Journal of Information Hiding and Privacy Protection
DOI:10.32604/jihpp.2021.027871

Tech Science Press

Article

A Performance Study of Membership Inference Attacks on Different Machine Learning Algorithms

Jumana Alsubhi[1], Abdulrahman Gharawi[1] and Mohammad Alahmadi[2,*]

[1]Department of Computer Science, University of Georgia, Athens, GA, 30602, USA
[2]Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah, 23890, Saudi Arabia
[*]Corresponding Author: Mohammad Alahmadi. Email: mdalahmadi@uj.edu.sa

Received: 26 January 2022; Accepted: 03 March 2022

Abstract: Nowadays, machine learning (ML) algorithms cannot succeed without the availability of an enormous amount of training data. The data could contain sensitive information, which needs to be protected. Membership inference attacks attempt to find out whether a target data point is used to train a certain ML model, which results in security and privacy implications. The leakage of membership information can vary from one machine-learning algorithm to another. In this paper, we conduct an empirical study to explore the performance of membership inference attacks against three different machine learning algorithms, namely, K-nearest neighbors, random forest, support vector machine, and logistic regression using three datasets. Our experiments revealed the best machine learning model that can be more immune to privacy attacks. Additionally, we examined the effects of such attacks when varying the dataset size. Based on our observations for the experimental results, we propose a defense mechanism that is less prone to privacy attacks and demonstrate its effectiveness through an empirical evaluation.

Keywords: Membership inference attack; data privacy; machine learning; security

1 Introduction

Machine learning (ML) algorithms have been widely used in many real-world applications as their computational ability has been improved significantly over the years. One central limitation of ML is data dependence, in which many ML models need massive amounts of labeled data during the training process to work properly [1,2]. The data could contain sensitive information such as health-related records [3], crimes [4], and credit cards [5]. Yet, training a model with such sensitive information makes the data exposed to the membership inference attacks (MIA) [6,7].

The first membership inference attack on machine learning models is presented by Shokri et al. [8], which has shown the possibility of exposing information through an empirical investigation. Yet, the proposed approach does not use the same ML models as used in this paper and only used synthetic dataset without varying their sizes which could be the root cause issue of the MIA attack. Salem et al. [9] proposed a follow-up investigation which showed the possibilities of the MIA attack even without the knowledge of the target model and developed a defense mechanism against the attack. Yet, the experiments were not performed with dataset that have different sizes across different ML models.

The MIA attack aims to find out whether a target data point is used to train a certain ML model, which can raise security and privacy concerns [10]. In other words, given sensitive information such as an image or a text, MIA can find if that piece of information was used to train a given ML model. Some ML models are more immune to the MIA attack than others. Besides, there might be other factors that are important to make the model more robust to such a privacy attack. As a result, it is important to know which machine learning algorithms are more secure in terms of leakage of membership information. This helps us to understand which machine learning algorithms can be applied in each specific situation. In addition, increasing the dataset size can be a very effective countermeasure against membership inference.

In this paper, we investigate the performance of membership inference attacks against four machine learning algorithms, which are k-Nearest Neighbors (KNN), random forest (RF), support vector machine (SVM), and logistic regression (LR) using three datasets. Furthermore, we propose a defense mechanism for membership inference attacks by studying the effect of varying the dataset size. The datasets we utilized in our study are Breast Cancer Wisconsin dataset [11], MNIST database of handwritten digits [12], and Adult dataset [11].

In summary, we make the following contributions:

- To the best of our knowledge, we conducted the first study to compare the performance of membership inference attacks against different machine learning algorithms to determine which machine learning algorithm is more immune to such privacy attacks.
- We also investigated the relationship between the dataset size and the performance of these attacks for each machine learning algorithm.
- Our findings shed a light on new ways to defend against membership inference attacks.

The rest of the paper is organized as follows. Section 2 provides a background information about the membership inference against ML models and the datasets used in this paper. Section 3 and Section 4 present our methodology and evaluations of the four different ML algorithms using the three datasets, respectively. In Section 5, we introduce the defense mechanism. Section 6 concludes the paper and discusses potential future work.

