The Effectiveness of NN-based Defense Models Against Membership Inference Attacks

AMIT SARKER*, MASHRUR RASHIK*, and ERFAN ENTEZAMI*, University of Massachusetts Amherst, USA

Machine learning models are prone to memorizing sensitive data, making them vulnerable to membership inference attacks in which an adversary aims to guess if an input sample was used to train the model. Adversaries can reach users' sensitive information using membership inference attacks. As an example, Imagine if an adversary has access to a machine learning application, which is trained by patient data of a hospital. If he could check whether a user record is used in training data, he can understand that the user was once a patient in that hospital. To test the effectiveness of NN-based defense models, we will benchmark membership inference privacy risk by using non-neural network-based inference attacks on NN-based defense mechanisms. Furthermore, we will assess the effectiveness of a new privacy risk score that uses the training data's sampling probability to measure the risk of an attack. Our work is based on the Usenix 21 paper - "Systematic Evaluation of Privacy Risks of Machine Learning Models [12]."

CCS Concepts: • **Computer systems organization** → **Privacy**.

Additional Key Words and Phrases: membership inference, neural networks

ACM Reference Format:
Amit Sarker, Mashrur Rashik, and Erfan Entezami. 2024. The Effectiveness of NN-based Defense Models Against Membership Inference Attacks. 1, 1 (May 2024), 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

According to previous research, it is shown that machine learning models can memorize some information from their training data. In some cases, it may not seem a critical problem but in the case that the model is trained using users' sensitive information, it would be a significant privacy risk [2, 3]. In this project, we focused on membership inference attacks where an adversary tries to figure out whether a data sample was used to train a model or not [10, 14]. Thus, it can disclose users' sensitive information. For example, if adversaries figure out that a user's information is being used to train a machine learning model for a healthcare application, they can realize that users have been patient once. Since membership inference attacks can reveal the presence of user information in the training data using the target model, they can be considered as a good tool to evaluate the quality of privacy implementations [4]. Membership inference attacks can be divided into two main categories including black-box attacks and white-box attacks which will be explained in the next section. To reduce the privacy risk, several defenses against membership inference attacks have been proposed. In [1] membership inference attack is considered in the training phase. They tried to train the target model with high accuracy for prediction and low accuracy for membership inference attacks. Memguard [5] is another defense method that does not require retraining the model. We will review how previous works evaluated the membership inference privacy risk of machine learning models and explain two limitations in the previous works: first, many of the previous works proposed defense models using custom-trained NN-based classifiers to perform membership inference attacks. Second, the evaluation phase in previous works just focused on aggregate concepts of privacy risks faced by all data samples. In this project, our main contributions are as follows:

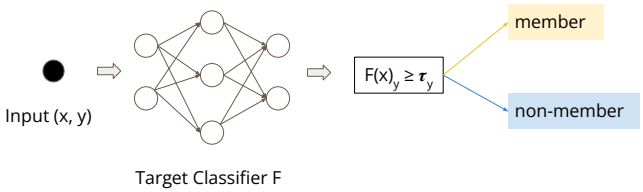

Fig. 1. Improving existing attacks with class-dependent thresholds.

- We propose a suit of metric-based attacks and use them to supplement existing neural network (NN) based MI attacks. We also evaluate multiple attack strategies and report the worst-case privacy risks.
- We review two defense models [5, 7]that are mentioned previously and demonstrate that they are not so efficient.
- We define a new metric called privacy risk score in order to evaluate the privacy risks of machine learning models in a fine-grained manner.
- We apply fine-grained analysis in conjunction with existing aggregated analysis for a thorough evaluation of privacy risks. Our implementation is publicly available online[1].

2 BACKGROUND

As mentioned previously, in membership inference attacks, adversaries try to figure out if a given data was used to train the target model or not, thus it would be a privacy risk for the user whose information is part of the training set of the target model. In this section, we express two categories of membership inference attacks and some previous works for each group. In the next section, we will talk about two state-of-the-art defense methods including adversarial regularization [7] and MemGuard [5].

