4th IEEE International Conference on AI in Cybersecurity (ICAIC), 5-7 February 2025, University of Houston, Texas, USA

Quantifying Classification Metrics in Black Box Membership Inference Attacks

Neil Dixit
Palo Alto, CA, USA
neildixit10@gmail.com

***Abstract*— This paper investigates how attackers can adjust the thresholds of their classification to optimize classification metrics in their membership inference attacks. By using Monte Carlo methods, we modeled the distribution of scoring functions for both True Positive and True Negative values. Then we calculated classification metrics (FPR, FNR, MA, and AR) as a function of threshold value and found a sigmoid relationship, verified by linearizing our data (R^2=0.997). From this, we found relationships for most metrics as a function of threshold value and how to optimize them. We found that the relationship for FPR and FNR as a function of threshold value, T, follows a translated sigmoid function. Our findings provide information on methods attackers can use to fine-tune their thresholds to optimize their attack with minimal computational power. Our findings demonstrate the importance of altering aggregate statistics with Differential Privacy to mitigate Membership Inference Attacks. The major limitation of our model is that the attacker needs to know the underlying distribution of data, which we have assumed is Gaussian. In addition, we have only taken the case where data is binary. Additional research is needed to adjust for or reject these limitations.**

Keywords—Membership Inference Attacks, Classification, Machine Learning, Data Privacy, Metrics, Data Science

I. Introduction

Data privacy is a relatively young field and membership inference attacks are a new topic in academia. In this section we review and summarize prior literature regarding the topic, with emphasis on the attacker's hypothesis testing for classification.

A. Membership Inference Attacks

A membership inference attack is when an attacker attempts to infer whether a person is in a dataset. Usually, the attacker has an auxiliary information vector for a person's data, denoted as **a** [1], [2]. Each component of the vector represents a value for a given attribute. The attacker needs another vector, **p**, which represents the aggregate statistics for a large-scale population [1]. The attacker also needs a third vector, **y**, which represents the aggregate data for each attribute for the specific database they are examining [1].

By cross-referencing as many attributes as possible between aggregate statistics of the group, **y**, the auxiliary information, **a,** and large-scale population aggregate statistics, **p**, the attacker can compare the attributes known in all 3 pieces of data [1].

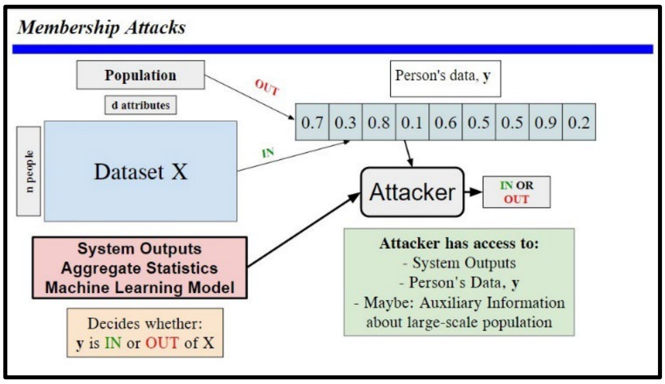


Figure 1: Visual Representation of how a Membership Inference Attack is carried out. Slide based on one from [1].

B. Metrics used in Machine Learning Classification

When attackers use Classification, they are not going to get every Classification correct. The success of the Classification can be measured in several ways, outlined below [3]-[7].

False Positive Rate (FPR):

$$FPR = P(Classified\ IN\ /\ OUT)\ (1)$$

This determines how likely the attacker is to classify data as IN the dataset when the data is not IN the dataset.

False Negative Rate (FNR):

$$FNR = P(Classified\ OUT\ /\ IN)\ (2)$$

The converse of FNR, this determines how likely the attacker is to classify data as not IN the dataset when it is IN the dataset.

Attack Precision (AP):

$$AP = P(IN\ /\ Classified\ IN)\ (3)$$

Attack Precision gives the probability that, given the attacker classifies the data as IN the dataset, the data is IN the dataset.

Attack Recall (AR):

$$AR = P(Classified\ IN\ /\ IN)\ (4)$$

Attack Recall indicates the proportion of the target dataset **y** that is classified correctly as IN for the attack.