2 Background

In this section, we first introduce the definition of membership inference attacks. Then, we present the datasets we used for our performance evaluation, and provide background information for the models we used during our evaluations.

2.1 Membership Inference Attack

Membership inference attack has been identified as a possible confidentiality violation and privacy hazard to training data by organizations such as the ICO (UK) and NIST (US) [13]. Membership inference attack in ML arises when an adversary aims to find out whether a target data point is used to train a specific ML model or not, and the adversary has black-box access to the ML model [14,15]. The attack model is a binary classifier, which outputs 0 or 1. The output 0 means the data point is not a member while 1 means that the data item is a member of the ML model's training dataset. Thus, this attack is considered one of the simplest privacy attacks [16].

2.2 Dataset Description

In this paper, we utilize three different datasets ranging from image to text to conduct our experiments. Two of those datasets were used in a previous study [8] which are the MNIST and Adult datasets, whereas we added a new dataset Breast Cancer Wisconsin, to investigate the performance of the MIA attack on a smaller dataset. The first dataset is the Breast Cancer Wisconsin dataset, which has 569 samples total. It has 30 attributes and 2 classes, Benign and Malignant, with 212 and 357 samples each, respectively. The second dataset is the MNIST database of handwritten digits. It has around 180 samples of images per class, so in total it has 1797 samples with 625 dimensions. Finally, we used the Adult dataset, which has 48,842 samples with 14 attributes, which is used to determine whether a person earns over 50 K a year. We intentionally selected datasets with different sizes to explore the performance of membership inference attacks against different machine learning algorithms.

2.3 K-Nearest Neighbors (KNN)

KNN seeks to discover the k training samples closest to a new element in terms of distance and then predicts the label of that new element based on the k-nearest points. Any similarity function can be used to calculate the distance. Despite its simplicity, KNN is typically successful in classification situations with a highly irregular decision boundary [17].

2.4 Random Forest (RF)

RF is a collection of many independent decision trees, each of which is constructed from a sample selected from the training set with replacement. The split that is chosen when dividing a node during tree construction is no longer the optimal split across all features. Instead, the best split among a randomly selected subset of the features is chosen. As a result of this randomness, the forest's bias usually rises significantly [18].

2.5 Support Vector Machine (SVM)

SVM is a supervised machine learning technique that can be used for binary classification, regression analysis, and other tasks in multidimensional data spaces, such as outlier detection. In a high-dimensional feature space, an SVM looks for a hyperplane that separates two classes of data by the greatest possible margin. The hyperplane parameters may be mathematically demonstrated to be dependent only on a subset of the training samples, known as support vectors. A test sample is first projected to the feature space, then assigned a class based on which side of the hyperplane it lies on [19].

2.6 Logistic Regression (LR)

In predictive modeling, LR is used to examine big datasets where one or more independent variables can influence the outcome. The result is represented as a dichotomous variable with two possible outcomes. In essence, logistic regression calculates the mathematical probability that a given instance belongs to a certain class or not [20].

3 Methodology

In order to determine the vulnerability of different machine learning algorithms to membership inference attacks, we explore the performance of such attacks on different machine learning models. We first implemented the membership inference attacks using Python. There are many ways to construct the attack model. However, the one used in this paper is the attack model proposed by [9] in the third adversary, which is based on a threshold. It works without any shadow model, and no training procedure is required. The attack only relies on the outcomes obtained from the target model when querying it with target data points. This attack model is implemented as an unsupervised binary classification. Firstly, we obtain the target model's output posteriors after querying the target data point. The highest posterior is then extracted and compared against a threshold. If the answer is yes, then we predict the data point is a member of the training set of the target model and vice versa. Moreover, we used a threshold-choosing method in which the top 10 percentile of the data points output can serve as a good threshold since it is shown that the top 10 percentile generalizes across all the datasets. Thus, we used the top 10 percentile for each dataset to automatically determine the value of the threshold. This simple attack can achieve practical inference. Therefore, in this paper, we use this method to perform membership inference attacks in ML. We also collected the datasets and applied several preprocessing techniques to clean the data and make it proper for the ML model. Then, KNN, random forest, SVM, and logistic regression are used to evaluate the performance of the attack on these ML models. Additionally, in order to see the effect of increasing the size of the training data, we used 60%, 80%, and 100% of each of the three datasets when training the aforementioned ML models. Fig. 1 shows an overview of the overall methodology.