*All authors contributed equally to this research.

Authors' address: Amit Sarker, asarker@umass.edu; Mashrur Rashik, mrashik@umass.edu; Erfan Entezami, eentezami@umass.edu, University of Massachusetts Amherst , Amherst, Massachusetts, USA, 01002.

XXXX-XXXX/2024/5-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

[1]https://github.com/mashrur29/MIA-Evaluation

2.1 Black Box Membership Inference Attacks

In black box attacks, Adversary only observes the prediction outputs of the target model. In [11] Shokri et a. Investigated black box membership attacks against the machine learning models. Salem et al. [9] showed that even if only a single shadow model existed, membership inference attacks can perform successfully. In addition, there are some non-NN membership inference attacks that work with custom metrics on the predictions of the target mode. Leino [6] proposed a model using prediction correctness as a sign of being a member or not.

2.2 White Box Membership Inference Attacks

In white box attacks, the adversary has full access to the target machine learning model and knows the model architecture and model parameters. In [8] Nasr et al. reviewed the white box membership inference attacks. They claimed that combining the predictions of the target model together with its intermediate computations can improve the accuracy compared to black-box attacks. In this project, we demonstrate that the gap between the accuracy of black box membership inference attacks and white box membership inferences attacks is less than what claimed in the [8]

3 EVALUATING MEMBERSHIP INFERENCE ATTACK

The existing neural network-based membership attacks train classifiers to distinguish between members and non-members. However, the accuracy of such attack models largely depends on the accuracy of the classifier. Therefore, we use metrics in this work to compute the privacy risk of Machine Learning (ML) models. These metrics depend on class-dependent thresholds set using shadow training.

3.1 Defense Methods

This work uses two popular neural network-based defense methods: Adversarial Regularization [7] and MemGuard [5].

3.1.1 Adversarial Regularization. In adversarial regularization, an adversary is trained with the target model, which tends to maximize the accuracy of the membership inference attacks. On the other hand, the target model uses the output of the adversary to regularize its loss to minimize the accuracy of the attack model. So, we get a min-max objective, where an adversary is trained with the target model.

3.1.2 MemGuard. In MemGuard, the training process remains unchanged. Instead, the output of the target model is obfuscated with predefined noise to confuse the attack model. The objective is to reduce the distance between the actual target model prediction and the noisy prediction so that the final predicted class remains the same.

3.2 Trained Models

This work uses three network architectures with Tanh and ReLU activations and RNN unit. We use the Texas-100 and Purchase 100 datasets to train the models. We trained each model for 20 epochs with a batch size of 128. We used a 3 : 1 split for the train and test set. The hyperparameters were tuned using grid search. The models were trained on a single 1050ti over a span of 1 week. Figure 2, shows the results of train and test accuracy of our trained models. We obtained a reasonable test accuracy for each model, except for the RNN unit with MemGuard defense. We speculate that this is caused by insufficient hyperparameter tuning due to resource constraints. Following the work done in [12], we used class-dependant thresholds to infer membership (see Figure 1).

3.3 Evaluating SOTA Defense with NN-based Adversarial Attacks

We performed adversarial attacks on the model during the training process to evaluate the privacy leakage of the trained models (see Figure 3). On average, the privacy leakage of the models was close to random guessing. Furthermore, the models trained on the texas dataset had a relatively greater privacy leakage.

3.4 Metric-based Membership Inference Attacks

In this work, we supplement the output of neural network-based attacks with metric-based attacks. These metrics are easy to compute and are learned from the shadow models, where the neural network is trained to replicate the behavior of the target model. Following [12], we used prediction correctness, confidence, entropy, and modified entropy. Prediction correctness computes the correctness of the membership inference and prediction confidence uses the confidence of the probability with which the model infers membership. Prediction entropy computes the prediction entropy distribution, and the modified entropy incorporates the ground truth label.