Membership Advantage (MA):

$$MA = AR - FPR\ (5)$$

The use of randomized guessing (classify IN with 50% chance, OUT with 50% chance) would result in an FPR of 50% and Recall of 50%. In this case Membership Advantage

979-8-3315-1888-2/25/$31.00 ©2025 IEEE

= 0. MA is a measure of how much better the classification is than pure guessing.

F1-Score:

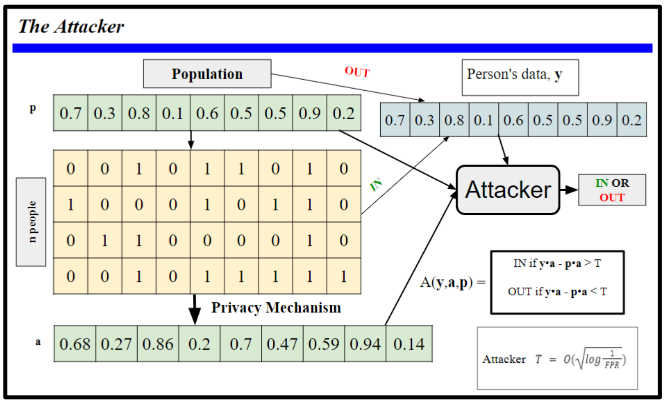
$$F_1 - score = \frac{2 \cdot AR \cdot AP}{AR + AP} \quad (6)$$

The F1-score is the harmonic mean of AR and AP. This considers the optimization of both AP and AR simultaneously.

C. Scoring Functions in Membership Inference Attacks

A scoring function is to give each vector, **a,** a score-based on which the attacker classifies the data [1]-[2]. This is usually done by choosing a threshold value, T, such that if the scoring function returns a value more than T, the attacker classifies **a** as IN and otherwise the attacker classifies **a** as OUT [1]-[2].

There have been other uses of the idea of scoring functions, such as using a scoring function to establish a confidence interval for the classification rather than a direct classification, employing probabilistic and entropy-based methods [3], [8].


The Attacker

p	0.7	0.3	0.8	0.1	0.6	0.5	0.5	0.9	0.2
	0	0	1	0	1	1	0	1	0
	1	0	0	0	1	0	1	1	0
	0	1	1	0	0	0	0	1	0
	0	0	1	0	1	1	1	1	1
a	0.68	0.27	0.86	0.2	0.7	0.47	0.59	0.94	0.14

Person's data, **y**: 0.7 | 0.3 | 0.8 | 0.1 | 0.6 | 0.5 | 0.5 | 0.9 | 0.2

A(y,a,p) = IN if **y•a** - **p•a** > T; OUT if **y•a** - **p•a** < T

Attacker $T = O(\sqrt{log\frac{1}{FPR}})$

Figure 2: Detailed diagram of the method of Membership Inference, introducing the Attack Threshold Value, T, and the Scoring Function. Based on a slide from [1].

D. Creating Synthetic Data for Data Privacy Research

Testing attack methods in the field of data privacy is difficult as researchers do not have access to uncensored, large-scale datasets. Hence, it is not unusual for researchers to create synthetic datasets for research [10].

E. Modeling Datasets as Binary

In real life, many attributes are binary in datasets and if not binary, many of the rest are discrete values, Therefore, it is reasonable to start research by modeling datasets as binary [1]-[3], [8], [10].

F. Metrics as a Function of Threshold Value

Depending on the threshold value used for the classification, the different metrics (FPR, FNR, AP, AR, MA, F1-score) will vary. The asymptotic nature of the relationship between the FPR and the value of the threshold is as follows [1]-[2]:

$$T = O\left(\log\left(\sqrt{\frac{1}{FPR}}\right)\right) \quad (7)$$

Manipulating this yields the following:

$$FPR = O\left(e^{-T^2}\right) \quad (8)$$

We also note the following:

$$AR = P\left(Classified\ IN\ \ / IN\right)$$

$$= 1 - P\left(Classified\ OUT\ \ / IN\right)$$

$$\Rightarrow AR = 1 - FNR \quad (9)$$

Using the definition of MA:

$$MA = AR - FPR = 1 - FNR - FPR$$

$$\Rightarrow MA = 1 - (FPR + FNR) \quad (10)$$

In addition, prior literature has found that the upper bound for the Membership Advantage (MA) and Attack Precision (AP) follows a sigmoid relationship [6]- [7], [9]-[10].

