Towards Understanding Data Vulnerability to Membership Inference Attacks

Zhiyi Wang

4th Year Project Report
Computer Science and Management Science
School of Informatics
University of Edinburgh

2023

Abstract

Membership inference attacks (MIA) enable attackers to determine the presence of specific data in a machine learning model's training set, which exploits the confidentiality and privacy of the data contributor. Our study examined factors contributing to data vulnerability in MIA for logistic regression models. The data vulnerability in our context refers to the difficulty of MIA in violating data privacy. We emphasized the importance of visualizing privacy violations to identify data vulnerability, which the outlying degree can not always capture. Our study discovered that removing some vulnerable data points from the training dataset may increase data vulnerability in MIA, while others can decrease it. But, explaining this phenomenon proved challenging, highlighting the complexity of explaining data vulnerability. We also examined the effectiveness of differential privacy (DP) in protecting data privacy, revealing that different data points possess varying levels of data vulnerability under different DP strengths. We explained this variation through the influence of DP on the model's decision boundary. Identifying the hard-to-protect vulnerable data points is essential since we showed removing a specific proportion of them can improve data privacy and protection efficiency if DP is applied to protect the target model. These insights help understand data vulnerability to MIA, providing valuable research for future machine learning privacy protection measures.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Zhiyi Wang*)

Acknowledgements

I'm grateful to everyone who has contributed to my project's success and personal development, especially my supervisor, Rik Sarkar. His guidance and mentorship have helped me develop a deep interest in research, particularly in differential privacy, and his support in my personal growth has been invaluable. Thanks to him, I've navigated significant turning points in my life, and I'm forever grateful.

I would like to express my heartfelt gratitude to Lauren Watson for her invaluable support and expertise in differential privacy and membership inference attacks. It was always a pleasure discussing my findings with her, and her insights played a significant role in shaping my research.

I am deeply grateful to my dear friends MengJie Li, JingYuan Wang, and YiNing Hou for their emotional support throughout this journey. Without their presence, I could not have accomplished this achievement.

Finally, I extend my heartfelt gratitude to my family, especially my parents, for their unwavering support and belief in my abilities to achieve my goals. I also thank my beloved pet dog, HaLei, for bringing joy and comfort to me throughout this project.

Table of Contents

Chapter 1

Introduction

An increasing number of machine learning models have been created, deployed and commercialised in various sectors that process sensitive personal information. Natural language processing [39, 13, 65, 28], biomedics [55, 61, 40, 8] and image classifications in disease diagnostics [53, 36, 18, 60] are common examples. The security of the models and the privacy of the sensitive data learned by the models has gained immense attention because of the potential of privacy violations on this sensitive information suggested in [15].

Among the range of privacy attacks, membership inference attack (MIA) was the focus of this study since it targets user-specific privacy [62]. In MIA, the adversary aims to investigate whether a given data point is inside the training set for the target machine learning model. Knowing whether training a model involves a specific instance potentially introduces privacy risks to the data contributor. For example, suppose a hospital trained and published a machine-learning model to diagnose patients based on their medical records and symptoms. Suppose an adversary is interested in the sensitive information this model has learned. This adversary could fake a patient data set to build his attack model for MIA. If the attack model suggests a high probability on a given data instance, then this instance is highly likely inside the confidential training set kept by the hospital. As a result, MIA violates the privacy and confidentiality of the patients whose sensitive information gets leaked to the attacker.

Recent studies on privacy issues in machine learning models realised that data points in the training set are not equally hard to attack with MIA [9, 58]. This problem is easier to see when outliers exist in the training set of a complex neural network. For instance, Carlini et al.[9] found that MIA exposed the outlying training data points at a low false-positive rate more frequently than non-outliers to the attacker when training a deep neural network on an image dataset. Similar results are also found in the study of Watson et al.[58] on a model of the same type, trained on the same dataset as Carlini et al. These findings led to a phenomenon that different subsets of the training set bear various difficulties to being attacked by MIA, a.k.a. they are unevenly vulnerable to MIA.

Knowing which part of the training set is more vulnerable to MIA is extremely helpful in

designing a protection mechanism to be applied to the model using protection techniques like differential privacy (DP) [17]. When differential privacy is applied to protect a machine learning model, it perturbs noises generated from a well-tuned distribution to the target of protection. This way, the presence of an instance inside the training set for this model will not significantly increase its risks of being exposed by MIA. But when protecting the model from privacy violations, it trades the utility of the model at the same time [23]. For instance, when the protection given to the model is set too strong, the model is hardly likely to function as it was supposedly because the model has become too noisy. On the other hand, if the noise is not big enough, the model might fail to be privacy-preserving while still useful for its tasks.

Thus, as a new direction in protecting data points' privacy in machine learning models, if we can identify which subset of training points are more vulnerable to MIA, we can know the proportion of training points that require more protection. This knowledge will allow practitioners to better decide on the scale of privacy protection based on their tolerance for the loss of privacy and the utility of their product.

1.1 Challenges and Related Work

Identifying the more vulnerable subset of training data to membership inference attack (MIA) is challenging, especially in complex models trained on high-dimensional datasets. Although the out-of-distribution data in the training set is often vulnerable, this does not imply that the frequently-identifiable data are always an outlier. Moreover, previous studies held opposite opinions regarding the correlation between data point vulnerability and outlier status, adding to the complexity of the problem.

One approach to detecting data points' vulnerability to MIA in recent scholars is to check if they are outliers in the target model's training set. Carlini et al.[9] found that outliers in the training dataset significantly impact data point privacy. They removed the easiest-to-attack outliers from the training dataset to train a convolutional neural network (CNN) trained on CIFAR10. They then observed the resulting privacy implications on the remaining training set. They found that removing these data points leads to a *Privacy Onion Effect* where the new model trained on the remaining set exposed a previously hidden layer of vulnerable data points. This finding suggested a close relationship between a data point's vulnerability to MIA and its outlying degree. On the other hand, Watson et al. [58] showed that being an outlier is not crucial in identifying data points' vulnerability to membership inference attacks. They examined the CIFAR10 dataset on a CNN model and observed that the data points frequently exposed to MIA were not necessarily extreme-outliers in the training set. Furthermore, they noted that a weak differential privacy (DP) could adequately protect these data points, which contradicts Carlini's perspective on data point vulnerability to membership inference attacks.

Previous works have revealed the challenges in arriving at a concrete answer to the problem of our interest. These challenges might attribute to the lack of explainability in the complex model architecture for neural networks [52, 4] and the difficulty in defining outliers in a high-dimensional training set [41, 2, 56, 59]. Since these research areas still lack promising results, readers with relevant interests can refer to the citations for

more information. Furthermore, optimizing neural networks' non-convex objective functions poses additional technical challenges. These optimization techniques introduce randomness into the training process, leading to different performance and attack results from multiple models trained under the same initial settings [22, 63, 32]. This uncertainty in inference outcomes and lack of explainability further complicates the question of why some specific data points are more vulnerable to MIA than others.

Previous works have shown that using neural networks trained on high-dimensional data is inadequate for our specific problem in this study. Because our research will need to face additional challenges, including the non-deterministic optimization results of the target models, the difficulty of explaining the behaviour of complex models, and the challenge of defining outliers in high dimensions. However, despite these challenges, it is still worthwhile to investigate the relationship between a data point's vulnerability to MIA and its degree of being an outlier. Since understanding the factors that might contribute to data vulnerability, we can take steps to enhance privacy protection in machine learning models.

Therefore, given the challenges identified, this project aims to tackle the problem in a less complex setting than the previous works: logistic regression trained on low-dimensional datasets that are small to medium size for binary classification tasks. In the next section, we will provide a detailed description of the problem we will address.

1.2 Research Focus and Questions

Based on the motivation and the challenges identified, our focus of the project will be centring around the following two questions:

- *Can we explain a data point's vulnerability more convincingly than how much the point is out-of-distributed in the training sets?*
- *Which subsets of data points in a training set are more vulnerable to membership inference attacks than others?*

We have discussed that answering these two questions is difficult in a complicated problem setting. Therefore, to answer them more comprehensively, we choose to lower the difficulty of the context to a convex learning problem – logistic regression, for binary classification tasks on small to middle-sized low-dimensional datasets. The challenges to solve in this project are:

1. *When the target model is logistic regression, is the data easier to attack still the out-of-distribution ones in the training set of the target model?*
2. *Can we still observe the privacy onion effect due to outlying data removal in the context of logistic regression?*
3. *Can we explain data points' vulnerability to be attacked by membership inference attack using its ease to be protected by differential privacy?*

With the focus of the project stated, experiments are planned and conducted. Following are the main contributions we made in attempting to answer them.

1.3 Main Contribution

Figure 1.1 summarize the most important part of our contributions using a synthetic data set generated from a Gaussian distribution. Detailed explanations of these contributions can be found listed below these illustrations.

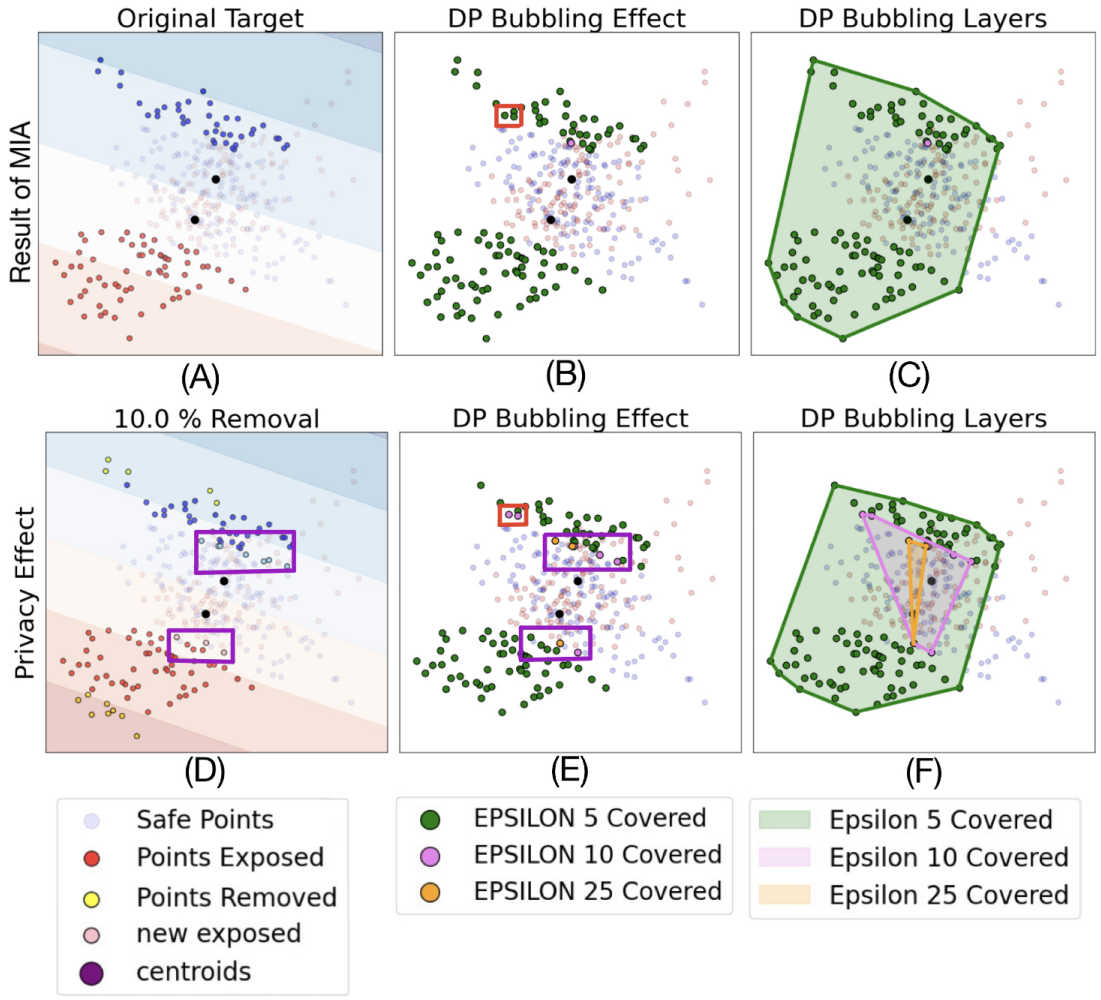

Figure 1.1: **(A)**: MIA on the target model exposed the privacy of the denser-coloured data to the attacker, while lighter-coloured data remained hidden. This pattern of privacy violation and the model's decision boundary in the background indicates that data vulnerability is related to the nature of the model (Contribution 1); **(D)**: Removing the yellow dots from the training set violated the privacy of the data inside the purple frames. This illustrates a part of our second contribution regarding the privacy impact of data removal; **(B)**: By demonstrating how the exposed data points in (A) are obscured when one level of DP is applied to the model at a time, we provide evidence for our fourth contribution that not all data points in the training set are equally difficult to protect; **(E)**: Our attempt to conceal the exposure in (D) leads to our fifth contribution: removing the yellow dots in (D) protected the data in the red box with weaker DP measures and the new exposure framed with purple boxes can be covered with a weak DP as well; **(C)&(F)**: We draw convex hulls on data covered at various DP levels in (B) and (E) respectively, to demonstrate that the difficulty of protecting a training set is layered, with easier layers nested inside harder ones.

1. Observing the pattern of the privacy violation in membership inference attacks (MIA) on logistic regression training sets gave us insights that outliers are not

always vulnerable (Plot (A)), and data vulnerability depends on the characteristics of the model. Knowledge of these properties fosters our understanding of data vulnerability in the current context. Thus, direct observation of the results of MIA helps us understand data vulnerability.

2. We discovered two additional privacy effects on the training set resulting from data removal in logistic regression, other than the *Privacy Onion Effect* shown in Plot (D). Data removal may either *expose previously safe data (Privacy Onion Effect)*, *hide previously unsafe data*, or *have no impact* on the remaining training data. However, we found no clear guidance on where to remove data for privacy effects on the remaining data, but there is guidance on the number of vulnerable data points to remove for privacy benefits. These findings emphasize the potential risks and benefits associated with data removal from the training set.

3. We also identified a potential contributor to the privacy effects of data removal: the scale of the shift in the decision boundary caused by the modified training set. While our experiments only provided empirical evidence of a relationship between these two factors, we do not yet know in what way the shifts in the decision boundary might lead to each kind of privacy effect. Our results suggest understanding this factor as a potential future direction for research in understanding data vulnerability.

4. The difficulty of protecting data points in a training set can vary, as demonstrated by their ability to be covered by different scales of differential privacy (DP), as shown in Plots (C) and (F). This bubbling cover effect suggests that we can use the difficulty of protecting data points in defining their vulnerability to MIA if we consider the easier-to-protect data as the less vulnerable one to attack. Additionally, we explained the variation in difficulty to protect with the scale of the shift in the decision boundary brought by DP at various levels.

5. Further, when combining data point removal with DP, we found that DP can effectively cover the results of the *Privacy Onion Effect*, as shown in the purple boxes in plots (D) and (E). Additionally, we found that data removal helped lower the level of DP needed for some instances to be protected from MIA, as shown in the red boxes in plots (B) and (E). This finding emphasises the importance to identify which subset of data points should be removed from the training set to benefit the privacy of the remaining data, which may include the hard-to-protect data points described earlier.

Our contributions answer the three research questions proposed earlier in a more focused and deterministic problem setting than previous works, with logistic regression and low-dimensional datasets. We learned from previous work to avoid overcomplicating the problem for clear insights into the reasons behind data vulnerability. As a result, we gained more comprehensive knowledge about data vulnerability and how it can lead to a more efficient privacy protection mechanism. Our results are illustrative and are rare in academia, providing clarity to the research community. We recommend that future researchers explore the potential benefits of combining the removal of hard-to-protect data points and a weak differential privacy perturbation to protect the remaining easy-to-protect data points in the training set from MIA attacks.

1.4 Report outline

- **Chapter 2** provides the theoretical background necessary to understand the rest of the project and defines the concepts and terminologies for the experiments.
- **Chapter 3** outlines the hypotheses, experimental pipelines, and technical details, followed by the performance of the crucial steps in our pipeline.
- **Chapter 4** presents our findings from the experiments and provides explanations for some of them.
- **Chapter 5** concludes our findings with a critical review for potential improvement on our work and ends with future directions for further explorations.

Chapter 2

Background and Definitions

In this chapter, we will provide an overview of the key preliminary concepts needed to comprehend the challenges addressed in this project. We will also introduce the terminologies defined for use in the experiments.

2.1 Machine Learning Preliminaries

This section will briefly explain the relevant machine learning concepts and the rationale for choosing logistic regression as the focus of our study. For a more comprehensive theoretical understanding of machine learning, readers may refer to [50].