3.5 Evaluating SOTA Defense with Metric-based Attacks

Table 1 shows the results from our metric-based attacks. The results show that the metric-based attacks have comparable accuracies to the NN-based attacks. Moreover, the metric-based attacks show a relatively greater privacy leakage than the NN-based attacks. In particular, the Purchase-100 with the defense has a privacy leakage of 57% confidence, and the Texas 100 with a privacy leakage of 68% confidence. This suggests the relevance of metric-based MIA attacks in determining privacy leakage as an alternative to NN-based attacks.

4 PRIVACY RISK SCORE

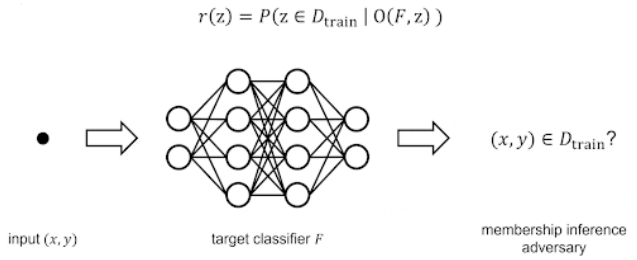

Fig. 4. Fine-grained privacy Analysis

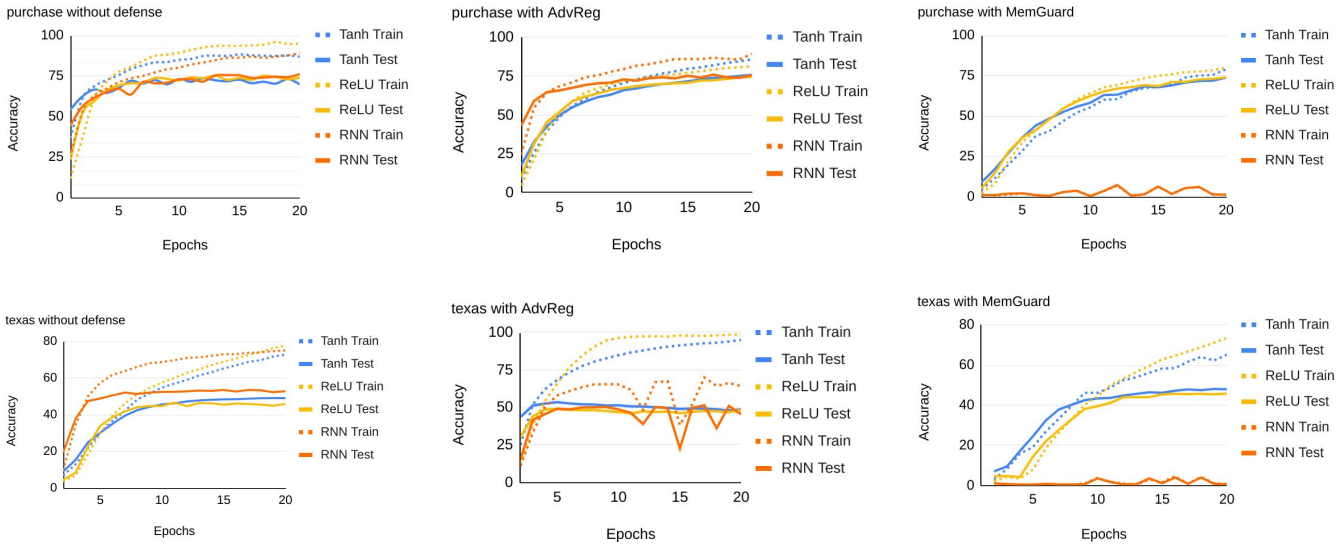

Fig. 2. Trained Models

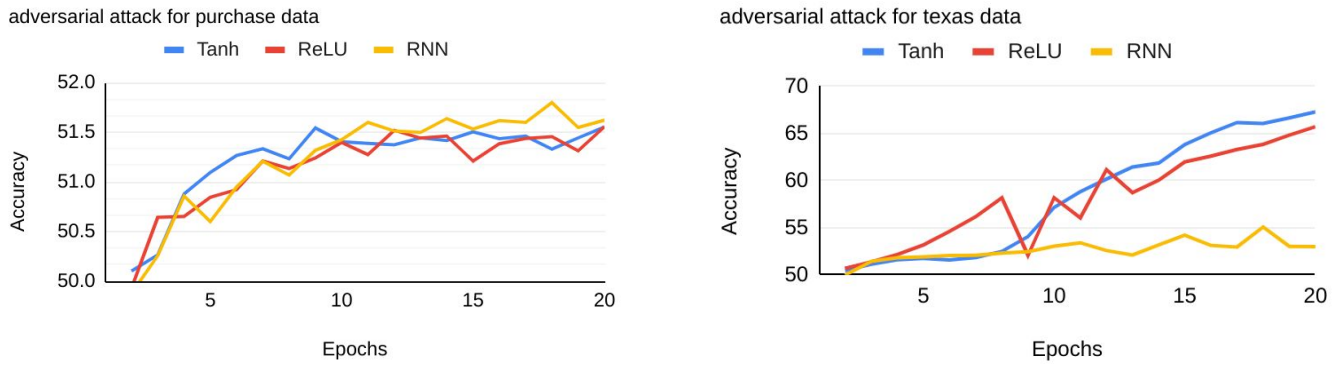

Fig. 3. Adversarial Attacks

Table 1. Benchmark of Membership Inference Attacks on neural network based defense models.

dataset	using defense?	train acc	test acc	correctness	confidence	entropy	modified entropy
Purchase 100	no	89.7	73.5	0.58	0.59	0.49	0.53
Purchase 100	yes	88.4	76.4	0.57	0.57	0.49	0.51
Texas 100	no	76.9	48.7	0.64	0.68	0.53	0.57
Texas 100	yes	76.1	53.6	0.61	0.64	0.51	0.54

Privacy analysis is not a new concept, and in [5, 7, 8, 10, 13] these works, the authors concentrate on an aggregate assessment of privacy threats by presenting overall attack accuracy or a precision-recall pair averaged across all samples. However, for the heterogeneity of the target ML model's sample, a fine-grained analysis of the privacy risk model is necessary. The privacy risk of a training member