II. Further Background

If the vector **a** is very similar to the population vector **p** and is very different to the group data **y** then it's quite unlikely that **a** is IN the dataset. However, if vice versa is true then we would expect that Alice's data is IN the group [1]-[2].

Conveniently, the inner product of two vectors is a measure of how similar two vectors are. Based on the definition of the inner product, it follows that the higher the value of **(y - p)•(a - p),** the more likely Alice is in the dataset [1]-[2]. We can rewrite **(y - p)•(a - p)** as **y•a** - **p•a** + constant terms. Therefore, the attacker will only use **y•a** - **p•a** for the classification. Therefore, the higher the value of **y•a** - **p•a**, the more likely **a** is IN the dataset.

By setting a boundary or threshold value, T for the inner product to determine whether **a** is IN or out of the dataset, the attacker can classify data [1]-[2]. However, the attacker needs to determine the value of T to use for the classification and would want to maximize their one of the metrics of their classification [1]-[3].

III. Aim Of The Paper

In this paper, we suggest a direct relationship between FNR and FPR and T. Specifically we claim that FPR(T) and FNR(T) are sigmoid relationships. We also show how attackers can use this information to use optimization techniques to minimize the FNR and FPR. We then show that the attacker can very easily optimize each metric.

IV. Materials and Methods

We use a Monte Carlo simulation to find the distribution of the score function for data that is both IN and OUT of the dataset. We can take a normalized cumulative distribution function to represent FPR and FNR from this. Below we outline the steps needed to carry the Monte Carlo simulation:

Method to run Monte Carlo simulation:

1. Initialize vector **p** to all zeros.

2. Assign each attribute of **p** a random value between 0 and 1.

3. Add some small but nonnegligible Gaussian noise to each attribute of **p**, and define this as the vector **y.**

4. Define a function that takes in the vectors **y**, **a** and **p** and returns **y•a** - **p•a.**

5. Add some small but nonnegligible Gaussian noise to each attribute of **y**, and define this as the vector **a**.

6. Find the value of **y•a** - **p•a** and append it to a list of values.

7. Repeat Steps 5 and 6 many times to create many examples and plot the values on a histogram.

8. Add some small but nonnegligible Gaussian noise to each attribute of **p and** define this as the vector **a’**.

9. Find the value of **y•a’** - **p•a’** and append it to a list of values (separate from those for **a**).

10. Repeat Steps 8 and 9 many times to create many examples and plot the values on a histogram.

11. For many values of T, find the fraction of vectors with a score below T for both datasets, and make plots. These will approximate cumulative probability distribution functions.

V. Results

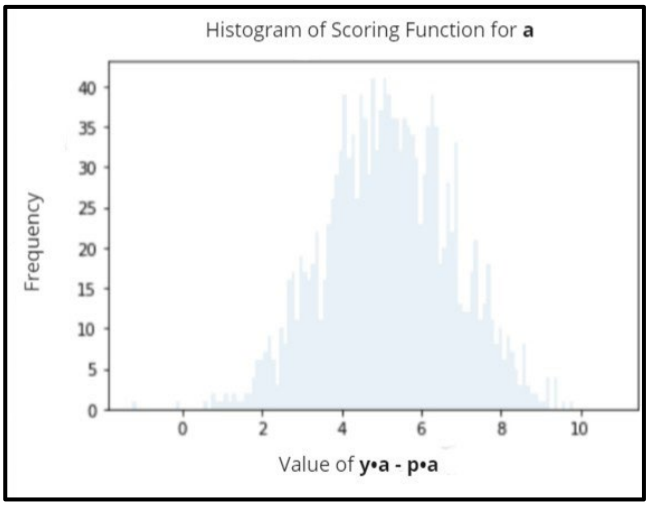


Figure 3: Histogram of Scoring Function Output, difference in inner products, for all vectors, **a**, in the dataset. d=1000, n=900.