2.1.1 Machine Learning in General

Machine learning can be conceptualized as a human learning system, consisting of two crucial components: **memorization**, which involves the ability to recall information learned, and **generalization**, which involves the ability to apply learned knowledge to new situations.
Suppose we refer the model as a learner, the fundamental aspects to teach this learner is to formulate the learning system mathematically, which can be defined with the following elements:

- A domain set X includes the instances we want the learner to learn and predict.
- A label set Y to categorise the instances in the domain set.
- A training set $S = (x_1, y_1), ..., (x_m, y_m)$ as a sequence of labelled instances that is accessable to the learner.
- A hypothesis $h : X \to Y$ as a prediction rule to infer the label of an input domain, and a correspondence hypothesis class H where $h \in H$.
- A learning algorithm A to select the hypothesis h from the hypothesis class H based on the training set S s.t. $h = A(S)$.

- A probability distribution D over X where the domains in the training set is generated from, and a perfectly correct labelling function $f : X \to Y$, s.t. $y_i = f(x_i)$ for all i, both D and f are unknown to the learner.
- A loss function as a measure of the learner's success in its classification $L_{D,f}(h) := \underset{x\sim D}{P}(h(x) \neq f(x))$.

With these terms, given the learner access to S with each instance inside has been labelled correctly by f. The learning process for the learner is to use A to find a h that minimise the loss L w.r.t. D and f.

Problem of overfitting: Overfitting occurs when a model fits the training set S too well, resulting in an inability to react properly to instances outside the training set s.t. $h_S(x) = y_i$ iff $x = x_i$ for $i \in |S|$. This indicates that the model has memorised the training set instead of generalising on it. [42] states that an increase in the gap between a model's empirical loss $L_S(h)$ and its true loss (generalisation error/prediction loss) $L_{D,f}(h)$ after a continual decrement can indicate overfitting. This gap, referred to as the model's **generalisation gap**, is commonly used to measure the model's performance in generalising.

$$\textit{Generalization gap} = |\textit{True loss} - \textit{Empirical loss}| = |L_{D,f}(h) - L_S(h)| \tag{2.1}$$

How Overfit leaks Privacy: MIA is easier when the model is overfitting, as Yeom et al.[62] found a strong correlation between the model's ability to generalize and the adversary's advantage in MIA. Specifically, the loss on instances inside the training set is lower than the ones outside when the model is overfitting. Yeom et al. formalized this property as loss-based MIA and used it to distinguish members of the training set which we will cover more in Sections 2.3.2.

2.1.2 Why Logistic Regression

A logistic regression model is a sigmoid function mapping real value instances to a probability between 0 and 1. Given an input $s = \{x, y\}$, the model f output the probability $prob(x) = \frac{1}{1+e^{-x}}$ of the instance x belongs to class y.

The choice of the logistic regression model as the target of studying is based on its desirable properties. First, logistic regression is a supervised learning problem, which makes it a suitable choice for studying MIA on machine learning models that are typically trained using supervised learning algorithms [14]. Second, logistic regression is a convex learning problem, this means the lost function has a global minimum, making it easier to train [7]. Finally, logistic regression can be trained using gradient-based optimisers, which are computationally efficient and widely used in machine learning [46].

2.1.2.1 Supervised Learning

Supervised learning involves a model learning to predict output labels based on input data [14]. The model is trained on a labelled training set S, and a testing set is kept

unknown to the model. There are two types of supervised learning problems: classification and regression. Classification tasks have discrete and categorical output labels y_i, while regression tasks have continuous and numerical output labels. Although logistic regression is named after "regression", it is often used for classification tasks as its outputs represent the probabilities of an input being classified as each possible output label.

2.1.2.2 Convexity

A learning problem is said to be convex when its hypothesis class is a convex set, and its loss function is a convex function [7]. When the hypothesis class H is a convex set, for any two functions h_1,h_2 in H, all other functions in H can be written as $\alpha h_1 + (1-\alpha)h_2$ where $\alpha \in [0,1]$. When the loss function L is convex it satisfy the following property for any two data points u, v and $\alpha \in [0,1]$:

$$L(\alpha u + (1-\alpha)u) \leq \alpha L(u) + (1-\alpha)L(v) \tag{2.2}$$

A convex function has only one optimal value, Figure 2.1 illustrate this property.

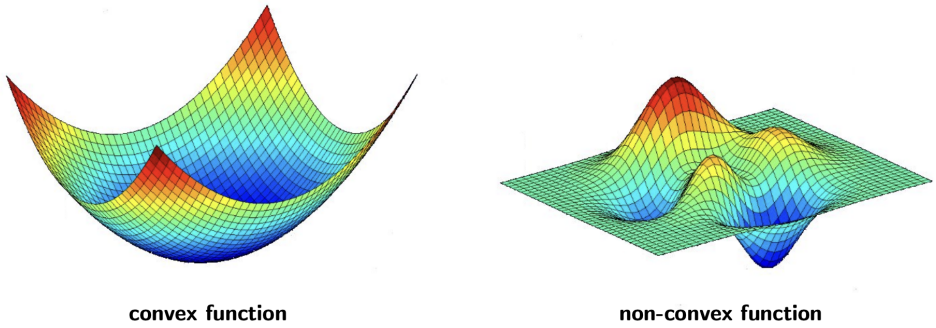

Figure 2.1: 3D representations of the Loss landscape for **LHS**: convex function & **RHS**: non-convex function; When the learning is convex, the loss would be gradually optimised to the bottom of the bowl and that will be the global minimum. On the other hand, non-convex learning has the problem of being stuck in the local minimum because of the jerky landscape.[25]

Combining the property of the hypothesis class being a convex class and the loss function being a convex function allows a fast convergence for the loss to the global minimum.

Logistic Regression is Convex: Logistic regression is proved by [45] to be a convex model because of the logistic regression function, the sigmoid function, is convex. Furthermore, if cross-entropy loss (see Definition 2.6) is chosen to be the loss function for logistic regression. Then it guarantees the training process is convex because this loss function is also convex.

2.1.2.3 Gradient Based Optimizer

Gradient-based optimizers gradually minimize the loss function by calculating and updating the gradients of the loss function on the samples at each step of the optimization process. There are three categories of gradient-based optimizers that differ in the method

of sampling from the training set during the gradient update process, which can be seen in Figure 2.2.
Among the three options, stochastic gradient descent (SGD) is popular for complex models with large datasets [6], while batch-gradient descent is still able to efficiently converge on small-to-medium scale problems. Thus, the decision of which kind to use in the training process depends on problem scenario.

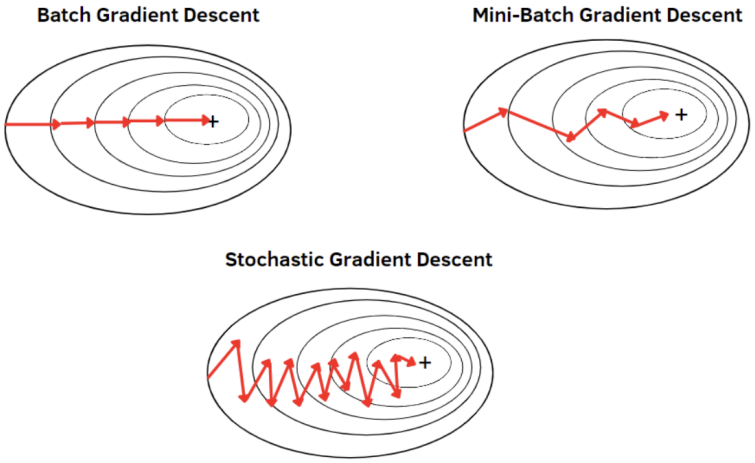

Figure 2.2: Traces of the process in minimising the loss function with batch gradient descent (BGD), stochastic gradient descent(SGD) and Mini-batch; Mini-batch and SGD are faster than BGD but may not guarantee convergence. BGD takes the entire training set in one step, ensuring convergence. SGD samples randomly, resulting in a jerky path to the optimal.[49]

After explaining the basics of machine learning and the reasons for selecting logistic regression as the target of study, the next section will delve into the details of the privacy attacks employed in this study. This will begin with an overview of privacy attacks that target a model's training set.

2.2 Privacy Attacks on Training Set

Known privacy attacks on the training set of a model based on the setting of a black-boxed target model and assuming the attacker knows the data distribution D of the private training set S, includes: attribute inference attacks (a.k.a. model inversion attacks) [19, 64, 19], property inference attack[20, 33], and membership inference attack[62, 51].

In an attribute inference attack proposed by Fredrikson et al. [19] , the attacker uses available information about users to infer their disclosed sensitive attributes. An example of such an attack is when an attacker uses a person's public social media activity to dig the hidden information about their sexual orientation. Similarly, property inference attacks infer hidden characteristics about targets. For instance, an attacker could get to know targets' financial status based on their browsing history in the web browser.

Although both attribute and property inference pose great threats on users' privacy since they are able to extract users' information without consent. None of them infer directly

the existence of an instance in the training set, which is achievable via the membership inference attack explained in the following.

2.3 Membership Inference Attack

As briefly introduced in Chapter 1, Membership Inference Attack (MIA) reveals the presence of a user in the model's training set, which poses a significant threat to data contributors' privacy. It can lead to serious privacy breaches includes:

1. Knowledge of sensitive data existence, such as medical records, violates confidentiality [54], which is often regulated by data protection laws like GDPR [57].
2. The attacker can rebuild a training set for an easier attribute or property inference attack using the results from MIA[62].
3. The attacker can establish correlations between known instances in the training set and other public databases for more information.

Shokri et al. [51] proposed the first application of MIA on machine learning models through shadow model training techniques. This method requires the attacker to train and use multiple shadow models to mimic the behaviour of querying the target model with instances carrying different labels. Then, the attacker use the output of these shadow models and the attacker's public dataset to train an attack model to infer members of the target training set. However, Yeom et al. [62] showed that utilising the overfitting property to attack the model can perform similarly to the shadow training approach for the reason we explained at the end of Section 2.1.1. For efficiency and explainability, this project will use the score-based MIA method.

Following is a general definition for membership inference attack by [62].

2.3.1 Definition

Given adversary the public API access (a.k.a. a black box) to the target model M; a target data entry s=*(x,y)* where x is the input domains and y is the label. s would be sampled from the target model's training set based on a coin flip decision; the size of the training set $n = |S|$ and the distribution D where the target model's training set is drawn from. Let A denote the attack model in MIA, A can be defined as:

$$A(M,s,D,n) = \begin{cases} 1, & \text{if } s \in S. \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

This means the model's output being 1 if the entry s is indeed inside the training set and 0 otherwise. The technical definition for each type of MIA differs by category, like the shadow model training and the score-based MIA showing distinctive approaches. The following section will cover more details about the score-based MIA due to its usage in the experiments.

2.3.2 Scored Based MIA

Because of target models' overfitting property, score-based attacks react distinctively on data points in the training set to the ones outside the training set (see Section 2.1.1). Relying on this property means the score-based attack is more efficient than the Shadow-training approach because no additional model is needed. Score-based MIA assigns a membership score $score(s_i, M)$to each instance s_i in the dataset as to how likely each is inside the training set. Based on a preset threshold value τ obtained from the process of threshold selection. The attacker can classify data points as members by comparing their membership scores to the threshold.

Threshold Selection is a crucial setup in loss-based MIA. Given the attacker the access to the distribution D,M, and the size of the target's training set $n = |S|$. The attacker randomly sampled a public dataset S_{pub} of size n from D. Then, the attacker obtains a collection of threshold values using S_{pub} and M, as a sequence of 1000 or more values within the range of the membership scores for the S_{pub}. Then, the attacker tries each of the threshold values τ in this collection in the loss-based attack by applying the following threshold function on each target data s_i:

$$score(s_i, M) > \tau \tag{2.4}$$

Data points having a membership score higher than this threshold would be classified as members and vice versa. The loss based MIA A_{loss} can be formulated as:

$$A_{loss}(M, s, D, n, \tau) = \begin{cases} 1, & \text{if score}(s_i, M) > \tau \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

High Precision Score-based MIA [58] require a threshold value that would lead to a high-performing inferencing result. Ideally, in a membership inference attack, an attacker would aim to infer a significant number of members inside the training set, achieving high precision and recall. However, this ideal scenario is rarely achievable, and it is more common to accurately infer only a few members, resulting in high precision and low recall. Privacy attack with high precision poses a significant threat to users' privacy since privacy attack also obey the rubrics in many areas of computer security[26, 27, 34, 24]. Thus, this project focuses only on high-precision score-based attacks.

Selection of Membership Score determines the category of a score-based attack. Currently, the common uses of membership scores are: prediction loss [62], prediction confidence [47], prediction entropy [47], and gradient norm [37]. Among them, we chose **loss-based MIA** for this study because the logistic regression model's output on a given instance is its cross-entropy loss, which can be directly used as input for the loss-based MIA. This allows us to leverage the strengths of both the logistic regression model and the loss-based MIA approach for effective membership inference attacks.

The following sections outline our selection of membership scores used in the loss-based attacks and provide theoretical frameworks for implementing high-precision loss-based attacks.

2.3.2.1 Prediction Loss as Membership Score

Yeom et al. [62] proved the correlation between an instance's prediction loss (generalisation error) on a target model and its privacy risk in an MIA in Section 3.2 of his paper. This correlation suggested that the higher the model's generalisation gap, the higher its privacy risk. One contributor to the large gap is having high prediction loss and low empirical loss (Overfitting). Hence, an instance's prediction loss from the target model is used as part of its membership score in a loss-based attack. With the selection of cross entropy loss L_{CE} as the loss function for our target model, L_{CE} can be defined as follows:

Assume the model is classifying data into n classes. Let $s_i = (x_i, y_i)$ denote a data entry; M be the model s.t. given x_i, it outputs the probabilities p_i that x_i been in each class i $\in$ n; Let t_i denote the one-hot encoded representation of s's true label – t_i is a vector of length n with the y_{th} entry been 1 and else been 0. Then the cross entropy loss $L_{CE}(s_i, M)$ of s_i from the model M is defined as:

$$L_{CE}(s_i, M) = -\sum_{i}^{n} t_i \log p_i \tag{2.6}$$

And the membership score can be defined as:

$$score(s_i, M) = -L_{CE}(s_i, M) = \sum_{i}^{n} t_i \log p_i \tag{2.7}$$

Notice that the membership score holds the opposite sign to the cross-entropy loss, this setting allows MIA to distinguish the members of the training set since they usually have a low empirical loss.

However, loss-based attacks might be poor performing in the case of distinguishing data of high dimensionality. This is because both easy-to-predict non-members and hard-to-predict members can achieve high scores using plain loss as the membership score[58]. Taking into consideration this potential drawback, we implement another version of loss-based MIA as a resolution.

2.3.2.2 Calibrated Loss as Membership Score

Watson et al.[58] proposed calibrating each instance's difficulty to attack to their membership score to improve the drawback of loss-based MIA. As a result, the easy-to-predict non-members became distinguished from the hard-to-predict members, which led to better attack performance. The calibrated membership score can be defined based on the empirical loss as follows:

Following the same assumptions and notations used to define the loss-based MIA. Additionally, let D_{shadow} denote a shadow dataset drawn from the same distribution of where S is drawn from, and let R denote a randomised algorithm that samples over a collection of models trained on D_{shadow}, the calibrated membership score is defined as:

$$score_{cal}(s_i, M) = score(s_i, M) - E_{M' \leftarrow R(D_{shadow})}[score(s_i, M')] \tag{2.8}$$

The framework of score-based MIA and the two versions of membership scores to use in our study have been introduced. We have covered all the relevant knowledge about privacy attacks in this study. Thus, the following is going to explain the protection mechanism used in the study.

2.4 Differential Privacy

What is Differential Privacy: Differential privacy (DP) was formally introduced by Dwork et al. in *The Algorithm Foundation of Differential Privacy* [17]. DP is scalable privacy protection by randomly introducing noises on the outputs from a machine learning model, ensuring that the publication of those outputs will not suffer the privacy of the original inputs. Since with DP, models trained on datasets differ in one instance would behave similarly. This outcome implies that the presence of the instance which distinguishes the two datasets becomes less obvious in the training set, and hence its privacy is protected. In a word, DP allows models to learn an instance without its presence in the training set.

What Differential Privacy Promises and Not: Differential privacy allows the learner to learn about the target of privacy protection, which seems to be absent from the content due to the protection mechanism. For example, suppose a company wishes to publish the mean salary statistics each month while keeping individual salaries private. DP allows the disclosure of statistics as if some employees are not inside the company, but they are. Hence, their personal information is kept safe from potential privacy violation in the publication of the statistics.

However, differential privacy does not promise the conclusion reached by the learner would not disclose facts or related information about the individual irrelevant to their presence. For instance, suppose the company has a history of giving bonuses to its employees at the end of the year. Thus, the average monthly salary at the end of each December would be higher than in other months. Then it's enough for an outsider to learn that this company might have such a bonus mechanism even if the average salary was protected by DP.

Randomness in Differential Privacy: Being inspired by the requirement for randomness in semantic secure cryptosystem [38, 35] and encryption scheme [5, 29, 3] in computer security. Dwork et al. also redeemed that randomness is essential to define differential privacy as 'privacy preserving'. Consider the following randomised algorithm $M : X \to \Delta(Y)$ which outputs the probability space $\Delta(B) = \{x \in R^{|Y|} | x_i \geq 0 \text{ for } \forall \text{ i and } \sum_{i=1}^{|Y|} x_i = 1\}$ on the input domain $x_i \in X$. Then, differential privacy can be defined as stated below.