originates from the distinguishability of its model prediction behavior with non-members in membership inference attacks. Therefore, the privacy risk score of an input sample $\mathbf{z} = (\mathbf{x}, y)$ for a Machine Learning model F is defined as the posterior probability that it is from the training dataset D_{tr} after observing the target

model's behavior over that sample denoted as O(F, **z**). Figure 4 describes the process to calculate the privacy risk score.

$$r(\mathbf{z}) = P(\mathbf{z} \in D_{tr}|O(F, \mathbf{z})) \tag{1}$$

We can apply Bayes' rule and get the following equation:

$$r(\mathbf{z}) = \frac{P(\mathbf{z} \in D_{tr}).P(O(F, \mathbf{z})|\mathbf{z} \in D_{tr})}{P(O(F, \mathbf{z}))} \tag{2}$$

and,

$$P(O(F, \mathbf{z})) = P(\mathbf{z} \in D_{tr}).P(O(F, \mathbf{z})|\mathbf{z} \in D_{tr}) + P(\mathbf{z} \in D_{te}).P(O(F, \mathbf{z})|\mathbf{z} \in D_{te}) \tag{3}$$

So, we can further get:

$$r(\mathbf{z}) = \frac{P(\mathbf{z} \in D_{tr}).P(O(F, \mathbf{z})|\mathbf{z} \in D_{tr})}{P(\mathbf{z} \in D_{tr}).P(O(F, \mathbf{z})|\mathbf{z} \in D_{tr}) + P(\mathbf{z} \in D_{te}).P(O(F, \mathbf{z})|\mathbf{z} \in D_{te})} \tag{4}$$

The authors of this paper assume that their model only needs black-box access to the target machine-learning model. In a black-box membership inference attack, O(F, **z**) = F(**x**). So we can write:

$$r(\mathbf{z}) = \frac{P(\mathbf{z} \in D_{tr}).P(F(\mathbf{x})|\mathbf{z} \in D_{tr})}{P(\mathbf{z} \in D_{tr}).P(F(\mathbf{x})|\mathbf{z} \in D_{tr}) + P(\mathbf{z} \in D_{te}).P(F(\mathbf{x})|\mathbf{z} \in D_{te})} \tag{5}$$