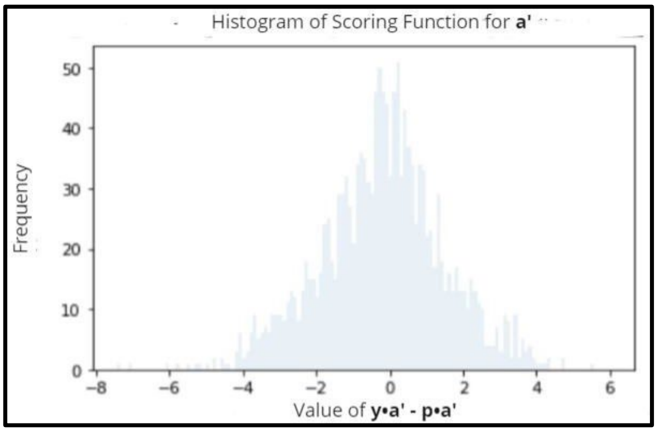


Figure 4: Histogram of Scoring Function Output, difference in inner products, for all vectors, **a’**, not in the dataset. d=1000, n=900.

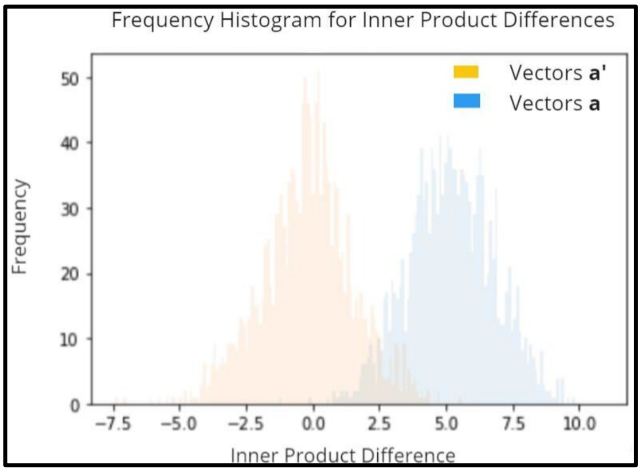


Figure 5: Histogram for scoring function, difference in inner products, on one graph for all vectors, **a**, in the dataset and all vectors, **a’**, not in the dataset. d=1000, n=900.

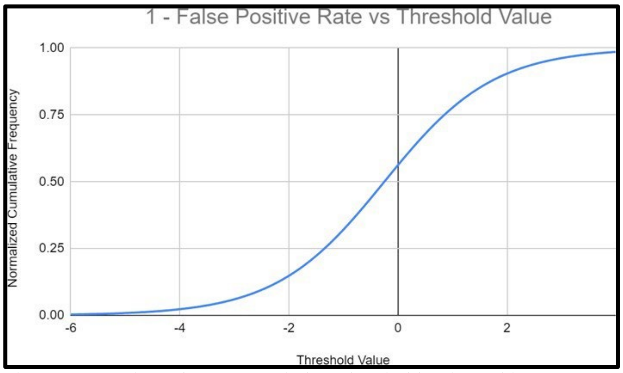


Figure 6: Fraction of vectors **a’** below threshold value. This represents a normalized cumulative frequency graph, which is equivalent to 1 – False Positive Rate.

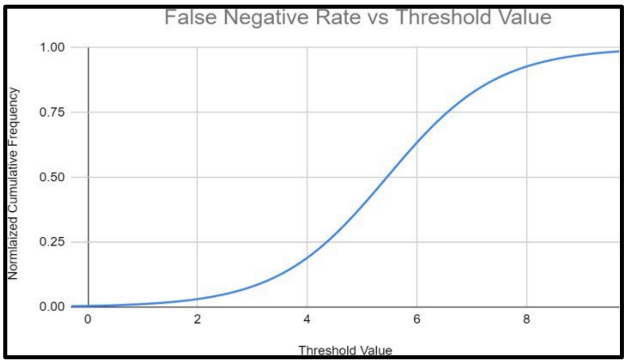


Figure 7: Fraction of vectors **a** below threshold value. This represents a normalized cumulative frequency graph, which is equivalent to False Negative Rate.