Differential Privacy Definition: Using L_1 norm of a dataset as a measure of its size $||s||_1 = \sum_{i=1}^{|S|} |s_i|$, the difference in the number of instances between two datasets (S, S') can be noted as $||S - S'||_1$.

Suppose there's a set of neighbouring datasets that differs in only one entry being denoted as (S, S'), such that the L_1 distance $||S - S'||_1 \leq 1$. A randomised learning

model M is said to be (ε, δ)-differential private if for $\forall Y \subseteq Range(M)$:

$$Pr(M(S) \in Y) \leq exp(\varepsilon) * Pr(M(S') \in Y) + \delta \tag{2.9}$$

Similarly, M is ε-differential private, if $\forall Y \subseteq Range(M)$:

$$Pr(M(S) \in Y) \leq exp(\varepsilon) * Pr(M(S') \in Y) \tag{2.10}$$

While this defines differential privacy in general, it may be too abstract to think of its applications in machine learning tasks. Thus, we will explain how DP is made applicable to machine learning models in the coming section.

2.4.1 Application of DP on Machine Learning Models

Dwork et al. also proved that noises randomly drawn from certain probability distributions also apply the definition of differential privacy. Perturbing such noises to the target function would also make the function privacy-preserving. These certain distributions could be: *Laplace*, *Exponential* or *Gaussian* depending on the actual use-case. This process of perturbing noises to the target model is called **Noise Perturbation**.
For instance, a ε-DP Laplace mechanism can be defined with a laplace distribution $Lap(*)$ as follows:
Given a hypothesis $h : X \to Y$ with the sensitivity of h being $\Delta(h)$ and the privacy parameter ε, the Laplace mechanism $M_{Lap}(x, h, \varepsilon)$:

$$M_{Lap}(x, h, \varepsilon) = h(x) + (R_1, ..., R_{|Y|}) \tag{2.11}$$

where R_i are drawn i.i.d. from $Lap(\frac{\Delta h}{\varepsilon})$.

Because the Noise Perturbation method described above matches the mathematical formulation of a machine learning problem, differential privacy became applicable to machine learning tasks. Common ways of realising such statistical mechanism include noise perturbations on the model's parameters during the training process [1]; noise perturbation on the model's outputs [44], and on the model's objective function [10]. It is hard to be clear about which method is prior to the other since each of them has its own pros and cons to consider under specific contexts. For this project, object perturbation is considered since it can be efficiently applied to logistic regression models.
DP through Objective Perturbation [11] is an efficient algorithm to introduce differential privacy on logistic regression models. This method perturbs well-tuned noises on the model's objective function before the start of the optimisation process. It was also proved to satisfy ε-differential privacy (Definition 2.10). Readers with interest in this algorithm can refer to Section 3.2 and Section 3.3.2 in [11].

2.5 Resolution to the Randomness

Randomness in DP set a stone on our way to draw concrete conclusions in our experiment since it leads to uncertainty in the observations. As a resolution, we proposed the following terminologies.

2.5.1 Average Protection Success Rate

Randomness in DP means there's the likelihood that two models trained with the same privacy settings on the same training set would perform differently in their protection. Thus, we should observe DP's average performance to determine its efficiency. We defined the term *Average Protection Success Rate* to measure the average performance of DP in protecting the points exposed in the non-private model as follows:

APSR: Given an original non-private target model as M trained on the set S which consists of n elements sampled randomly from the entire distribution D. Formally,

$$S \subseteq D, \text{ where } |S| = n.$$

Let A denote the adversary as defined in Section 2.3.1, and let S_{vul} denote the set of points exposed in the non-private model M such that:

$$A(M, s_i, D, n) = 1 \quad \text{for all } s_i \in S_{vul} \text{ and } S_{vul} \subseteq S \tag{2.12}$$

let M_ε denote the model trained with ε-DP on the set S. Suppose we train N number of M_ε to see ε-DP's average performance in protecting the points in S from been discovered by A. Let M_ε^i with $i \in \{0, 1, ..., N\}$ denote the i_{th} one of the N models. Then, we can define the *Average Protection Success Rate* of ε-DP on each data instance s in the Set S_{vul} as following:

$$APSR(s, M_\varepsilon, A) = \text{Prob}[A(M_\varepsilon, s, D, n) = 0] = 1 - \frac{1}{N}\sum_{i=1}^{N} A(M_\varepsilon^i, s, D, n) \tag{2.13}$$

for $\forall s \in S_{vul}$.

With each instance in the S_{vul} has been assigned an APSR score through a series of attacks on the ε-DP models. We measure the performance of ε-DP by how many data points can be protected more than half the time. Thus, if a data point's APSR score exceeds 0.5, it is **protected** by DP at ε level. We chose this thresholding value because of APSR's similarity to the definition of an average attack success rate in [9].

2.6 Degree of Outlying for Data Points

Because outlier is a central part of our study, we should define them according to our context. We use the term 'degree of outlying' to describe how distant a data point is, relevant to most of its class. We use two distance metrics to define the data point's degree of outlying in this study for the purposes stated in the following sections.

Distance to Cluster's Centroid is applicable when the data is two-dimensional, which allows the use of the centroid of these points as a target to compare with other data points in the training set for measuring their degree of outlying in the set. The following is how to calculate the centre of each cluster of data points.

Given a collection of training points of $class_a$: $S_a = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ of length n. The centriod of these points: $(\widetilde{x_a}, \widetilde{y_a})$ is calculated as:

$$(\widetilde{x_a}, \widetilde{y_a}) = (\frac{1}{n}\sum_{1}^{n} x, \frac{1}{n}\sum_{1}^{n} y) \tag{2.14}$$

Then, for each point $s_i = (x_i, y_i)$ in $class_a$, the distance d_i for $\forall i \in [1, n]$ to their centroid, $(\widetilde{x_a}, \widetilde{y_a})$ can be defined as the following:

$$d_i = \sqrt{(x_i - \widetilde{x_a})^2 + (y_i - \widetilde{y_a})^2} \tag{2.15}$$

Distance to the Decision Boundary of the Logistic Regression Model is another metric to measure data points' degree-of-outlying because data vulnerability might be relevant to the nature of the model. This distance can be defined using the model's hyper-plane. For instance, given a 2d data point $s = (x, y)$, with a logistic regression model defined as:

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \beta_2 y)}} \tag{2.16}$$

The point's distance d to the decision boundary can be defined as:

$$d = \frac{\beta_1 x + \beta_2 y + \beta_0}{\sqrt{\beta_1^2 + \beta_2^2}} \tag{2.17}$$

Since the denominator is a constant for each model, we use only the numerator as an estimate of the points' distance to the decision boundary in the actual experiments for efficiency. Thus:

$$d \approx \beta_1 x + \beta_2 y + \beta_0 \tag{2.18}$$

Chapter 3

Experiments

Experiments are needed to answer the questions stated in Chapter 1. In this chapter, we first formalise these questions into hypotheses, then plan the experiments accordingly into pipelines. We also explain the technical details for implementing the crucial steps in our pipeline and their performance.

3.1 Hypothesis and Pipelines

Hypothesis 1 (H1): ***Data Points' Exposure to MIA is Relevant to their Distribution****: On the logistic regression model, the set of points exposed by loss-based MIA tends to be out of distribution in the training set.*
Hypothesis 2 (H2): ***Removing Sensitive Points trigger Privacy Onion Effect in logistic regression****: Suppose the sensitive points in S of the model M are the points exposed by MIA that are more out-of-distributed than others. Removing a fraction of these sensitive points from S will expose another set of more inlining points to the attacker in logistic regression models.*
Hypothesis 3 (H3): ***Data Points' Protection Difficulty is Distance-Related****: On the logistic regression model, the more outlying the points exposed by loss-based MIA are. The larger the noises needed from DP to protect its privacy and vice versa.*

With the hypotheses stated, a set of experiments are designed to test each of the hypothesis in the following section.

3.2 Pipelines

We use the following set of flowcharts to illustrate each step in the experiment planned for testing each of the hypotheses. The resulting output at the end of each pipeline will be used to examine and analyse the three hypotheses respectively.

We are then going to cover the general set-up that is common for these experiments in Section 3.3.

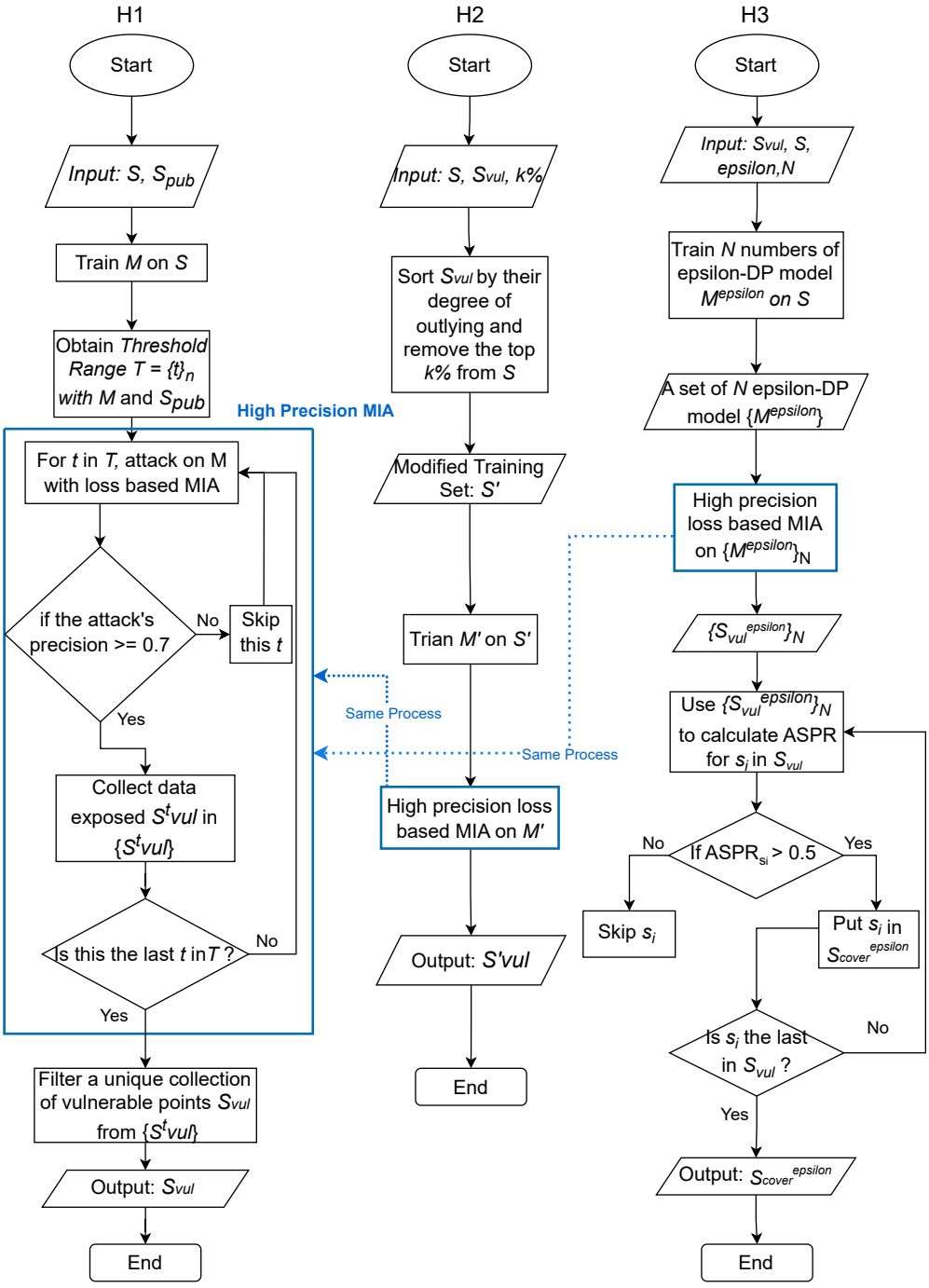

Figure 3.1: Flow charts of the steps for experimenting on the three hypotheses respectively; Each flow from the start to the end marks one run of the experiment. For instance, to do one experiment on testing H3 on one ε value, all the steps in the H3's flow need to be followed from the start to the end.

3.3 Experiment Setup

Following are the common setups in the target model architecture and the datasets for all three kinds of experiments.

3.3.1 Target Model Architecture

We mainly experiment with our hypotheses on a logistic regression model, with the reasons stated in Section 2.1.2. The learning algorithm used in the experiments is a batch-based gradient descent algorithm: *Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS)* algorithm [48], due to its efficiency in decreasing the cross-entropy losses to the optimal with time complexity of $O(n^2)$ where n is the size of the training set.

3.3.2 Datasets

Following are the datasets used in the experiments: Two synthetic datasets generated from different Gaussian distributions: *SynLR1* and *SynLR2*, each containing 400 entries of 2-dimensional data; Some benchmark datasets obtained from the UCI machine learning repository [16]: *Haberman's Survival* contains 306 entries of 4-dimensional data; *Cleveland Heart Disease* contains 303 entries of 13-dimensional instances, and *Adult* contains 48842 entries of 14-dimensional data. The small to middle sizes datasets allowed fast convergences of *LBFGS* optimiser on the objective function. The dimensionality of the data with more than 2 dimensions was reduced to 2 by principal component analysis (PCA) [31] to satisfy our problem setting on low dimensional data.

3.4 Target Model Training

Before the model is trained with the optimiser, the dataset is split equally into a public and private set based on the assumptions of the attack (in Section 3.5.1). Then, with the *LBFGS* optimiser, the model is trained to its convergence on the given private training set and tested with the private testing set.
At the start of experiments, we fix a public-private split for each distribution and train an original model. The following table records the corresponding statistics for these models respectively.

Original Target Models			
Model	DataSet	TrainAcc(%)	TestAcc(%)
LR	SynLR1	83.428	77.333
	SynLR2	66.857	61.333
	Adult	80.142	79.815
	Haberman	76.636	76.086
	Heart Disease	84.761	69.565

Table 3.1: The optimal performance of memorisation (Train Acc) and generalisation (Test Acc) of the target models trained on the default training set for each data distribution used in this study.

Then in the later experiments, we would randomly switch the public datasets accessible to the attacker and the private datasets accessible to the target model for a thorough inspection.

3.4.1 Modified Target Model Training

We need to modify the private training set for experiments on H2. This action can impact the model's empirical performance on the training set and overall performance on the testing set. Following the default settings of public-private data splits for each distribution, we progressively removed a bigger number of the most out-of-distribution points from each private training set, using both the definitions of data points' degree of outlying outlined in Section 2.6.
We trained new target models on the modified private training set and test them on the original private testing set, removing data points based on their distance from the decision boundary. Accuracy trends are depicted in Fig 3.2, with a similar trend for the other case of removing data points based on their distance from the clusters' centroids found in Appendix A, Fig A.1.

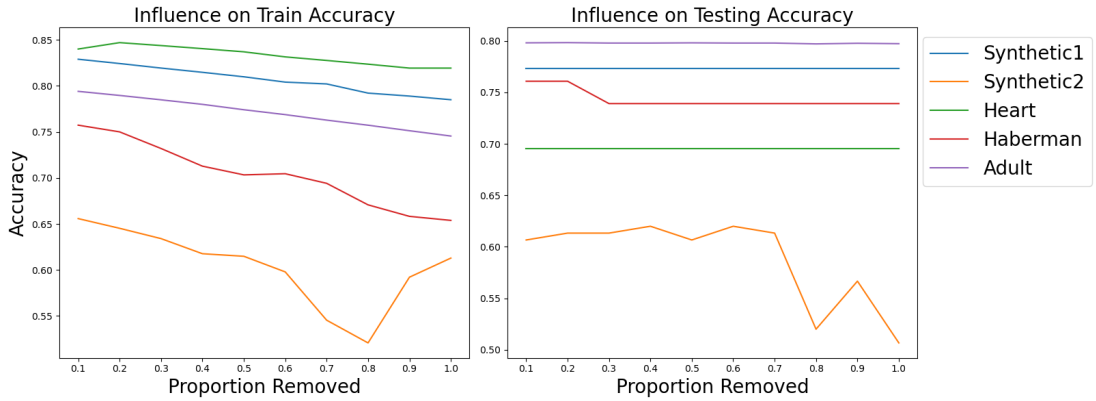

Figure 3.2: LHS: drop in training accuracy when an increasingly big proportion of data was removed; **RHS**: the drop in the testing accuracy according to the drops in training accuracy; Remove data points by their degree of outlying from the private training sets doesn't influence a lot on the model's training accuracy and testing accuracy on most of the data distribution, except the *SynLR2* until removed by 80%.

It's interesting to see a turning point at 0.8 in the continuous decrease of training accuracy on the model trained on the private training set sampled from *SynLR2* distribution. The most likely cause of the remaining training data points forms a new pattern that is linearly separable. Hence we scatter the modified training set and the corresponding classifier's decision boundary for inspection in Fig 3.3.

Thus, as the proportion of removal increases, the model might find new patterns in the training set and result in higher empirical accuracy, but this can't prevent the drop in testing accuracy. This drop in accuracy might give us unexpected results when we attack the privacy of the model training set with the methods described in the following sections.