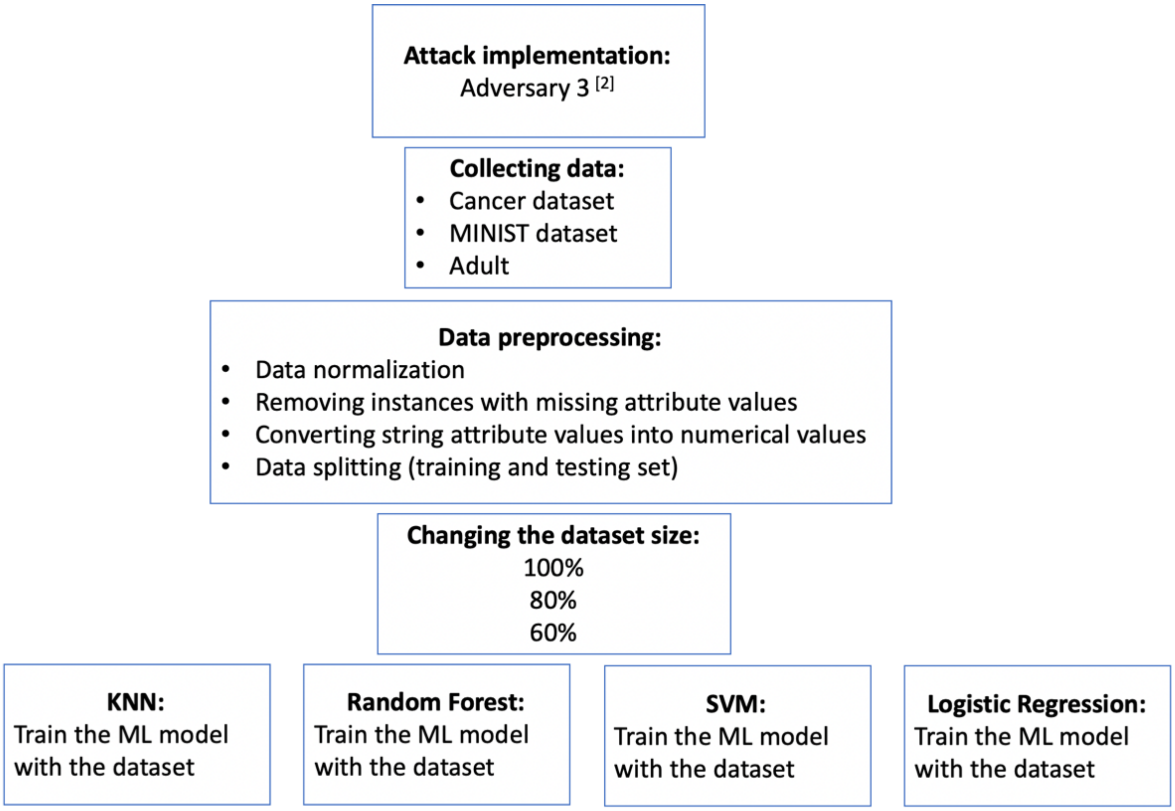
Attack implementation:
Adversary 3 [2]
Collecting data:
• Cancer dataset
• MINIST dataset
• Adult
Data preprocessing:
• Data normalization
• Removing instances with missing attribute values
• Converting string attribute values into numerical values
• Data splitting (training and testing set)
Changing the dataset size:
100%
80%
60%
KNN:
Train the ML model with the dataset
Random Forest:
Train the ML model with the dataset
SVM:
Train the ML model with the dataset
Logistic Regression:
Train the ML model with the dataset

Figure 1: Block diagram of the overall methodology

4 Evaluation

By examining the impact of the membership inference attacks on different machine learning models, namely KNN, random forest, SVM, and logistic regression, as shown in Table 1 to Table 4, we can see which machine learning algorithms are more immune to such attacks. The results show that random forest is the most vulnerable ML model amongst the other models with an average attack accuracy of more than 95%. On the other hand, SVM and KNN are considered the most immune ML models to such attacks with an average attack accuracy of about 78%.