We can see from Equation 5, the risk score depends on the $P(\mathbf{z} \in D_{tr})$, $P(\mathbf{z} \in D_{te})$, and $P(F(\mathbf{x})|\mathbf{z} \in D_{tr})$, $P(F(\mathbf{x})|\mathbf{z} \in D_{te})$. For the prior probabilities, the authors assume that the samples are either from the training set or the test set. So the probability is equal which is 0.5 probability. So, with this assumption we can write:

$$r(\mathbf{z}) = \frac{P(F(\mathbf{x})|\mathbf{z} \in D_{tr})}{P(F(\mathbf{x})|\mathbf{z} \in D_{tr}) + P(F(\mathbf{x})|\mathbf{z} \in D_{te})} \tag{6}$$

In this paper, the shadow training concept is used to find the conditional probability distribution $P(F(\mathbf{x})|\mathbf{z} \in D_{tr})$ and $P(F(\mathbf{x})|\mathbf{z} \in D_{te})$. This shadow training is performed using the following steps:

- In the first step, they train a shadow model to replicate the behavior of the target machine-learning model.
- They acquire the prediction outputs of the shadow model on shadow training and shadow test data in the second step.
- Finally, empirical evaluation of conditional distributions on shadow training and shadow test data are computed.
- Also the authors compute the distributions of model prediction across training and test data $P(F(\mathbf{x})|\mathbf{z} \in D_{tr})$ and $P(F(\mathbf{x})|\mathbf{z} \in D_{te})$ in a class-dependent way.

So, for calculating $P(F(\mathbf{x})|\mathbf{z} \in D_{tr})$, Equation 7 is used.

$$P(F(\mathbf{x})|\mathbf{z} \in D_{tr}) = \begin{cases} P(F(\mathbf{x})|\mathbf{z} \in D_{tr}), & \text{when } y = y_0 \\ P(F(\mathbf{x})|\mathbf{z} \in D_{tr}), & \text{when } y = y_1 \\ . \\ . \\ . \\ P(F(\mathbf{x})|\mathbf{z} \in D_{tr}), & \text{when } y = y_n \end{cases} \tag{7}$$

and $P(F(\mathbf{x})|\mathbf{z} \in D_{te})$ is calculated using the Equation 8.

$$P(F(\mathbf{x})|\mathbf{z} \in D_{te}) = \begin{cases} P(F(\mathbf{x})|\mathbf{z} \in D_{te}), & \text{when } y = y_0 \\ P(F(\mathbf{x})|\mathbf{z} \in D_{te}), & \text{when } y = y_1 \\ . \\ . \\ . \\ P(F(\mathbf{x})|\mathbf{z} \in D_{te}), & \text{when } y = y_n \end{cases} \tag{8}$$

Also, the authors demonstrate that by merely employing one-dimension prediction metrics like confidence and modified entropy, their suggested benchmark attacks achieve equivalent or even greater accuracy than NN-based attacks that use the entire prediction vector as features. As a result, they recommend that the multi-dimension distribution in Equation 8 be further approximated with the distribution of modified prediction entropy because employing modified entropy frequently results in the greatest attack accuracy across all benchmark assaults. Most of the time, both the modified entropy-based assault and the confidence-based attack provide the best attack performance. However, for undefended Location30 and Texas100 classifiers, the updated entropy-based approach delivers much greater attack accuracy.