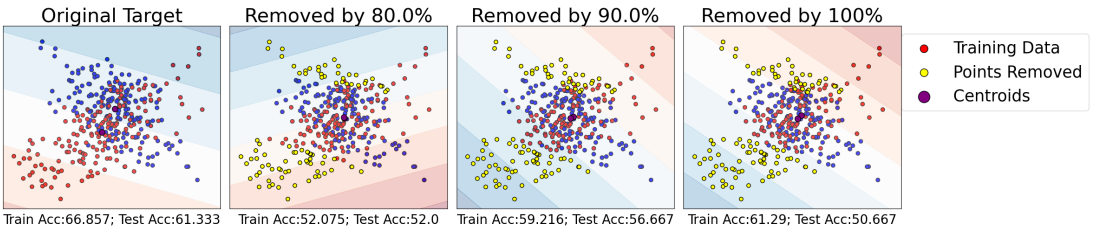

Figure 3.3: The data being removed (yellow dots) were scattered with the remaining training set (in red and blue) with the model's decision boundary in the background; Comparing the distribution of the modified training data by more than 80% to the decision boundaries indicate that the modified training sets are linearly separable.

3.5 Attack and Protect

This section detailed the procedures for implementing loss-based MIA and differential privacy protection on the target models trained. Starting with the attack, we will compare the performance of a loss-based attack to a calibrated loss-based attack to see which is more suited to our hypotheses in the sense of having a more balanced tradeoff between attack performance and efficiency. Efficiency is essential for getting our results due to the need for multiple attacks in some experiments.

3.5.1 General Assumptions on the Attack

We made the following assumptions about the attacker in the membership inference attacks:

- The attacker can only access the model as a black box.
- The attacker knows the data distribution of where the training set is drawn from.
- The attacker knows the type of the model.
- The size of the private training set.
- The attack will randomly sample a public training set of the same size as the private one for threshold selection.

3.5.2 Attack Performance

Loss-based attacks and calibrated-loss attacks are performed on the target model trained in Section 3.4.1. The following metrics are used to see if the attacks have the potential for high-precision attacks.

- **Highest Precision with its Recall**: High precision tells us the existence of threshold values for a high-precision attack, and recall value tells us the proportion of such high-precision thresholds among all the threshold values. $Precision = TP/(TP+FP)$ tells how much among all the 'true' responses from the attack are correct; $Recall = TP/(TP+FN)$ tells us how good the attack is in making correct inferences among all the 'true' instances.

- **Mean Accuracy**: The mean attack $accuracy = (TP+TN)/(T+N)$ averaging across all the thresholds, gives us a general performance of the attack.
- **Min Accuracy**: The poorest possible attack accuracy as a bottom line on the attack's performance.
- **AUC (Area Under the ROC Curve)**: the area under the **Receiver Operating Characteristics** curve tells us the overall performance of the attack in distinguishing the members of the training set from non-members; Unlike the measure of highest precision with recall, AUC tells the attack's performance across all the threshold values.

Following records the performance of each kind of attack using these metrics.

3.5.2.1 Uncalibrated Loss Attacks

The table below records the attack performance of the loss-based MIAs on the target models.

Loss Based Attack on Logistic Regression Models				
DataSet	Highest Precision w Recall	Mean Acc	Min Acc	AUC
SynLR1	0.735, 0.711	0.677	0.300	0.537
SynLR2	0.875, 0.02	0.611	0.302	0.533
Adult	0.710, 0.120	0.671	0.300	0.502
Haberman	1.0, 0.018	0.639	0.308	0.463
Heart Disease	0.746, 0.448	0.630	0.305	0.482

Table 3.2: Loss-based MIA attacks' performance on original target models specified in Section 3.4; We can see the potential of performing high-precision attacks on all of these models. Because the highest precision for all of these models are over 70%. As well as a good general performance from the mean accuracy values and the AUC.

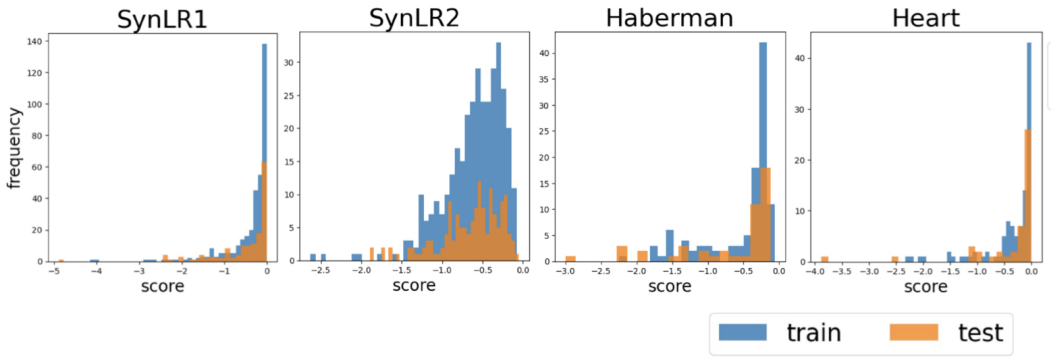

Figure 3.4: Distribution of the membership scores of attacking the training set (blue bars) and the testing set (orange bars) of the loss-based attacks on LR trained on the four kinds of data; All of these attacks have the thresholds for high precision, because of the existence of regions on the right-hand side of each graph where a moderate proportion of blue bars not being overlapped by the orange bars.

From the table, we can see that although loss-based attacks have a bad worse accuracy, it doesn't eliminate the chance for a high-precision attack at a tolerable accuracy.

Further, we can explain the reason for high precision attacks on the models by illustrating the distribution of the attack's score on the training set and the testing set as being inspired by Watson et al.[58].

From the loss distributions for attacks on each of the original models, we can see each of them has a section where only blue bars exist. Meaning most of the attack results going to be correct if the attacker could set a threshold there and classify all the data above this threshold as inside the training set. On the other hand, if the threshold selected has strong overlapping of the blue and the orange above that point, then it's hard to distinguish and hence results in a low precision attack.

3.5.2.2 Calibrated Loss Attacks

The table below records the performance of calibrated loss attacks on the target models using the same metrics as previous.
Similarly, the loss distribution of the calibrated attacks is plotted to explain the potential improvements brought by calibration.

Calibrated Loss Based Attack on Logistic Regression Models				
DataSet	Highest Precision w Recall	Mean Acc	Min Acc	AUC
SynLR1	0.722,0.837	0.699	0.656	0.225
SynLR2	0.759,0.686	0.696	0.628	0.423
Adult	0.707,0.386	0.699	0.458	0.435
Haberman	0.729,0.981	0.702	0.699	0.151
Heart Disease	0.706,0.953	0.690	0.514	0.361

Table 3.3: Calibrated Loss Attacks on the target models; After calibration, we can see a clear improvement in the attack's recall with the highest precision unaffected much. As well as a lift in the minimum attack accuracy. But a limit in improvement on the attack's AUC.

However, the calibration's capability differs in logistic regression from neural networks. To see the difference in the effectiveness of calibration on these two kinds of models. We trained a convolutional neural network with the same architecture as the CNN used in the study of Watson et al.[58] on CIFAR10 for comparison.

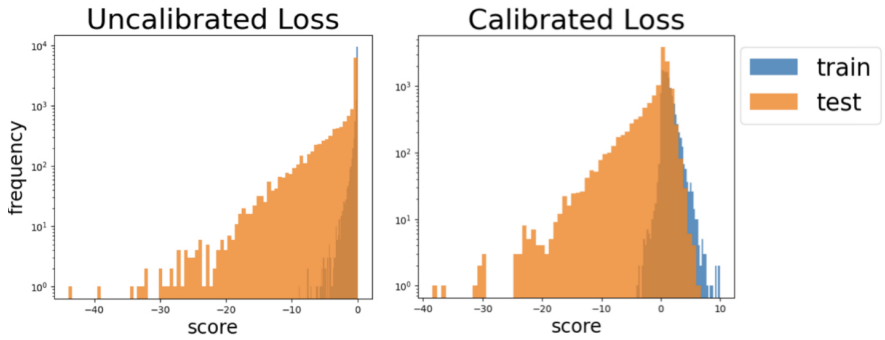

Figure 3.5: LHS: Uncalibrated loss attack and **RHS**: Calibrated loss attack on the CNN trained on the CIFAR10 dataset; Calibrate attack difficulty to the membership score improve the attack precision significantly on neural network.

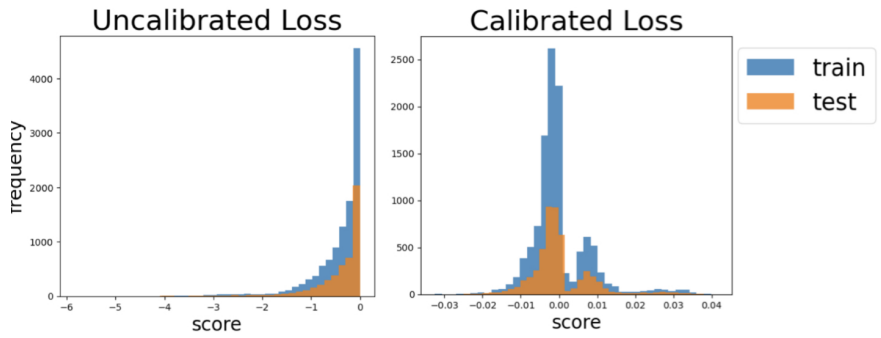

Figure 3.6: LHS: Uncalibrated loss attack and **RHS**: Calibrated loss attack on the LR model trained on the *Adult* dataset; Compare to the significant improvement on CNN, the effect of difficulty calibration is less obvious on logistic regression.

While on logistic regression, it's still able to observe an overlap in the loss distribution after difficulty calibration. Unlike the clear separation in the case of CNN. Thus, empirically demonstrate the drop in calibration's effectiveness on logistic regression models.

Comparing the performance on both uncalibrated and calibrated loss attacks. We decided to use the uncalibrated loss-based attacks in the experiments for the reason: 1. Avoid potential drawbacks in difficulty calibration; 2. Computational efficiency without calibration; 3. Uncalibrated loss attacks can also yield high-precision attacks.

3.5.3 DP's Protection Performance

Since experiments on testing H3 require privacy protection on the target model from differential privacy. *Diffprivlib* library [21] was used to efficiently introduce noises on the logistic regression model through objective perturbation. We had two metrics to examine the effectiveness of this DP mechanism. The first is to compare the model's average training accuracy before and after the privacy protection due to the privacy utility trade-off[23]. Averaging is vital due to randomness in DP. And it is achieved by training and averaging the result over 1000 models at the same privacy level quantified by ε.

Privacy Preserved Target Models' Average Training Accuracy (%)							
Dataset	Epsilon						
	0.1	5	10	25	100	1000	None
SynLR1	50.825	50.645	82.708	82.930	83.415	83.434	83.428
SynLR2	50.513	50.662	65.945	66.787	66.653	66.781	66.857
Adult	58.662	80.029	79.741	80.095	80.135	80.142	80.142
Haberman	49.628	62.069	73.895	75.317	75.727	75.700	76.636
HeartDisease	50.952	51.146	80.593	85.096	85.273	84.910	84.761

Table 3.4: Objective Perturbation on logistic regression models' average training accuracy; We can see that with the increase in epsilon's value from 0.1 to 1000, the training accuracy will tend to be closer to the model without differential privacy.

Another popular way to measure the effectiveness of DP is through membership in-

ference attacks [12, 43, 23]. Thus, the following distribute the number of training data points exposed in the privacy-preserving models trained at each level of ε for a collection of epsilon values. Comparing the distribution for each level of ε, we see

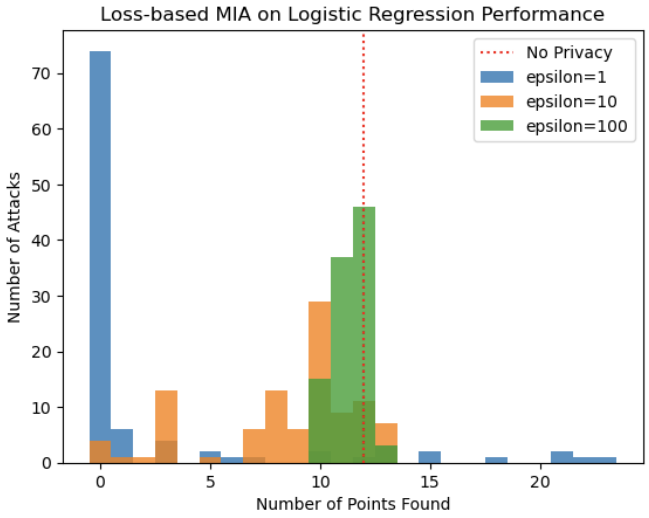

Figure 3.7: Distribution of the frequency (y-axis) of the attacks exposing each number of points (x-axis) on models with each ε-DP; The dotted red line indicate the number of points found in the non-private model; The bigger the ε, the closer the distribution gets to the non-private model. Whereas the smaller the ε, the more frequent it is to have a model with less number of data points exposed.

that the method we used to add differential privacy protection on the target models is effective since the trend at each level of ε follows the property of DP – the higher the epsilon, the weaker the protection. Thus, the objective perturbation is feasible to examine our hypotheses.

Figure A.8 and Figure A.9 in Appendix A.4 illustrate point-wise privacy protection from DP to MIA to demonstrate the randomness in such protection method for further interests.

Use of APSR in DP: From the unstable distribution of DP's covering effect as illustrated above, we can also see the randomness in this privacy protection. Thus, when examining any ε-DP's point-wise privacy protection, it's unavoidable to calculate the APSR (Average Protection Success Rate) for these points to determine whether they were covered successfully. The default setting of the number of repeats N in this calculation is 1000 for the logistic regression model. Since the model itself is deterministic, which makes DP is the only source of randomness in the experiments. In other words, we count training data in the target model as being covered successfully by an ε-DP. If and only if the data point can not be inferred by loss-based MIA on this model with more than 500 times.

Chapter 4

Results and Analysis

This chapter visualizes and analyses the experimental results from the previous chapter to provide answers to our three research questions. We also include additional experiments and results to thoroughly investigate some of our main contributions, followed by miscellaneous results discovered during our exploration.

4.1 Pattern of Privacy Violation and Data Vulnerability

We can see the pattern of privacy violation on the training set for logistic regression models on the illustrations of the results from experiments on H1 (Fig4.1). The pattern suggests that vulnerable data is not always the out-of-distribution ones, and they spread along the decision boundary.

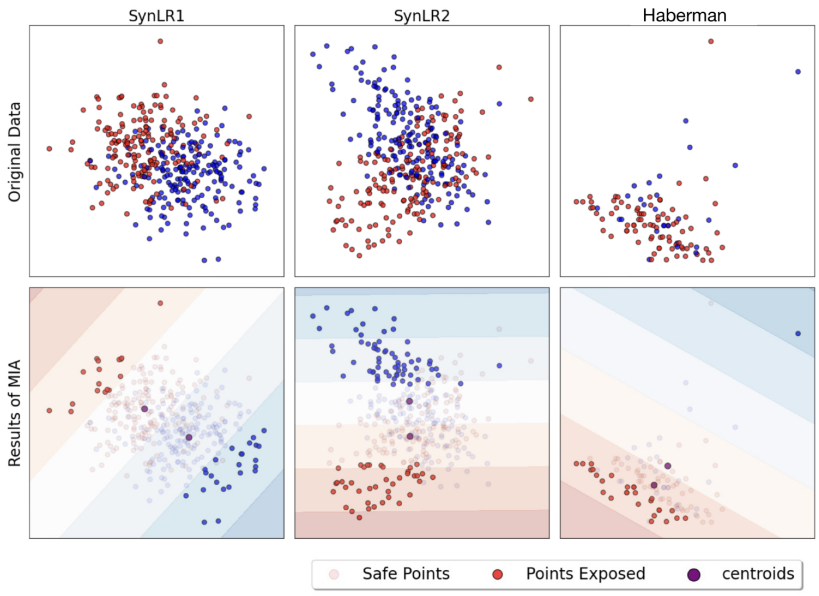

Figure 4.1: Top Row: The training set input to the logistic regression models: *SynLR1*, *SynLR2*, *Haberman*; **Bottom Row**: As a result of loss-based MIA on the training sets, the data being exposed are scattered with a denser colour than the data that is safe from the attack. The decision boundaries of the models are plotted at the back for observation.

Observing the pattern of privacy violation gives us insights into which part of the training set might be more vulnerable to membership inference attacks (MIA) than others. In the case of logistic regression models, the ones that are easy to separate with a line are more vulnerable since the loss-based attack marks the easy-to-separate data using higher membership scores. The violation spreads along the decision boundary because data along would bear similar prediction loss, making the loss-based attack identify these points altogether. Thus, the pattern of exposure in MIA gives us some insights into the reasons for these exposures and helps us interpret data vulnerability within the context of the problem.