Furthermore, Fig. 2 to Fig. 4 show the performance of membership inference attacks using different sizes of the datasets against different ML models using Cancer, MNIST, and Adult datasets, respectively. It can be clearly seen that the performance of membership inference attacks against different ML models increases as the size of the training dataset decreases. Experiments on multiple datasets show that increasing the dataset size can be a very effective countermeasure against membership inference. For example, on the Adult dataset, increasing the size of the dataset used for training the random forest model from 30% to 100% of data can decrease the performance of the attack from 0.99 precision and 1.00 recall to 0.75 and 0.93, respectively. Also, the attack accuracy increases from 80% to 99%.

Table 1: The experimental results for K-Nearest Neighbor (KNN) (Average of 10 Runs)

Target classifier	Dataset	Dataset size	Accuracy	Attack accuracy	Attack precision	Attack recall
KNN	Cancer	100	98.28	74.86	84.86	99.01
	Cancer	80	98.53	75.83	84.93	100
	Cancer	60	95.26	75.99	84.73	100
	MINST	100	98.4	74.95	74.95	75.95
	MINST	80	98.88	86.64	84.88	97.95
	MINST	60	99.16	98.37	99.17	100
	MINST	30	97.3	81.11	80.88	98.9
	Adult	100	82.07	70.9	74.98	99.4
	Adult	80	83.29	71.74	74.91	99.97
	Adult	60	82.85	74.33	79.62	100
	Adult	30	85.3	94.97	80.81	100

Table 2: The experimental results for RF (Average of 10 Runs)

Target classifier	Dataset	Dataset size	Accuracy	Attack accuracy	Attack precision	Attack recall
RF	Cancer	100	99.75	96.66	95.73	99.09
	Cancer	80	98.53	75.83	84.93	100
	Cancer	60	99.36	97.86	96.61	100
	MINST	100	98.52	98.99	96.88	100
	MINST	80	98.48	98.07	98.53	99.81
	MINST	60	99.02	99.08	98.75	100
	MINST	30	99.04	99.6	98.17	100
	Adult	100	99.97	99.9	100	100
	Adult	80	88.57	80.71	75	93.35
	Adult	60	89.39	82.67	95.37	95.9
	Adult	30	91.3	98.7	97.42	100

Table 3: The experimental results for Support Vector Machine (SVM) (Average of 10 Runs)

Target classifier	Dataset	Dataset size	Accuracy	Attack accuracy	Attack precision	Attack recall
SVM	Cancer	100	97.38	68.4	97.72	84.74
	Cancer	80	97.08	78.54	90.08	63.74
	Cancer	60	97.81	75.26	81.98	67.56
	MINST	100	99.89	78.01	77.32	95.22
	MINST	80	99.7	78.89	78.49	99.29
	MINST	60	99.85	80.7	77.88	99.87
	MINST	30	99.54	86.34	89.55	99.6

	Adult	100	80.08	74.47	74.99	99.97
	Adult	80	80.08	74.76	74.99	100
	Adult	60	79.99	80.99	88.32	100
	Adult	30	81.47	89.81	90.84	100

Table 4: The experimental results for Logistic Regression (LR) (Average of 10 Runs)

