents accuracy, while y-axis represents PLR. PLR is a metric to help quantify privacy. When $PLR_{0.001} = 100$, it means that when we ensure that no more than $0.1\%$ of non-members are falsely identified as members, if we identify 101 instances as members, we can expect that 100 are indeed members, and 1 of them is a false positive.

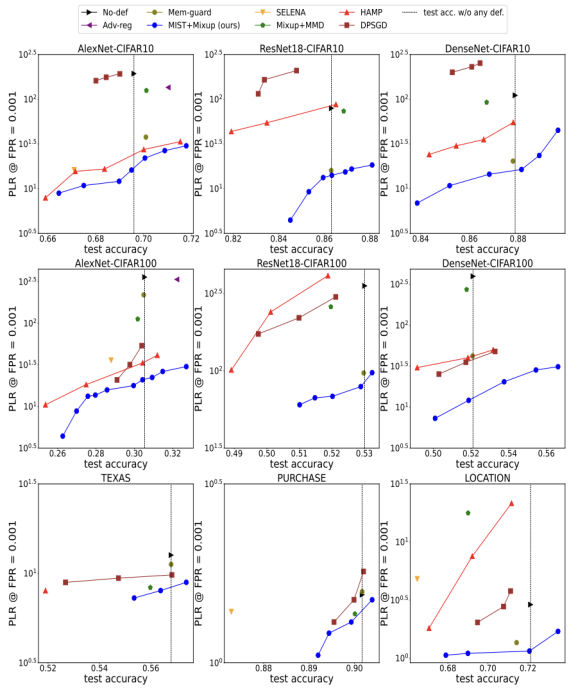


Fig. 1. Comparing all defenses using the highest MI attack PLR @ 0.001 FPR among all evaluated attacks. Defenses placed at the lower right corner are better (high test accuracy and low PLR). Notice that for PURCHASE, TEXAS and LOCATION datasets the mixup data augmentation is not applied. We exclude some results when the testing accuracy drop is larger than 5for clearer comparison

## V. Conclusion

This report has delved into the critical domain of membership inference privacy, highlighting the significant privacy risks that Membership Inference Attacks (MIAs) pose to machine learning (ML) models. As demonstrated, MIAs exploit the overfitting tendencies of ML models to discern whether a specific data instance was part of the model's training dataset, thus breaching individual privacy. The exploration of these attacks underlines the urgent need for robust defense mechanisms that can safeguard sensitive data used in training ML models against such privacy intrusions.

The literature review presented in this report showcases a range of defense strategies, from theoretical frameworks like Differential Privacy Stochastic Gradient Descent (DP-SGD) to innovative machine learning-based methods. Among these, Membership-Invariant Subspace Training (MIST) emerges as a particularly promising technique, offering a novel approach that deviates from conventional reliance on differential privacy (DP) and purely machine learning-based defenses. MIST's potential to provide a robust defense against MIAs, without compromising the utility of the ML models, marks a significant advancement in the field.

In conclusion, while MIAs present a formidable challenge to privacy in the ML domain, the ongoing development of sophisticated defense mechanisms like MIST offers a beacon of hope. Continued research and innovation in this area are imperative to develop more effective and efficient ways to protect privacy in the ever-evolving landscape of machine learning. The insights gained from this literature review underscore the importance of a multi-faceted approach to privacy preservation, combining theoretical insights with practical ML techniques to fortify defenses against MIAs and ensure the ethical use of data in machine learning.

## References


[1] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:207229839

[2] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, "Deep models under the GAN: information leakage from collaborative deep learning," *CoRR*, vol. abs/1702.07464, 2017. [Online]. Available: http://arxiv.org/abs/1702.07464

[3] N. Carlini, C. Liu, J. Kos, Ú. Erlingsson, and D. Song, "The secret sharer: Measuring unintended neural network memorization & extracting secrets," *CoRR*, vol. abs/1802.08232, 2018. [Online]. Available: http://arxiv.org/abs/1802.08232

[4] R. Shokri, M. Stronati, and V. Shmatikov, "Membership inference attacks against machine learning models," *CoRR*, vol. abs/1610.05820, 2016. [Online]. Available: http://arxiv.org/abs/1610.05820

[5] C. A. Choquette-Choo, F. Tramèr, N. Carlini, and N. Papernot, "Label-only membership inference attacks," *CoRR*, vol. abs/2007.14321, 2020. [Online]. Available: https://arxiv.org/abs/2007.14321

[6] J. Li, N. Li, and B. Ribeiro, "Mist: Defending against membership inference attacks through membership-invariant subspace training," 2023.

[7] V. Shejwalkar and A. Houmansadr, "Reconciling utility and membership privacy via knowledge distillation," *CoRR*, vol. abs/1906.06589, 2019. [Online]. Available: http://arxiv.org/abs/1906.06589