We also observe from the pattern that MIA does not violate the privacy of some specific outliers in the training set of logistic regression, which are the ones in the cluster of data of the other label. This behaviour is explainable by the outliers on the side of their classes will bear a lower prediction loss (higher membership score) than the outliers on the other side. Hence, the pattern of privacy violation lets us know that the data's degree of outlying can not solely explain data vulnerability in logistic regression models.

In summary, the pattern of privacy violation in MIA on logistic regression models is relevant to understanding data vulnerability. From this pattern, we can identify vulnerable parts of the training set and gain insights into how MIA exploits privacy in the specific context of the data and model. This knowledge can inform measures to protect against MIA, making the pattern of privacy violation a crucial factor to consider in assessing data vulnerability.

4.2 Privacy Effects of Data Removal

Illustrations on the results from experiments on H2 suggested that removing data points from the training set of logistic regression models would result in more complex privacy effects than the *Privacy Onion Effect*. The privacy effect has three cases: 1) when data removal helps cover the previously-exposed data points, 2) when the removal has no effect on the data points remaining, and 3) when the removal leads to new exposure (Privacy Onion Effect). This observation empirically suggests the potential that data removal might still be helpful to the remaining training set privacy.

Additionally, we found that we can not use the location of the data removed from the training set to explain the cause of the privacy effects since each effect happened regardless of where we take out the data from the training set. Evidence using training sets sampled from *SynLR2* distribution is in Fig 4.2. This finding emphasises that data removal from the training set of a model is risky to the privacy of the remaining data because the removal of even random data points can cause these privacy effects on the rest of the training set.

However, we might be able to explain the privacy effects in more detail if we consider the shifts in the decision boundary caused by data removal from the training set since we previously suggested a relationship between data vulnerability and decision boundary in the context of logistic regression. Besides, shifts in the decision boundary might also be one of the reasons why the location of the removed data can not be a cause of the privacy effects: Is it perhaps a specific kind of shift in the decision boundary leads to

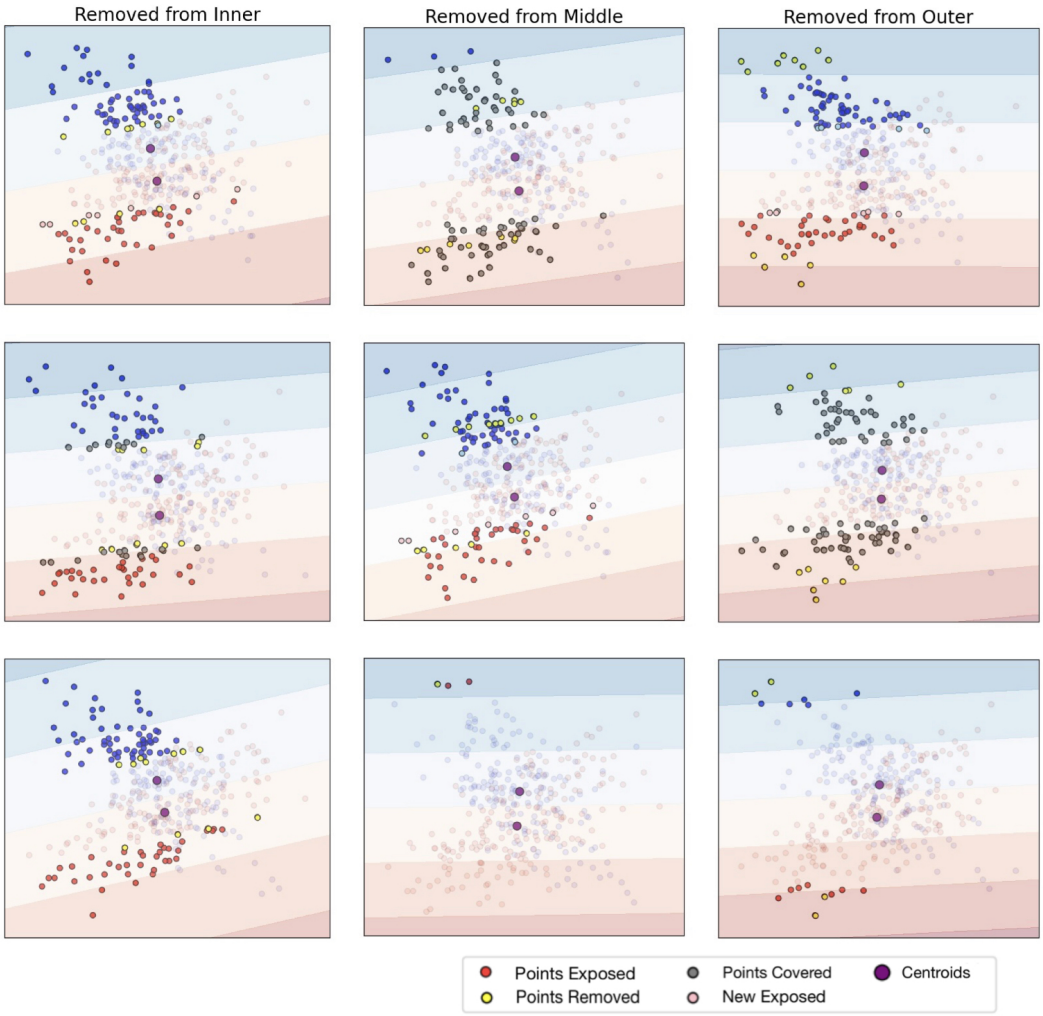

Figure 4.2: Each plot demonstrates the result of MIA and privacy effects as a result of 10% of the previously exposed data points being removed from the training set of data *SynLR2*. Each column indicate one location to remove data from the training set and it shows that the three privacy effects occurs regardless of the location of the points removed used-to-be.

each type of privacy effect? This hypothesis leads us to consider whether the decision boundary contributes to the privacy effects, which motivates further investigation. We conduct extended experiments, with the results stated in the coming section.

4.2.1 The Boundary was Shifted Closer to Most of the Privacy-Affected Data Points

We focused our analysis on the shifts in the decision boundary that affect the privacy of newly-exposed and newly-covered data points because these effects have an impact than having no effect. Let's denote the data points affected by the two targeted privacy effects as "privacy-affected data points." We can measure the shift in the decision boundary resulting from these privacy effects by using the distances of the privacy-affected data points to the decision boundary.

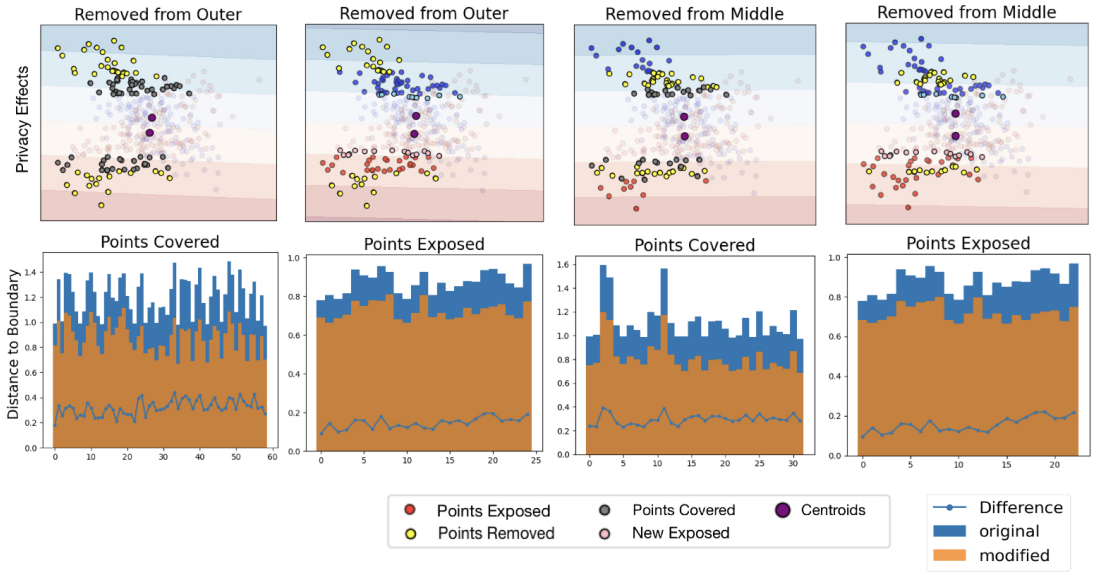

Figure 4.3: Top Row: As a result of removing 10% of the most-outlying or less-outlying sensitive data points from the training sets of *SynLR1*, both the cases would cause newly exposed (light pink and light blue) and newly covered (grey); **Bottom Row**: The bars in the plots represent the change in the distances of the data points that have their privacy being affected to the decision boundary. Where each bar represent one single data point. The change can be seen as the difference between the blue bars (before removal) and the orange bars (after removal). Blue bars are over the orange bars most of the time, meaning the boundary was often shifted closer to the remaining data after the training set is modified.

As a result, we found that most of the privacy-affected data became closer to the boundary than before (Fig 4.3). Exceptions also exist, like in one case when both privacy effects occurred (Fig 4.4), and the decision boundary shifts in a way that makes it moves closer to the new covered and away from the new exposed. Hence, based on our current observation, it is hard to draw detailed conclusions on the movement that lead to each kind of privacy effect. But we can conclude that data removal from the training set leads to a shift in the decision boundary, which might lead to the three privacy effects. Considering the proportion of data removed (10%) is not big enough to see concrete results in our experiments. We planned the following experiments for thorough inspections.

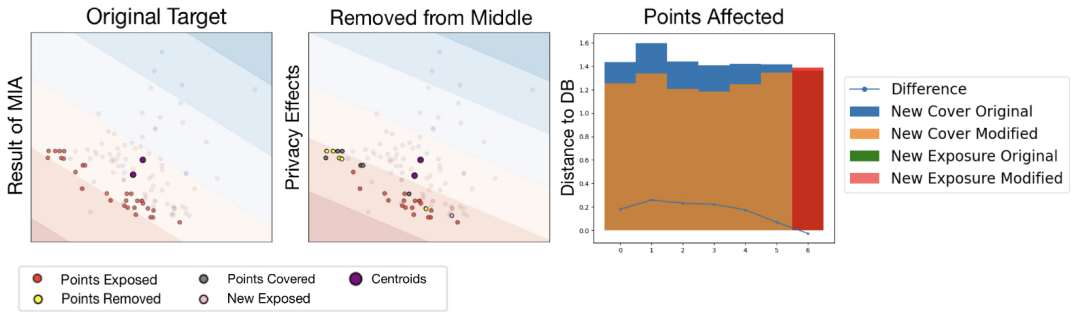

Figure 4.4: A rare case was observed on a training set of *Haberman* data: when two privacy effects happen on the same training set, the decision boundary is shifted in a way that it's closer to the points covered and away from the point exposed.

4.2.2 Data Removal within a Range Might Improve Privacy

We continue the experiments on data removal's impact on privacy with an increasing proportion from 20% to 100% because of our previous limitation in the results due to the removal of only 10% of the sensitive data points.

As a result of removal with increased proportion, we see big-scaled shifts in the decision boundary most of the time shown in Fig 4.6 and Fig 4.7. Exceptions exist like (Fig 4.5) when removing the entire set of previous-exposed data impacts little on the boundary. Thus, it is hard to guarantee a positive correlation between the removed data size and the scale of the shift in decision boundary. But we confirmed that the bigger the set gets removed from the training set, the more likely it is to have a big-scaled shift in the decision boundary.

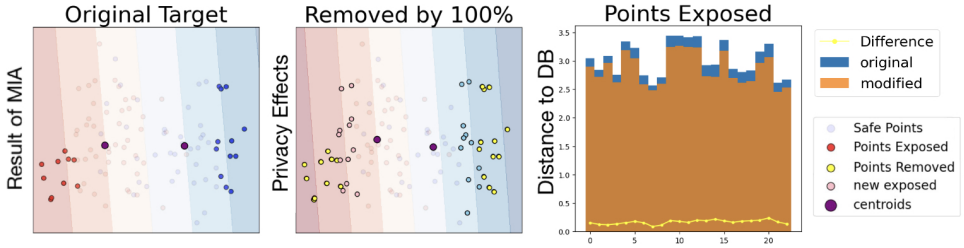

Figure 4.5: **LHS**: The data points in denser colours are exposed as a result of MIA on the original training set sampled from *Heart* data; **Middle**: Removing all the data exposed in the original training set creates new exposure in the remaining training set; **RHS**: Privacy effect in this context still cause a small change in distances of the newly exposed data points to the decision boundary before and after the removal action.

We also found that when removing data according to a proportion within a range for the context of each training set, the removal would only cover previously exposed data points with no new exposure. Evidence is in Fig 4.6 and Fig 4.7 where in each case the range is from 40%-90% in Fig 4.6 and 30%-80% in Fig 4.7 respectively. This finding suggests that data removal might be helpful to the privacy of the remaining training set if such a range can be observed and utilised in practice.

However, it's worth noticing the decrease in model performance might be tremendous when large numbers of data get removed. Like in the case of Fig 4.6, the drop in training

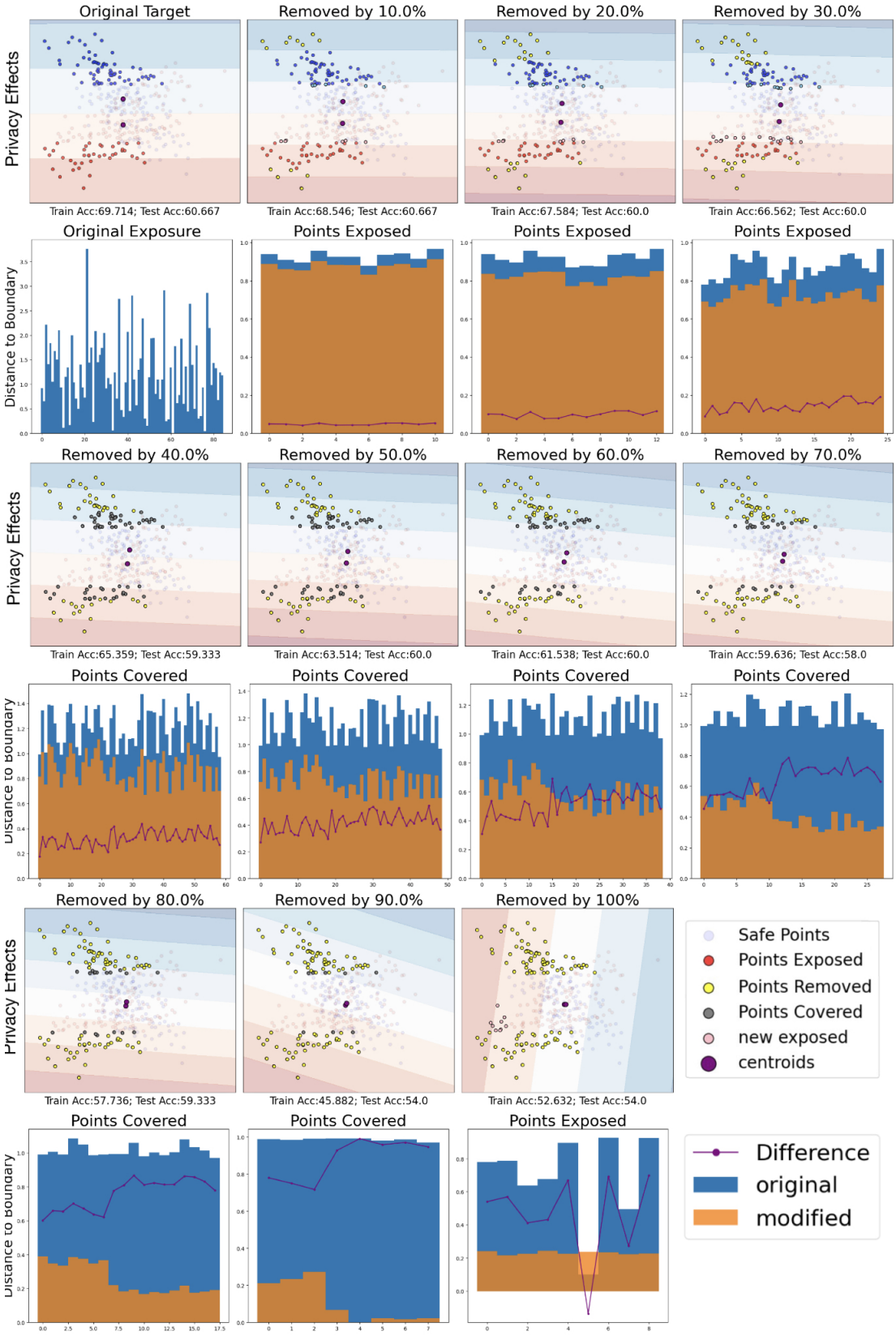

Figure 4.6: As the proportion of sensitive data points gets removed from a training set of *SynLR1* increased. We see that the privacy effect on the remaining training set gets stabilised to 'covering old exposure' when the proportion is within 40% to 90%. The gap in the distances of the affected data points to the boundary caused by training set modification gets bigger shown as a growing purple line.