Target classifier	Dataset	Dataset size	Accuracy	Attack accuracy	Attack precision	Attack recall
LR	Cancer	100	97.14	92.96	92.4	93.88
	Cancer	80	96.42	93.84	97.83	99.8
	Cancer	60	98.21	95.5	99.64	99.52
	MINST	100	99.55	88.68	88.74	88.93
	MINST	80	99.81	97	96.61	98.91
	MINST	60	99.97	99.03	99.74	100
	MINST	30	99.89	99.59	100	100
	Adult	100	80.05	70	70	100
	Adult	80	81.81	72.51	75.1	100
	Adult	60	83.68	75.02	75.01	100
	Adult	30	83.1	80.99	80.6	99.77

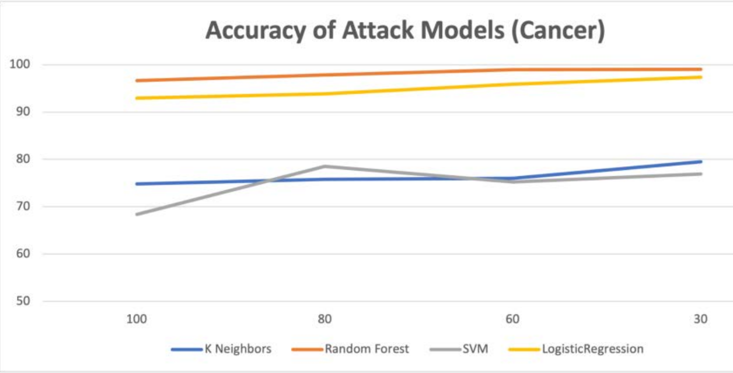

Figure 2: Accuracy of the attack model (Cancer dataset)

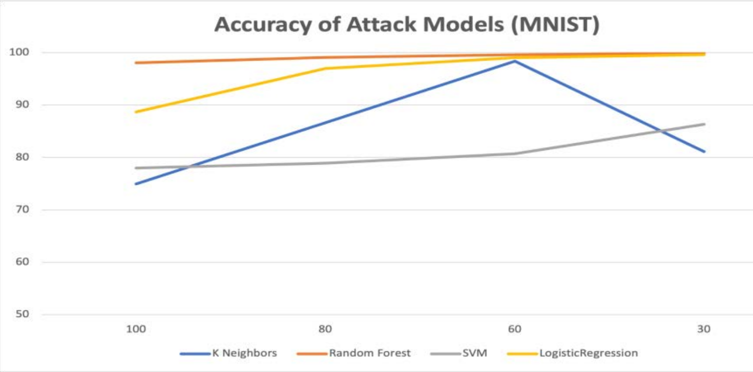

Figure 3: Accuracy of the attack model (MNIST dataset)

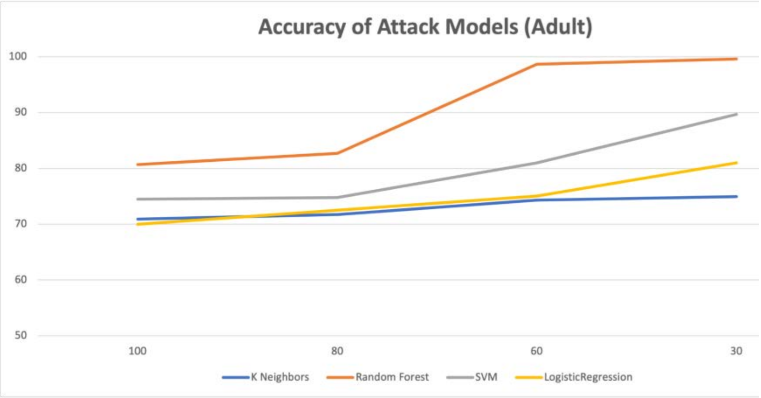

Figure 4: Accuracy of the attack model (Adult dataset)

5 Defense

Our findings illustrate a new potential way to defend against membership inference attacks. Through extensive experiments, we show that increasing the dataset size can be an effective defense mechanism. In the black-box setting, the success of membership inference attacks is substantially linked to the target model's overfitting [21]. Therefore, to defend against membership inference attacks, we may increase the dataset size when training different ML models. More data helps in avoiding overfitting, and there are many ways to address the problem of fewer data [22,23]. Several techniques have been proposed to increase the data size without collecting more data such as data augmentation with flips, crops, and brightness [24].