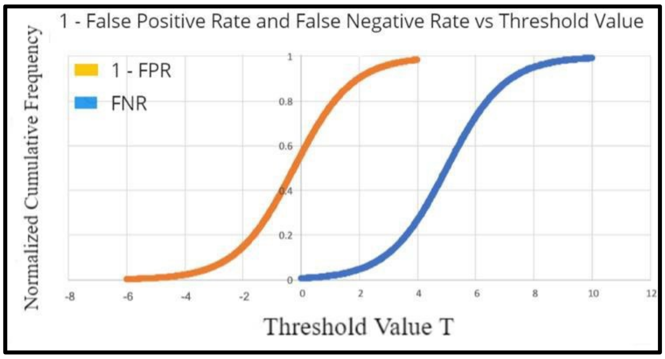


Figure 8: Scatter Plot of Normalized Cumulative Frequency Plot (proportion of vectors below Threshold Value, T).

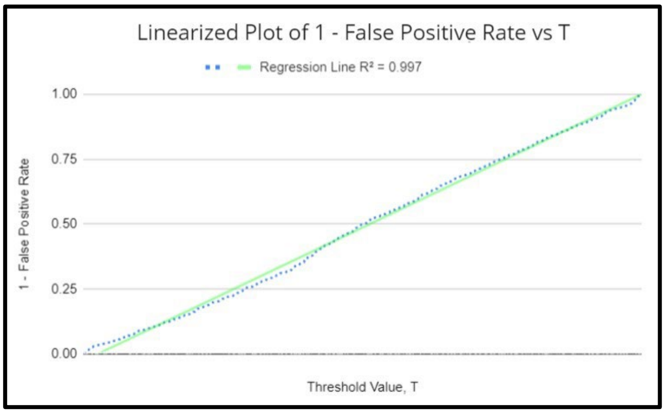


Figure 9: When graphing 1 - FPR vs σ(T $- m_1$), we found a near perfect bijection. ($R^2 = 0.997$). d=1000, n=900.

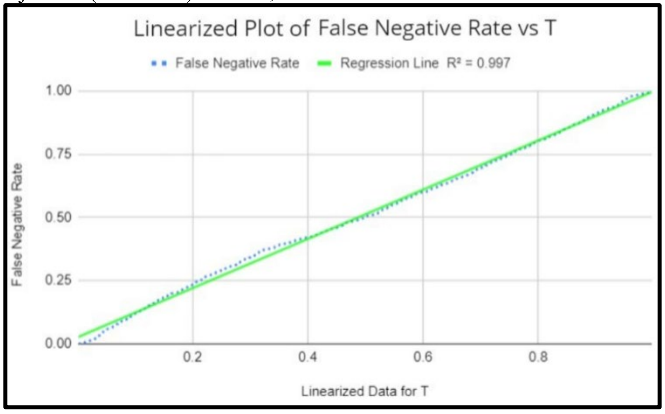


Figure 10: When graphing FNR vs σ(T $- m_2$), we found another near perfect bijection. ($R_2 = 0.997$ again). d=1000, n=900.

VI. Results Breakdown

Figures 3, 4 and 5 show the histograms of the score functions. As expected, vectors IN the dataset generally have a higher score than those OUT of the dataset. However, as Figure 5 shows, there is some overlap, meaning there is no threshold value that is 100% effective. Figures 6, 7 and 8 show approximate cumulative probability functions. These represent 1 – FPR(T) in Figure 6, FNR(T) in Figure 7, and both on top of each other in Figure 8. As expected, the 1 – FPR curve is centered at a lower value than FNR. Upon visual inspection of the results, the functions appear to follow a sigmoid relationship.

We define m_1 as the value of T for which FPR=0.5, which is the center of the cumulative probability curve in Figure 6. Likewise, we define m_2 as the value of T for which FNR=0.5, at the center of the cumulative probability curve in Figure 7. As expected, we have $m_1 < m_2$. If we assume that we have a sigmoid function, we get the following 2 equations:

$$FPR(T) = 1 - \sigma\left(k_1\left[T - \frac{m_1}{k_1}\right]\right) \quad (11)$$

$$FNR(T) = \sigma\left(k_2\left[T - \frac{m_2}{k_2}\right]\right) \quad (12)$$

where k_1 and k_2 are some constants.