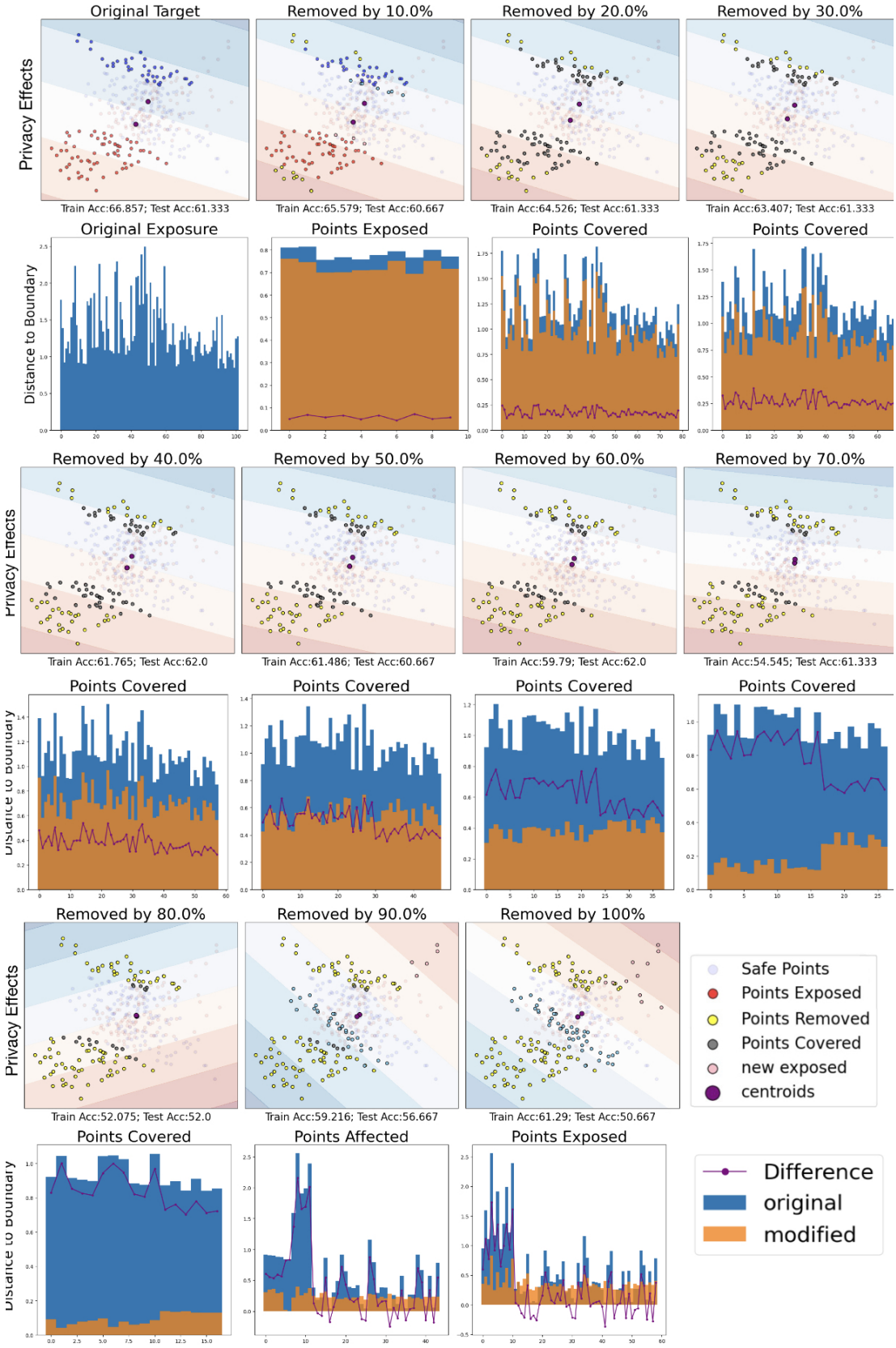

Figure 4.7: Same experiments as above on dataset *SynLR2* convey the same conclusion that when the proportion of sensitive data is kept within a range, this case:30%-80%, the privacy effect would be stabilised to 'covering old exposure'.

accuracy could be as most as roughly 17% when we remove the entire set of sensitive data. Thus, practitioners should notice the trade-off in utility when withdrawing data from the training set for the potential improvement in the training set privacy.

Concluding our exploration of the privacy effects caused by data removal in the training set, we confirmed the shift in the decision boundary leads to the privacy effects. The number, but not the location, of removed data points, may be related to privacy effects, and in certain scenarios, removing data within a specific range can enhance privacy protection. Identifying this range and the corresponding impact on model performance can be valuable in practice.

4.3 Data Vulnerability and Difficulty to be Protected by Differential Privacy

The vulnerability of data points to MIA attacks is not uniform when differential privacy is applied as a protection, as evidenced by our experiments on H3 (Fig 4.8) shows a bubbling effect in the protection strength, which is a difference in the level of privacy protection needed by individual data in a training set – easy-to-protect data points centralize around the decision boundary, the hard-to-protect data points locate more outlying. This finding demonstrates the susceptibility of data points to differential privacy protection differs in the training set, underscoring the uneven risks MIA poses on these data points.

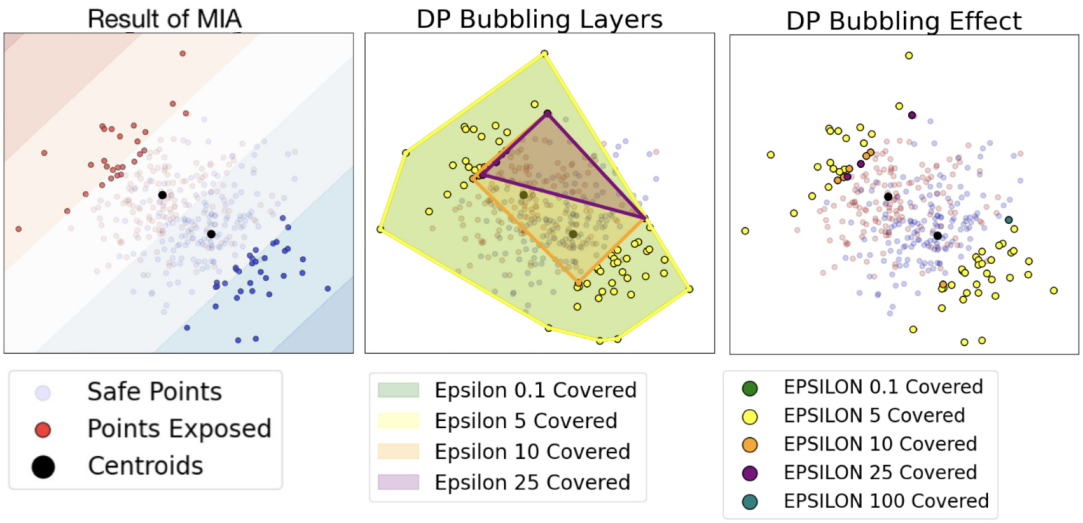

Figure 4.8: LHS: the result of MIA attack on logistic regression trained on a training set of *SynLR1* data; **Middle**: convex hulls on the data exposed that is coverable at each level of DP; **RHS**: the data points covered at each level of DP for more detailed inspections on covering effect; These plots shows that the protection of DP comes in layers: the bigger the privacy parameter ε, the weaker the protection and it's only capable of covering a smaller range of points concentrated around the decision boundary.

We might be able to explain the bubbling cover effect in protecting the training set by relating this effect to the resultant shifts in the model's decision boundary based on the conclusions we had so far on the relationship between data vulnerability and decision

boundary. Therefore, we conducted experiments to investigate this relationship further, and the results of these experiments are presented in the section below.

4.3.1 Protection Explained with Boundary Shifts

We can use the shift in decision boundary at each level of differential privacy to help understand the bubbling cover effect since the shifting scale for the decision boundary of the DP-private model at each level distinguishes, shown in Fig 4.9. This difference in shifting scales might be a reason for each layer of protection consisting of different data points in the training set. Additionally, the stronger the DP, the bigger the shifting scale in decision boundary and contra versa. Thus, we conclude that when DP cause a big-scaled shift in the decision boundary of the target model, it is more likely that this level of DP can cover more hard-to-protect data points in the training set.

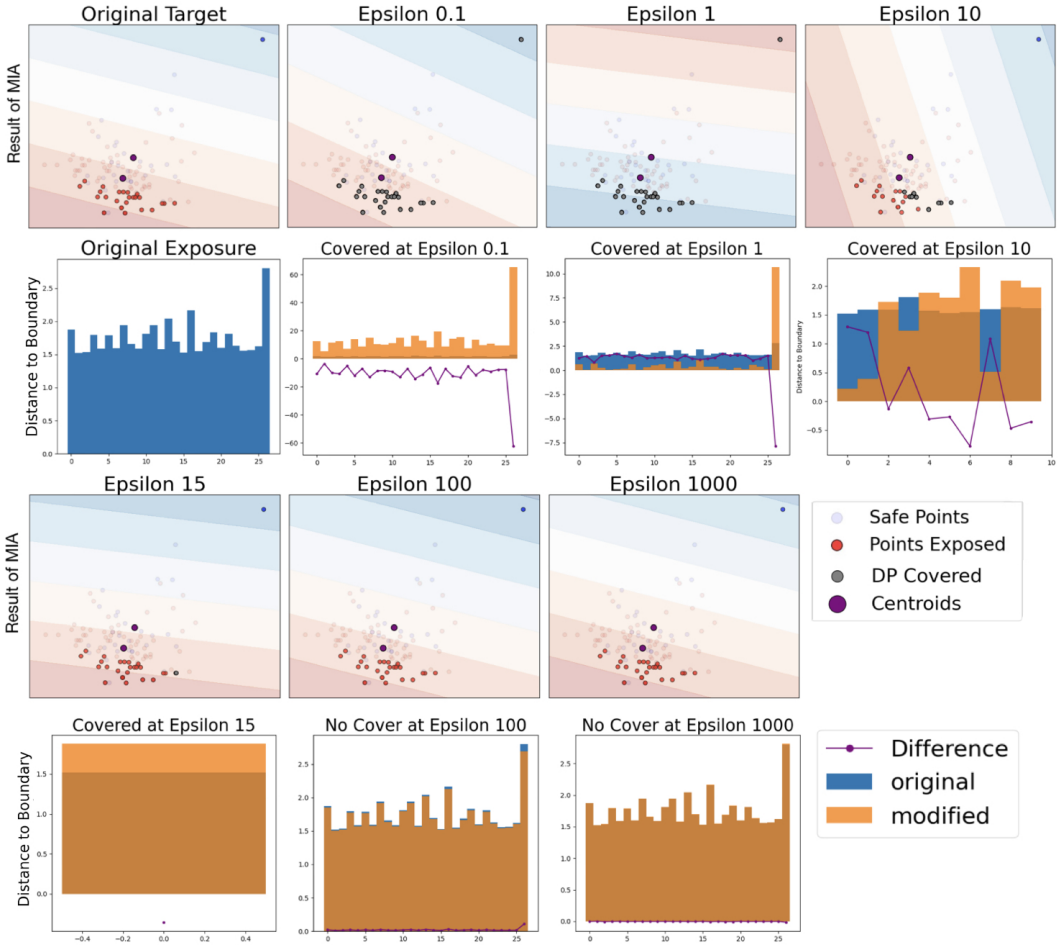

Figure 4.9: The scatter plots illustrate the point-wise result of MIA on the original and the DP-protected logistic regression models trained on *Haberman* data; The histograms illustrate the distance of the data points that were protected by a level of DP (or the data not being protected, if DP at that level failed to protect any) to the decision boundary of each model. We can see that the stronger the DP (smaller ε), the bigger the influence it brought to the decision boundary (shown as the big fluctuation in the purple line) while it protects more data from exposure and contra versa.

While we have observed shifts in the decision boundary caused by differential privacy,

we cannot determine a shifting direction for a successful DP protection in covering privacy violations in the training set, as both cases of moving closer or further away from covered points exist. Nevertheless, our understanding of the impact of differential privacy on the model's decision boundary at different scales allows for a new evaluation of its effectiveness in logistic regression. By observing the magnitude of the shift, we can estimate the protection's efficiency, which could be valuable for future research.

4.4 Data removal with Differential Privacy

Our previous experiments revealed that removing data points is not always sufficient to address privacy leakage in the training set. To resolve this, we suggest applying differential privacy (DP) on the model trained on the modified training set. As a result, we found that data removal aids DP's protection on the remaining training set in the following ways.

4.4.1 Weak Differential Privacy Suffices to Protect Newly Exposed Data Points

We observed that the scale of DP required to cover newly exposed data points is typically small, as shown in Fig 4.10 where we didn't use epsilon values no smaller than 25 to cover all the privacy violations. This finding is consistent with our previous conclusion that inner data points are easier to cover with DP, confirming that protection difficulty differs in layers within a training set. These findings further support the idea that data points closer to the decision boundary are easier to protect with DP, making them less vulnerable to MIA.

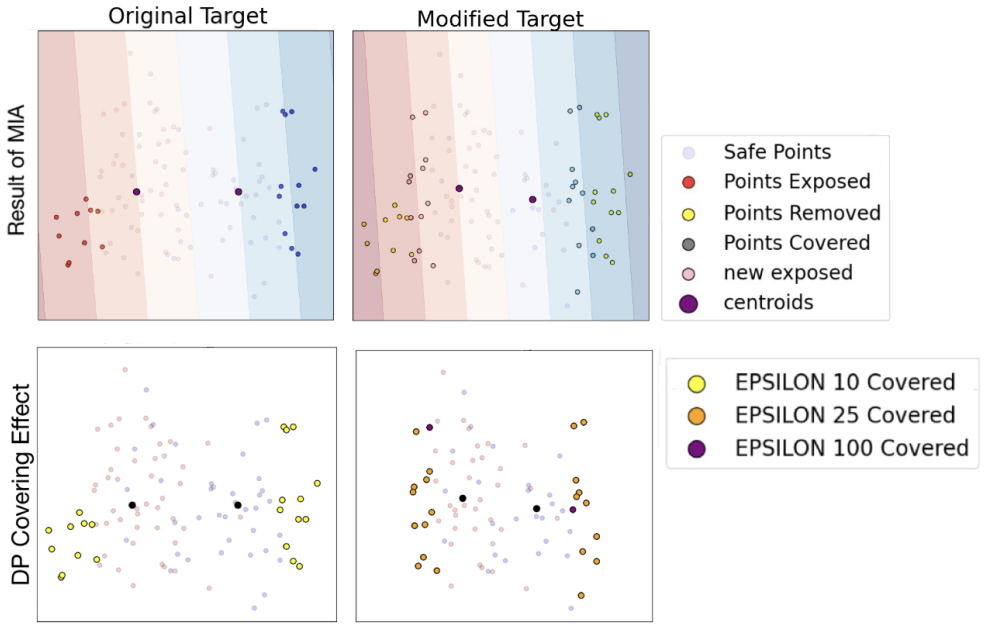

Figure 4.10: Top Row: When removing all of the data points exposed in the original target from the training set can still expose a new set of data points; **Bottom Row**: These newly exposed data points could be covered with a relevantly small scale of differential privacy with ε : 25&100.

4.4.2 Points Removal Supports Differential Privacy

Removing some of the previously unsafe data points from the training set can help DP cover the remaining training set more effectively. Our observation indicates that some other previously-unsafe data points can now be covered with weaker DP than before the removal, as demonstrated in Fig 4.11. This finding suggests that data removal is beneficial to privacy protection with DP.

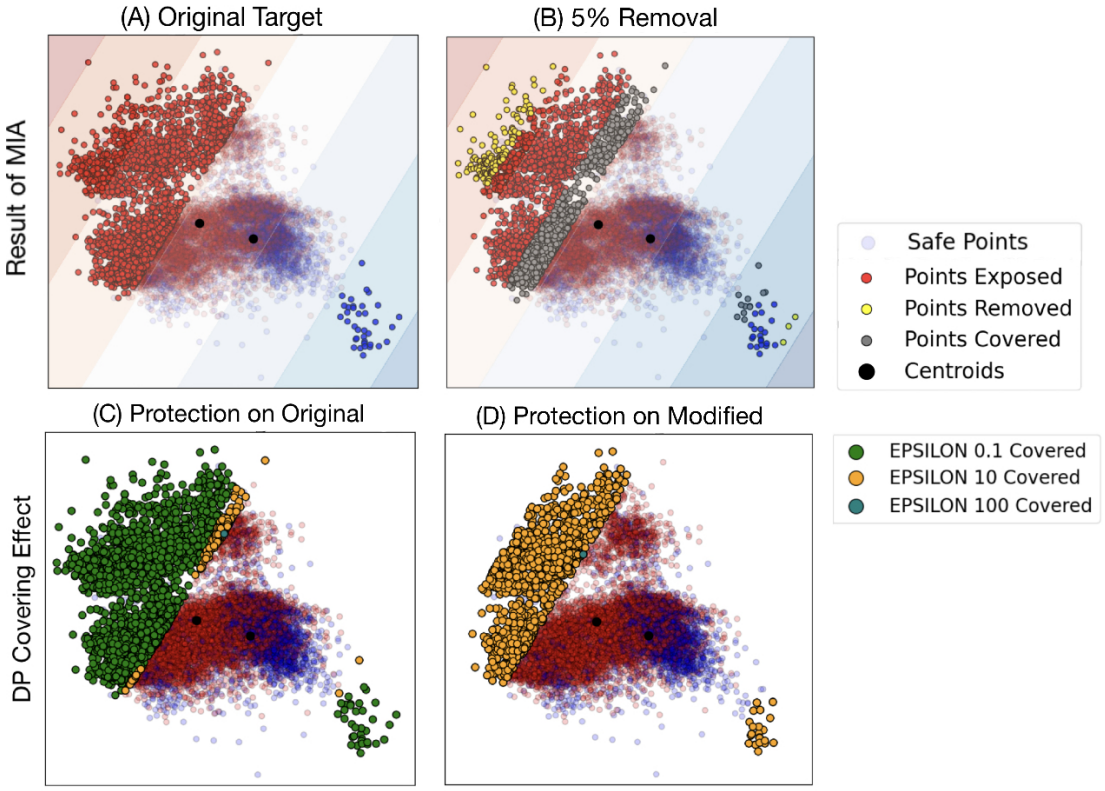

Figure 4.11: **(A)**: As a result of MIA attack on *Adult* dataset, the privacy of the points in denser colours are violated. **(B)**: As a result of the yellow data points being removed which consists on 5% of the data exposed shown in (A), it covers some of the previously unsafe data points (the grey points). **(C)**: The bubbling cover effect of DP on the exposure in (A). **(D)**: The bubbling cover effect of DP on the exposure in (B). Comparing (C) and (D), we can see that a lot of the data points that used to be protected with DP at $\varepsilon = 0.1$ can now be covered with $\varepsilon = 10$.