$$P(F(\mathbf{x})|\mathbf{z} \in D_{tr}) = \begin{cases} P(Mentr(F(\mathbf{x}), y)|\mathbf{z} \in D_{tr}, y = y_0), & \text{when } y = y_0 \\ P(Mentr(F(\mathbf{x}), y)|\mathbf{z} \in D_{tr}, y = y_1), & \text{when } y = y_1 \\ . \\ . \\ . \\ P(Mentr(F(\mathbf{x}), y)|\mathbf{z} \in D_{tr}, y = y_n), & \text{when } y = y_n \end{cases} \tag{9}$$

and, we can write Equation 7 as:

$$P(F(\mathbf{x})|\mathbf{z} \in D_{te}) = \begin{cases} P(Mentr(F(\mathbf{x}), y)|\mathbf{z} \in D_{te}, y = y_0), & \text{when } y = y_0 \\ P(Mentr(F(\mathbf{x}), y)|\mathbf{z} \in D_{te}, y = y_1), & \text{when } y = y_1 \\ . \\ . \\ . \\ P(Mentr(F(\mathbf{x}), y)|\mathbf{z} \in D_{te}, y = y_n), & \text{when } y = y_n \end{cases} \tag{10}$$

4.1 Validation of Privacy Risk Score

We validate the effectiveness of the privacy risk score here before releasing the comprehensive data. For the target machine learning model, we first compute the privacy risk scores using the procedure in Section 4 for all training and test samples. The complete range of privacy risk scores is then divided into numerous bins, and the number of training points (n_{tr}) and test points (n_{te}) in each bin are counted. The proportion of training points ($\frac{n_{tr}}{n_{tr}+n_{te}}$) in each bin is then computed, indicating the true probability of a sample being a member. If the privacy risk score actually correlates to the chance that a sample comes from the training set of a target model, we may anticipate the actual values of privacy risk scores and the proportion of training points in each bin to closely track each other. We show the distribution of the training sample's privacy risk score without defense in Figure 5 and with AdvReg in Figure 6. We then compare the privacy risk score and attack classifier's output with the real

Table 2. Benchmark of Membership Inference Attacks by varying the threshold values on privacy risk ratings, the (precision, recall) pair of membership inference attacks is presented for each threshold value.

Threshold values on privacy risk score	1.0	0.9	0.8	0.7	0.6	0.5
Attack Precision	88.2%	84.5%	82.6%	77.0%	71.3%	66.0%
Attack Recall	1.4%	7.6%	18.7%	43.7%	70.5%	99.9%

Table 3. Evaluation of the privacy risk score on different classifiers with and without defense mechanisms.

	MemGuard			AdvReg			Without Defense		
	Tanh	ReLU	RNN	Tanh	ReLU	RNN	Tanh	ReLU	RNN
Purchase100	0.547	0.504	0.519	0.501	0.502	0.509	0.513	0.510	0.541
Texas100	0.536	0.532	0.565	0.513	0.502	0.537	0.514	0.512	0.535

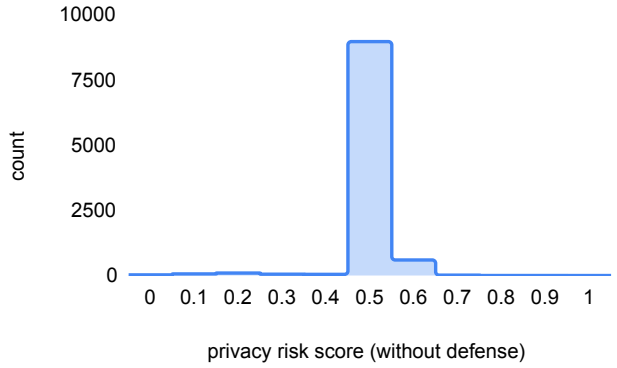

Fig. 5. Distribution of training sample's privacy risk score (without defense)

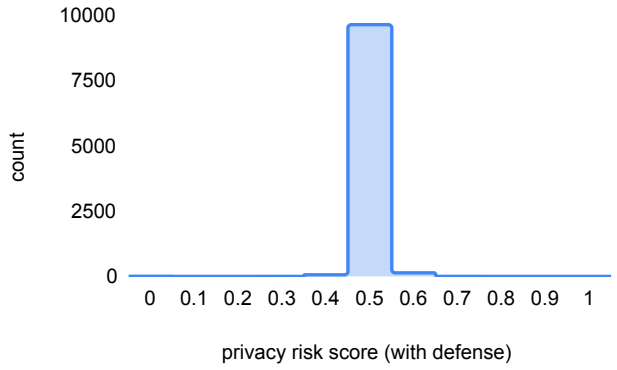

Fig. 6. Distribution of training sample's privacy risk score (with AdvReg)

probability of being a member. Figure 7 shows the comparison with the ideal case scenario. We have used the model without defense and the model with AdvReg. The red dotted line in Figure 7 represents