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. Best, J. Ott and E. J. Linstead, "Exploring the efficacy of transfer learning in mining image-based software artifacts," *Journal of Big Data*, vol. 7, no. 1, pp. 1–10, 2020.

[2] K. G. Liakos, P. Busato, D. Moshou, S. Pearson and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, pp. 2674, 2018.

[3] S. J. Mooney and V. Pejaver, "Big data in public health: Terminology, machine learning, and privacy," *Annual Review of Public Health*, vol. 39, pp. 95–112, 2018.

[4] S. Aggarwal, G. Dorai, U. Karabiyik, T. Mukherjee, N. Guerra *et al.*, "A targeted data extraction system for mobile devices," in *IFIP Int. Conf. on Digital Forensics*, Springer, pp. 73–100, 2019.

[5] R. Basnet, S. Mukkamala and A. H. Sung, "Detection of phishing attacks: A machine learning approach," in *Soft Computing Applications in Industry*, Springer, pp. 373–383, 2008.

[6] S. Truex, L. Liu, M. E. Gursoy, L. Yu and W. Wei, "Demystifying membership inference attacks in machine learning as a service," in *IEEE Membership Inf. Attacks and Defenses on Machine Learning Models*, 2018.

[7] Y. Long, V. Bindschaedler and C. A. Gunter, "Towards measuring membership privacy," arXiv preprint arXiv:1712.09136, 2017.

[8] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Sym. on Security and Privacy (SP)*, IEEE, pp. 3–18, 2017.

[9] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz *et al.*, "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models," arXiv preprint arXiv:1806.01246, 2018.

[10] Z. Li and Y. Zhang, “Label-leaks: Membership inference attack with label,” arXiv preprint arXiv:2007.15528, 2020.

[11] LECUN. [Online]. Available: http://yann.lecun.com/exdb/mnist/, 2021.

[12] UCI. [Online]. Available: https://archive.ics.uci.edu/ml/index.php, 2021.

[13] J. Ye, A. Maddi, S. K. Murakonda and R. Shokri, “Enhanced membership inference attacks against machine learning models,” arXiv preprint arXiv:2111.09679, 2021.

[14] B. Hitaj, G. Ateniese and F. Perez-Cruz, “Deep models under the gan: Information leakage from collaborative deep learning,” in *Proc. of the 2017 ACM SIGSAC Conf. on Computer and Communications Security*, pp. 603–618, 2017.

[15] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li *et al.*, “Socinf: Membership inference attacks on social media health data with machine learning,” in *IEEE Trans. on Computational Social Systems*, vol. 6, no. 5, pp. 907–921, 2019.

[16] C. A. Choquette-Choo, F. Tramer, N. Carlini and N. Papernot, “Label-only membership inference attacks,” in *Int. Conf. on Machine Learning*, pp. 1964–1974, 2021.

[17] A. Khan, B. Baharudin, L. H. Lee and K. Khan, “A review of machine learning algorithms for text-documents classification,” *Journal of Advances in Information Technology*, vol. 1, no. 1, pp. 4–20, 2010.

[18] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[20] R. E. Wright, “Logistic regression,” 1995.

[21] L. Song, R. Shokri and P. Mittal, “Membership inference attacks against adversarially robust deep learning models,” in *2019 IEEE Security and Privacy Workshops (SPW)*, IEEE, pp. 50–56, 2019.

[22] W. M. van der Aalst, V. Rubin, H. Verbeek, B. F. van Dongen, E. Kindler *et al*., “Process mining: A two-step approach to balance between underfitting and overfitting,” *Software & Systems Modeling*, vol. 9, no. 1, pp. 87–111, 2010.

[23] H. Jabbar and R. Z. Khan, “Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study),” *Computer Science, Communication and Instrumentation Devices*, pp. 163–172, 2015.

[24] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.