Given the nonlinear nature of the sigmoid function and how the two functions are very similar but just translated, we can conclude by observation that k_1=k_2.

Now, we verify that the functions are in fact sigmoid. We apply a sigmoid function to with the given shifts in Equations 12 and 13, and plot 1 – FPR and FNR. We find that we have a near perfect one-to-one correspondence for both cases. When running linear regression, we get an equation of *y*=*x* and R^2=0.997 in both cases as seen in Figures 9 and 10. These strong results confirm that the relationship is in fact sigmoid.

VII. Applications of Results

Now we see how the attacker can use the relationships for FPR(T) and FNR(T) to their advantage. We consider the optimization of FPR, FNR, AR, MA, AP, and F_1-score.

A. Optimizing FPR

Rearranging Equation 11 yields the following

$$T(FPR) = m_1 + \ln\left(\frac{1-FPR}{FPR}\right) \quad (13)$$

Therefore, the attacker can set T to achieve any desired FPR.

B. Optimizing FNR

Rearranging Equation 12 yields the following:

$$T(FNR) = m_2 - \ln\left(\frac{1-FNR}{FNR}\right) \quad (14)$$

Hence, just as with FPR, the attacker can set T to achieve any desired FNR

C. Optimizing AR

Using Equation 9 and Plugging into Equation 12 yields the following:

$$T(AR) = m_2 + \ln\left(\frac{1-AR}{AR}\right) \quad (15)$$

Therefore, like with FPR and FNR, the attacker can set T to achieve any desired AR.

D. Optimizing MA

Using Equations 10, 11, and 12, we get that

$$MA(T) = \sigma\left(k\left[T - \frac{m_1}{k}\right]\right) - \sigma\left(k\left[T - \frac{m_2}{k}\right]\right) \quad (16)$$

Given that this function is nonlinear, it isn't possible to optimize it through analytical methods. However, it is fairly simple to use numerical methods, such as the 1st Derivative Test and the Newton-Raphson method to get a very good approximation of the maximum MA, and the corresponding value of T.

E. *Optimizing AP*

The main problem with trying to optimize AP is that, as an attacker, we need to know the number of classifications that were successful. This would require the attacker having access to the dataset, which defeats the entire purpose.

We consider a rewrite of Equation 3 with Bayes' Theorem as follows:

$$AP = P(IN|Classified\ IN) \quad (3)$$

$$= P(Classified\ IN|IN) \times \frac{P(IN)}{P(Classified\ IN)} \quad (17)$$

$$= AR \times \frac{P(IN)}{P(Classified\ IN)} \quad (18)$$

We note that $P(IN)$ is a constant, so now we consider $P(Classified\ IN)$. For each value of T, we can find the probability that vectors IN and OUT of the dataset are classified IN, from which we can find $P(Classified\ IN)$ if and only if we know the ratio of the # of our vectors that are IN to OUT. The problem is that the attacker will never have this information ahead of time. Hence, optimizing AP isn't feasible with our model.

F. *Optimizing F_1-score*

Because the formula for F_1-score involves AP, and we can't calculate it, optimizing F_1-score is also not feasible.

VIII. Discussion

Using basic but powerful techniques, an attacker can use membership inference attacks through classification methods and in doing so can easily adjust their thresholds based on their purpose. They can set the boundary value for their classification, T, to a certain value by using optimization techniques or simple equations and can theoretically use that information to their benefit and the detriment of the wider community. We found that the relationship between False Positive and False Negative Rates and the threshold value T is a sigmoid function, supporting prior literature [7]-[10]. In addition, we have found expressions for other metrics such as Attack Recall (AR), Attack Precision (AP), Membership Advantage (MA) and F_1-score in terms of FPR and FNR. Currently, attackers cannot use any of the methods suggested as of right now, since the models used are simplistic and are reliant on some unrealistic assumptions. Examples include assuming that the underlying distribution of data being Gaussian, data being binary, and the attacker having data for many overlapping attributes with published aggregate statistics. Given real life datasets are usually large, they often converge to common assumptions, such as data distributions being Gaussian per the central limit theorem. For population scale aggregate statistics, attackers can use censuses, but governments use many techniques to distort aggregate statistics to render their use futile. Despite these limitations, the study provides an excellent model of an advanced black box membership inference attack that can easily be generalized to fit more realistic scenarios.