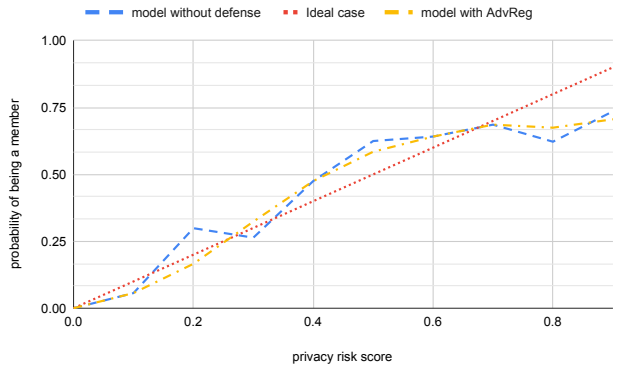

Fig. 7. Comparison of the privacy risk score with the real probability of being a member

the ideal case scenario. From this figure, we can see that privacy risk score closely aligns with the actual probability of being a member. [2]

4.2 Usage of Privacy Risk Score

From the findings in Section 4, the authors show that a data point's privacy risk score reflects its likelihood of being a member. Instead of chasing high average attack accuracy, now the adversary may discover which samples have high privacy risks and launch attacks with high confidence: a sample is inferred as a member if and only if its privacy risk score is above a particular probability threshold. We present the results with precision-recall values in Table 2. From the result, we can see that the Texas100 classifier has severe privacy risks. For example, 99.9% training members can be inferred correctly with a precision of 66.0%, and 7.6% training members can be inferred correctly with a precision of 84.5%.

[2]All the experiments in this report are done by us. We have not used any graphs or results from the actual paper

4.3 Evaluation of the Privacy Risk Score

We then evaluate the privacy risk score metric without defense and also by using MemGuard and AdvReg. We use the Purchase100 and Texas100 datasets for this evaluation. We report our results in Table 3 for Tanh, ReLU activation functions, and RNN. We can see that the privacy risk score is approximately close to 0.5. But for some cases, it is much higher than 0.5.

5 CONCLUSION

In this project, we have shown that measuring the privacy risks of membership inference privacy risk using NN-based attacks solely is not a reliable approach. We tried to benchmark the privacy risks of machine learning models using a suite of metric-based attacks including modified existing models and a newly proposed method. By using these benchmark attacks, two concepts are shown respectively:

- The defense approach proposed by Nasr et al. in [7] can only decrease privacy risks to a limited degree and it is not efficient enough.
- we showed that the performance of MemGuard which is proposed in [5] by Jia et al. is degraded with adaptive attacks

In addition, we presented a new metric called privacy risk score to evaluate and analyze individual samples' privacy risks. We showed that using a privacy risk score is a trustable approach to estimating the true likelihood of an individual sample being in the training set of a model. We also investigated the correlation between privacy risks and model properties. In conclusion, in this project, we emphasized on the importance of evaluation of privacy risks in machine learning models

REFERENCES

[1] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. 2016. Membership privacy in MicroRNA-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 319–330.

[2] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*. 267–284.

[3] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 1322–1333.

[4] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*. 1895–1912.

[5] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 259–274.

[6] Klas Leino and Matt Fredrikson. 2020. Stolen Memories: Leveraging Model Memorization for Calibrated {White-Box} Membership Inference. In *29th USENIX security symposium (USENIX Security 20)*. 1605–1622.

[7] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.

[8] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. 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)*. IEEE, 739–753.

[9] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).

[10] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[11] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[12] Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*. 2615–2632.

[13] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 241–257.

[14] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.