Furthermore, both experiments in Section 4.4.1 and Section 4.4.2 removed the hard-to-protect (the outermost lying) vulnerable data points from the training set, and they both proved to be beneficial to the privacy of the data remaining or to the protection efficiency of DP. This emphasised the importance to identify and remove the hard-to-protect instances that are vulnerable to MIA for wider privacy benefits for the remaining majority.

4.5 Miscellaneous Results

This section went through the findings we got during the experiments that are not very related to our main focus but are worth to be discussed in the main report. Figures mentioned in this section can be found in Appendix A.

4.5.1 The Public-Private Splits Influence the Privacy Effects of Data Removal

We found that for each set of public and private training sets generated from the same distribution, we can get distinctive results in the privacy effects as a result of removing data points based on the same criteria. Evidence can be found in Fig A.1 in Section A.1.

4.5.2 Outlier Definitions Based on Cluster Centroids and Decision Boundaries Yields Similar Conclusion

Because of the discovery that data vulnerability is more relevant to the decision boundary of the model. Our main results from the experiments involving data removal use data points' distances to the decision boundary to quantify their degree of outlying in the training set. While the same experiments were conducted on the other definition, we found that the results are similar. Evidence can be found in Section A.2.

4.5.3 Calibrated Loss May Not Enhance Attacks on Logistic Regression

We empirically found that calibrating each data point's difficulty to be attacked to their membership scores in loss-based MIA does not improve the attack performance a lot. Evidence can be found in Section 3.5.2.2. But our observation is empirical, further inspections are needed to draw concrete conclusions on the efficiency of calibrated attacks on logistic regression.

4.5.4 Challenges in Neural Networks and some Preliminaries

We also empirically examine the challenges in examining the problem of data vulnerability in the context of a neural network with a simple Multi-layered-perception model (MLP). As we mentioned in Chapter 1, the challenges came from the uncertainty in the attack results on the original target models. Evidence can be found in Figure A.6 in Section A.3.

We preliminarily identify that data points in the training set for the MLP model bear different difficulties to be protected by DP, shown in Fig A.7. These observations also suggest that it is easier to protect the more inner-lying data points than the outliers. But it is computationally costly to arrive at a more concrete conclusion because of the uncertainty in the attack results combined with randomness in DP.

Chapter 5

Conclusion

This chapter summarizes our conducted experiments and evaluated results. We also discuss the limitations of our work and suggest directions for further research.

5.1 Summary

In this report, we present our research on understanding data vulnerability to membership inference attacks (MIA) on logistic regression models trained on low-dimensional datasets. We focused on three main research questions: 1) Are outliers more vulnerable than others in logistic regression? 2) Will removing certain data in the training set reveal a new set of data points to the attacker in logistic regression? 3) Can we use differential privacy (DP) to explain data vulnerability?

Firstly, we discovered that certain outliers in the training set are as vulnerable to MIA as inliers from the pattern of privacy violation of MIA on the target models' training set. Secondly, we observed that removing these outliers could lead to the three privacy effects:*expose previously safe data*, *hide previously unsafe data*, or *have no impact* on the remaining training set, unlike the previous study which suggests data removal only leads to new exposure. What's more, we found that the consequence of random data removal is the same as removing these outliers, which we found currently hard to explain. Thirdly, we found that data vulnerability is explainable with the scale of DP individual data possessed in the training set. The required scale of DP varies by instance, indicating varying levels of difficulty to protect and vulnerability to MIA. We also found this variation is relevant to a shift in the model's decision boundary. Last but not least, we found the removal of a proportion of hard-to-protect vulnerable data points from the training set supports DP's privacy protection in the remaining training set, and this proportion is specific to the model and the training set. In conclusion, our results from the thorough investigation foster our understanding of data vulnerability to MIA in the context of logistic regression models trained on low-dimensional datasets.

5.2 Limitation and Future Works

Our findings are limited by assuming a balance between the number of accessible data to the attacker in MIA and the number of private data used to train the model. This assumption failed to consider a more real-world scenario where the attacker might be able to access significantly more data than the size of the training set. Thus, future researchers can gradually increase the size of the public dataset accessible to the attacker and evaluate its impact on data vulnerability.

Another limitation is that we only evaluate data vulnerability by using prediction loss as the membership score for the loss-based attacks, which may not fully capture all aspects of data vulnerability to MIA. For a thorough inspection, we should consider other metrics to define membership score, such as the prediction confidence [47] and gradient norm [37] should be considered as future work.

In addition to addressing these limitations, our contributions can guide future research in data vulnerability to membership inference attacks. Firstly, we present the findings in a setting of logistic regression models trained on low-dimensional datasets to highlight the importance of studying data vulnerability within the context of each model and dataset. To gain a more comprehensive understanding of data vulnerability, we encourage future researchers to use our findings as a bottom line to explore data vulnerability with a gradual increase in the complexity of the problem settings. Secondly, we suggest future researchers continue on our empirical findings on the relationship between the privacy effects of data removal and the influence on the decision boundary of the target model, to develop a detailed insight into what kind of shift contributes to each privacy effect. For a better understanding of the reasons behind the removal of random data points from the training set resulting the same as removing outliers. Thirdly, we suggest the development of a more efficient privacy protection mechanism that removes the hard-to-protect (more outlying) vulnerable points and then protects the remaining data points with a weak level of differential privacy. Last but not least, we discovered that the privacy effects as a result of data removal on the remaining data points are more complex than the previous *Privacy Onion Effect*, suggesting that researchers should examine data vulnerability more comprehensively in their specific contexts.

Bibliography

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Charu C Aggarwal and Philip S Yu. Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 37–46, 2001.

[3] Per Austrin, Kai-Min Chung, Mohammad Mahmoody, Rafael Pass, and Karn Seth. On the impossibility of cryptography with tamperable randomness. In *Advances in Cryptology–CRYPTO 2014: 34th Annual Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2014, Proceedings, Part I 34*, pages 462–479. Springer, 2014.

[4] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020.

[5] Carl Bosley and Yevgeniy Dodis. Does privacy require true randomness? In *Theory of Cryptography: 4th Theory of Cryptography Conference, TCC 2007, Amsterdam, The Netherlands, February 21-24, 2007. Proceedings 4*, pages 1–20. Springer, 2007.

[6] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

[7] Sébastien Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 15, 2014.

[8] Chensi Cao, Feng Liu, Hai Tan, Deshou Song, Wenjie Shu, Weizhong Li, Yiming Zhou, Xiaochen Bo, and Zhi Xie. Deep learning and its applications in biomedicine. *Genomics, proteomics & bioinformatics*, 16(1):17–32, 2018.

[9] Nicholas Carlini, Matthew Jagielski, Nicolas Papernot, Andreas Terzis, Florian Tramer, and Chiyuan Zhang. The privacy onion effect: Memorization is relative. *arXiv preprint arXiv:2206.10469*, 2022.

[10] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. *Advances in neural information processing systems*, 21, 2008.

[11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[12] Junjie Chen, Wendy Hui Wang, and Xinghua Shi. Differential privacy protection against membership inference attack on machine learning for genomic data. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 26–37. World Scientific, 2020.

[13] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[14] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008.

[15] Emiliano De Cristofaro. A critical overview of privacy in machine learning. *IEEE Security Privacy*, 19(4):19–27, 2021.

[16] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[17] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[18] Meherwar Fatima, Maruf Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.

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

[20] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 619–633, 2018.

[21] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. Diffprivlib: the ibm differential privacy library. *arXiv preprint arXiv:1907.02444*, 2019.

[22] Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

[23] Bargav Jayaraman and David Evans. When relaxations go bad:" differentially-private" machine learning. *arXiv preprint arXiv:1902.08874*, 2019.

[24] Gaganjot Kaur and Prinima Gupta. Hybrid approach for detecting ddos attacks in software defined networks. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE, 2019.

[25] Kalin S. Kolev. Convexity in image-based 3d surface reconstruction. 2012.

[26] J Zico Kolter and Marcus A Maloof. Learning to detect and classify malicious executables in the wild. *Journal of Machine Learning Research*, 7(12), 2006.

[27] Aleksandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 25–36. SIAM, 2003.

[28] Elizabeth D Liddy. Natural language processing. 2001.

[29] Wenjun Lu, Avinash L Varna, and Min Wu. Confidentiality-preserving image search: A comparative study between homomorphic encryption and distance-preserving randomization. *IEEE Access*, 2:125–141, 2014.

[30] Edward Lui and Rafael Pass. Outlier privacy. In *TCC (2)*, pages 277–305, 2015.

[31] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[32] Harsh Mehta, Ashok Cutkosky, and Behnam Neyshabur. Extreme memorization via scale of initialization. *arXiv preprint arXiv:2008.13363*, 2020.

[33] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[34] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes-which naive bayes? In *CEAS*, volume 17, pages 28–69. Mountain View, CA, 2006.

[35] Silvio Micali, Charles Rackoff, and Bob Sloan. The notion of security for probabilistic cryptosystems. *SIAM Journal on Computing*, 17(2):412–426, 1988.

[36] Abdullah-Al Nahid, Yinan Kong, et al. Involvement of machine learning for breast cancer image classification: a survey. *Computational and mathematical methods in medicine*, 2017, 2017.

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

[38] Tatsuaki Okamoto and Shigenori Uchiyama. A new public-key cryptosystem as secure as factoring. In *Advances in Cryptology—EUROCRYPT'98: International Conference on the Theory and Application of Cryptographic Techniques Espoo, Finland, May 31–June 4, 1998 Proceedings 17*, pages 308–318. Springer, 1998.

[39] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[40] Shuchao Pang, Anan Du, Mehmet A Orgun, and Zhezhou Yu. A novel fused convolutional neural network for biomedical image classification. *Medical & biological engineering & computing*, 57:107–121, 2019.

[41] AM Pires and JA Branco. High dimensionality: The latest challenge to data analysis. *arXiv preprint arXiv:1902.04679*, 2019.

[42] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[43] Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79, 2018.

[44] Vibhor Rastogi, Michael Hay, Gerome Miklau, and Dan Suciu. Relationship privacy: output perturbation for queries with joins. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 107–116, 2009.

[45] Jason DM Rennie. Regularized logistic regression is strictly convex. *Unpublished manuscript. URL people. csail. mit. edu/jrennie/writing/convexLR. pdf*, 745, 2005.

[46] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

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

[48] Dewi Retno Sari Saputro and Purnami Widyaningsih. Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). In *AIP Conference Proceedings*, volume 1868, page 040009. AIP Publishing LLC, 2017.

[49] Ravi Shah. Gradient descent and its types: Explained with examples. https://www.analyticsvidhya.com/blog/2022/07/gradient-descent-and-its-types/, 2022.

[50] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

[52] Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065, 2019.

[53] Fabio Alexandre Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)*, pages 2560–2567. IEEE, 2016.

[54] Elham Tabassi, Kevin J Burns, Michael Hadjimichael, Andres D Molina-Markham, and Julian T Sexton. A taxonomy and terminology of adversarial machine learning. *NIST IR*, 2019:1–29, 2019.

[55] Christian Tchito Tchapga, Thomas Attia Mih, Aurelle Tchagna Kouanou, Theophile Fozin Fonzin, Platini Kuetche Fogang, Brice Anicet Mezatio, and Daniel Tchiotsop. Biomedical image classification in a big data architecture using machine learning algorithms. *Journal of Healthcare Engineering*, 2021:1–11, 2021.

[56] Srikanth Thudumu, Philip Branch, Jiong Jin, and Jugdutt Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7:1–30, 2020.

[57] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[58] Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*, 2021.

[59] Xiaodan Xu, Huawen Liu, Li Li, and Minghai Yao. A comparison of outlier detection techniques for high-dimensional data. *International Journal of Computational Intelligence Systems*, 11(1):652, 2018.

[60] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big data*, 6(1):1–18, 2019.

[61] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

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

[63] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

[64] Benjamin Zi Hao Zhao, Aviral Agrawal, Catisha Coburn, Hassan Jameel Asghar, Raghav Bhaskar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson. On the (in) feasibility of attribute inference attacks on machine learning models. In *2021*

IEEE European Symposium on Security and Privacy (EuroS&P), pages 232–251. IEEE, 2021.

[65] Lina Zhou and Dongsong Zhang. Nlpir: A theoretical framework for applying natural language processing to information retrieval. *Journal of the American Society for Information Science and Technology*, 54(2):115–123, 2003.

Appendix A

First appendix

A.1 Miscellaneous Result 4.5.1

Figure A.1 illustrate the finding stated in Section 4.5.1.

A.2 Miscellaneous Result 4.5.2

Figures: A.2, A.3, A.4 and A.5, illustrate the findings stated in Section 4.5.2.

A.3 Miscellaneous Result 4.5.4

Figures: A.6 and A.7, illustrate the findings stated in Section 4.5.4.

A.4 Outlier Privacy

It's worth mentioning that Lui et al. [30] proposed a tailored-differential privacy mechanism that might be easily confused with our contribution. Their work introduced a novel privacy definition, called "Outlier privacy," which aims to address the privacy implications caused by outliers in the output of a calculation or learning model using their self-defined outlier definition. However, our contribution is on the broader topic of the vulnerability of data points in a training set for a machine learning model in general.

A.5 Differential Privacy's Randomness

Figure A.8 and A.9 demonstrated the randomness in DP's protection from the point-wise.

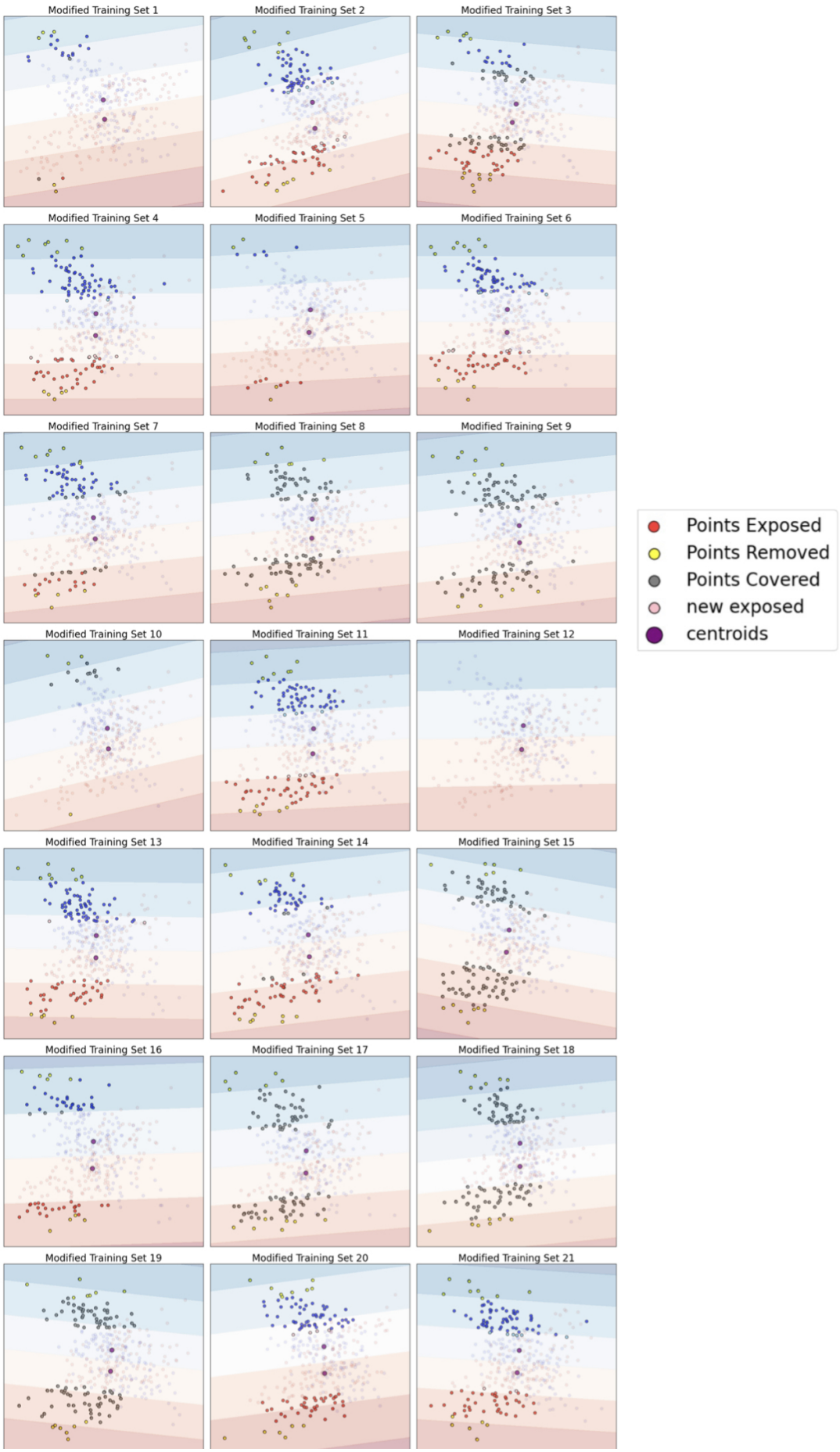

Figure A.1: Experiments of data removal on the different public-private split on *SynLR2* distribution; Each split results in a distinctive privacy effect from the privacy attack on the model trained on the modified training set.