IX. Conclusion

Data privacy is a young, growing field and one of the most important parts of data privacy research is simulating attacks to see their effectiveness. Prior findings have suggested that when carrying out membership inference attacks, attackers are able to adjust parameters to control their error rates when using classification techniques in Machine Learning [1]. The extent to which attackers can successfully do so is still unclear, but we have shown that doing so would theoretically require very little computational power. Doing research on privacy is difficult as there are very few open large real-life databases available to test research out on. More research is needed to verify the assumption that underlying distributions are Gaussian, as this is the main assumption that our method uses. All other assumptions are for simplicity, and the method can be easily modified to avoid needing them. The hope is that ensure that defense and privacy techniques make databases immune to black box membership inference attacks. The results of this paper demonstrate that with minimal computational power, attackers can easily fine tune their classification to the detriment of society. This strongly emphasizes the need for techniques such as cryptography, secure multiparty computation and differential privacy, as this prevents the attackers from even accessing meaningful aggregate statistics in the first place.

Acknowledgment

I would like to thank Dr. Wanrong Zhang, Professor of the John A. Paulson School of Engineering and Applied Sciences at Harvard University and Ms. Fatima Zaidouni of Kavli Institute at the Massachusetts Institute of Technology for their mentoring, guidance and suggestions in the writing of this research paper. In addition, I would like to recognize my former teacher, Mr. Christopher Kingcombe of Westminster School, London, UK for all the personal help and guidance he has given me which has shaped my journey. Above all, I would like to acknowledge my father Sunil Dixit for everything he has done for me throughout my life, as no words can explain how much it has made me the person I am. I am dedicating this paper, my first published research paper, to my former teacher, Mr. Chris Kingcombe and my father, Sunil Dixit.

References

[1] C. Dwork, A. D. Smith, T. Steinke, and J. Ullman, "Exposed! A Survey of Attacks on Private Data," Annual Review of Statistics and Its Application 4 (2017), vol. 4, pp. 61-84, March 2017.

[2] C. Dwork, A. D. Smith, T. Steinke, J. Ullman, and S. Vadhan, "Robust Traceability from Trace Amounts," 2015 IEEE Annual Symposium on Foundations in Computer Science, vol. 56, pp. 650-669, October 2015.

[3] H. Hu et al., "Membership Inference Attacks on Machine Learning: A Survey," ACM Computing Surveys (CSUR), vol. 54, pp. 1-37, September 2022.

[4] L. Song, R. Shokri, and P. Mittal, "Membership Inference Attacks against Adversiarially Robust Deep Learning Models," 2019 IEEE Security and Privacy Workshops (SPW), pp. 50-56, May 2019.

[5] J. Ye, A. Maddi, S. K. Murakonda, and R. Shokri, “Enhanced Membership Inference Attacks against Machine Learning Models,” 2022 ACM SIGSAC Conference on Computer and Communications Security, pp. 3093-3106, November 2022.

[6] A. Thudi, I. Shumailov, F. Boenisch, and N. Papernot, “Bounding Membership Inference,” unpublished.

[7] L. Watson, C. Guo, G. Cormode, and A. Sablayrolles, “On the Importance of Difficulty Calibration in Membership Inference Attacks,” unpublished

[8] N. Carlini et al., “Membership Inference Attacks from First Principles,” 2022 IEEE Symposium on Security and Privacy (SP), vol. 43, pp. 1897-1914, May 2022.

[9] T. Humphries et al., “Investigating Membership Inference Attacks under Data Dependencies,” 2023 IEEE Computer Security Foundations Symposium (CSF), vol. 36, pp. 473-488, July 2023.

[10] S. Mahloujifar, A. Sablayrolles, G. Cormode, and S. Jha, “Optimal Membership Inference Bounds for Adaptive Composition of Sampled Gaussian Mechanisms,” unpublished.