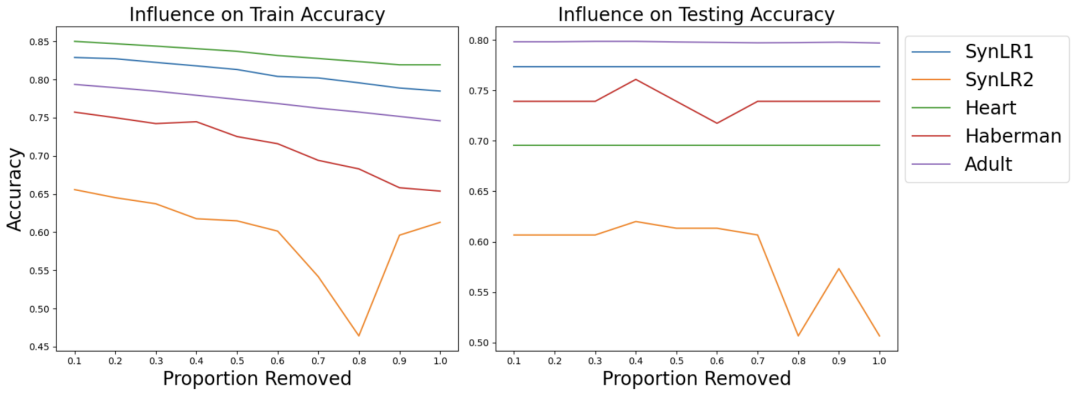

Figure A.2: LHS: drop in training accuracy at each proportion of data being removed; **RHS**: drop in the testing accuracy at each proportion of data being removed by their distance to the centre of each class; The effects of a training set modification on the utility of the model is similar to Fig 3.2.

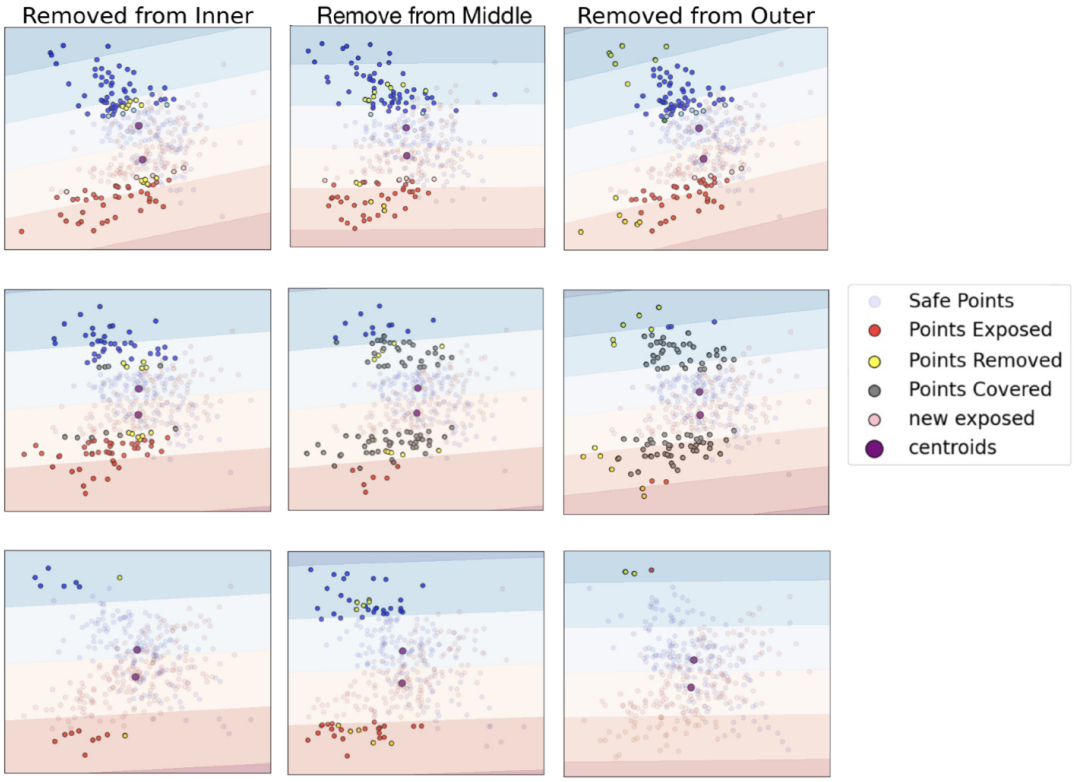

Figure A.3: Each plot represent a distinct training set sampled from *SynLR2* distribution with 10% of the previously-exposed training points being removed from **LHS**: the innermost; **Middle**: the middle; **RHS**: the outermost. With the degree of outlying measured by distances to the centre of each class. We can see that the same conclusion as Section 4.2 holds on this definition of degree of outlying as well.

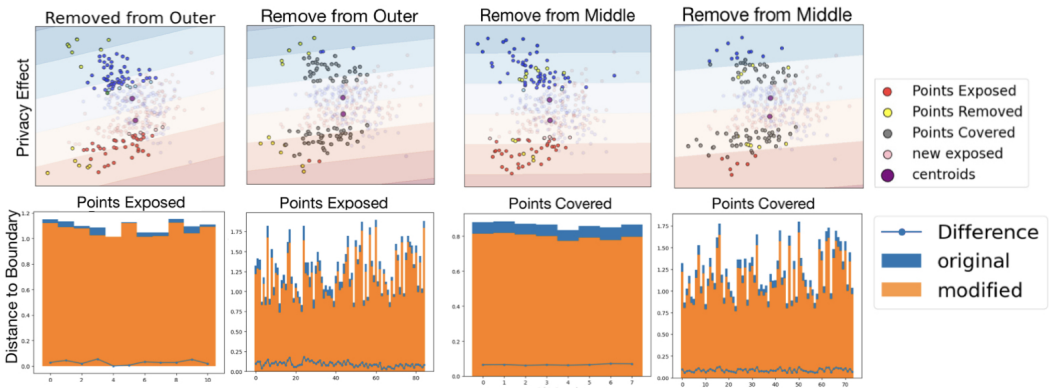

Figure A.4: Each scatter plot is a privacy effect resulting from removing 10% of data points at different degrees of outlying defined with the distances to the centre of each class, from the training sets selected from *SynLR1*. We can see that results from Section 4.2.1 also hold for this definition of the degree of outlying when data points' privacy is affected due to the training set being reduced. Those affected data were usually brought closer to the decision boundary.

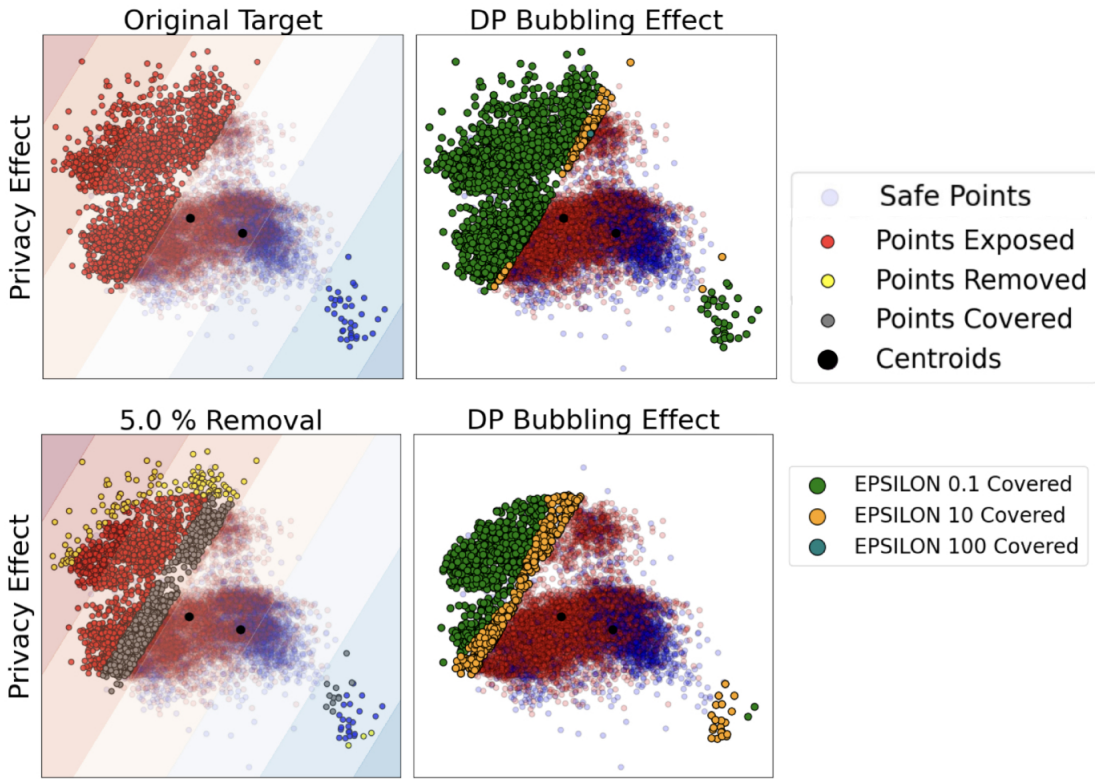

Figure A.5: LHS column: The privacy effect brought by removing 5% of the most outlying data points by their distance to the centres of their classes; **RHS** column: DP's bubbling covering effect on the corresponding exposure before and after the modification on the training set. Same conclusion as Section **??** can be drawn on removal by this definition of degree of outlying, that removing some of the most outlying data points helps the privacy of the remaining data points to be protected.

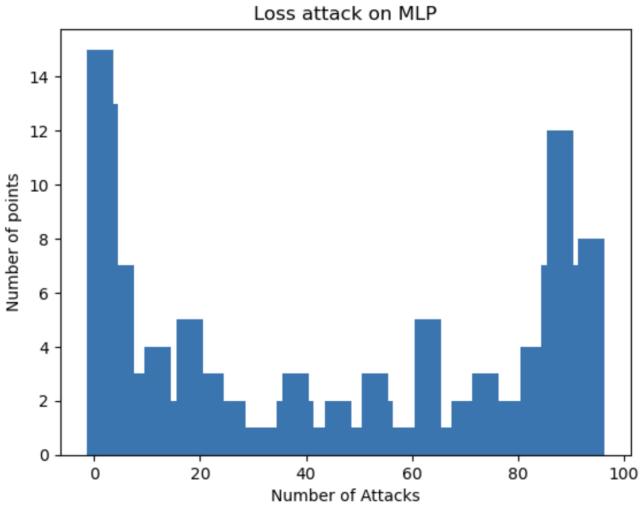

Figure A.6: Results of loss-based attack on MLP trained with SGD is non-deterministic; As we can see from the histogram of a number of data points being exposed (y-axis) vs the number of attacks having each number of exposure (x-axis), that each attack on a distinct model trained with the same setting differs.

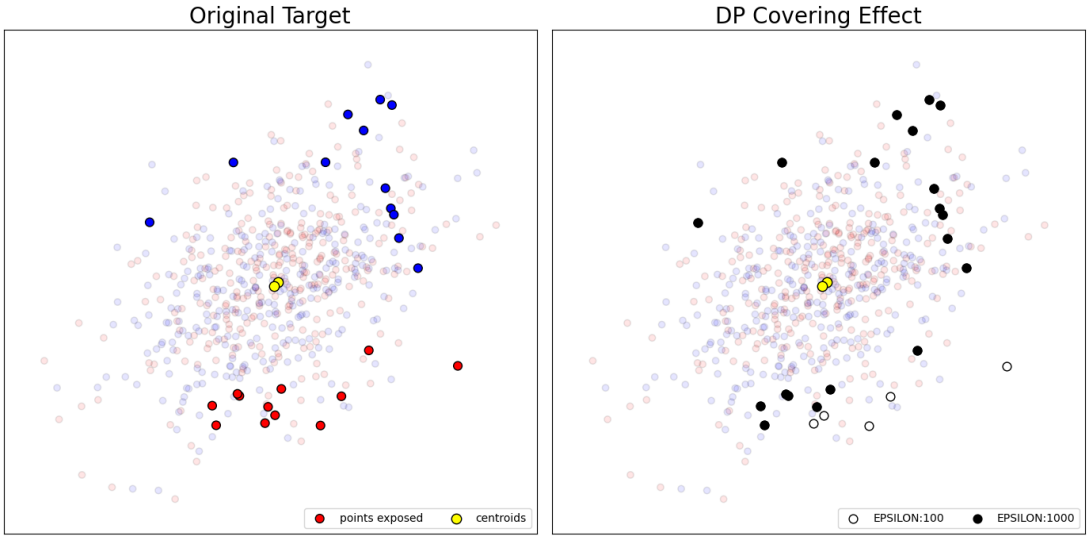

Figure A.7: LHS: The privacy of the data points with a denser colour was violated by loss-based MIA on Multi-Layer-Perception (MLP) models trained on a synthetic data set of Gaussian distribution. While the data safe from the attack were at the back; **LHS**: Use DP at different levels to cover the privacy violation shown in the RHS plot, showing us that data points in the training set for an MLP model are also not equally hard to be protected.

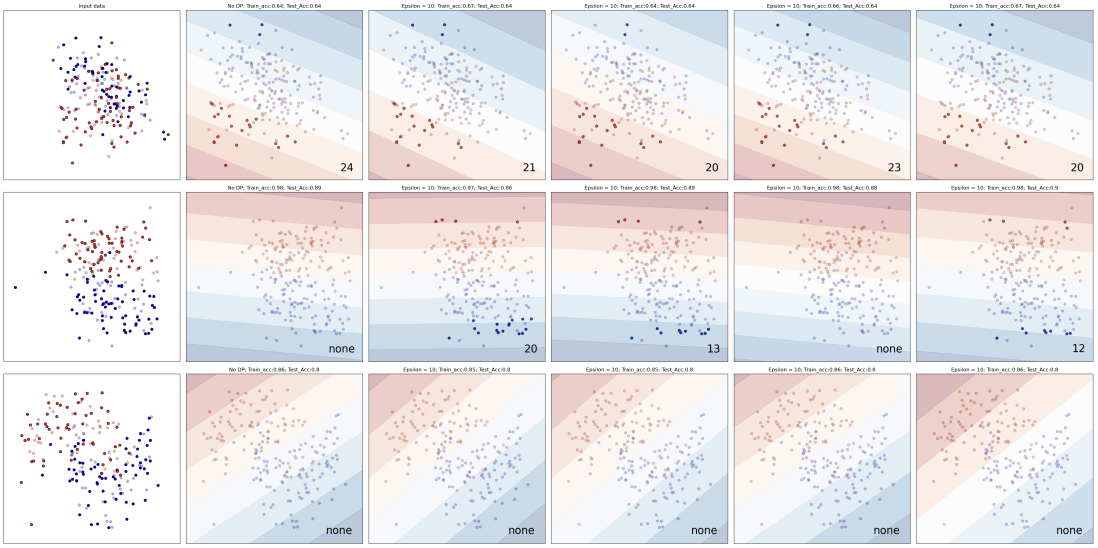

Figure A.8: Randomness in DP protection at epsilon 10 on the MIA on a model trained on the synthetic dataset: the first column plots the original datasets with colours used to distinguish the classes and each row is one dataset. Then each column afterwards is the attack result on one DP-private model. As a result of the attack, the data points with their privacy being exposed were scattered with a denser colour at the front with the total number noted at the bottom right corner. The decision boundary of each DP LR is also plotted in the background.

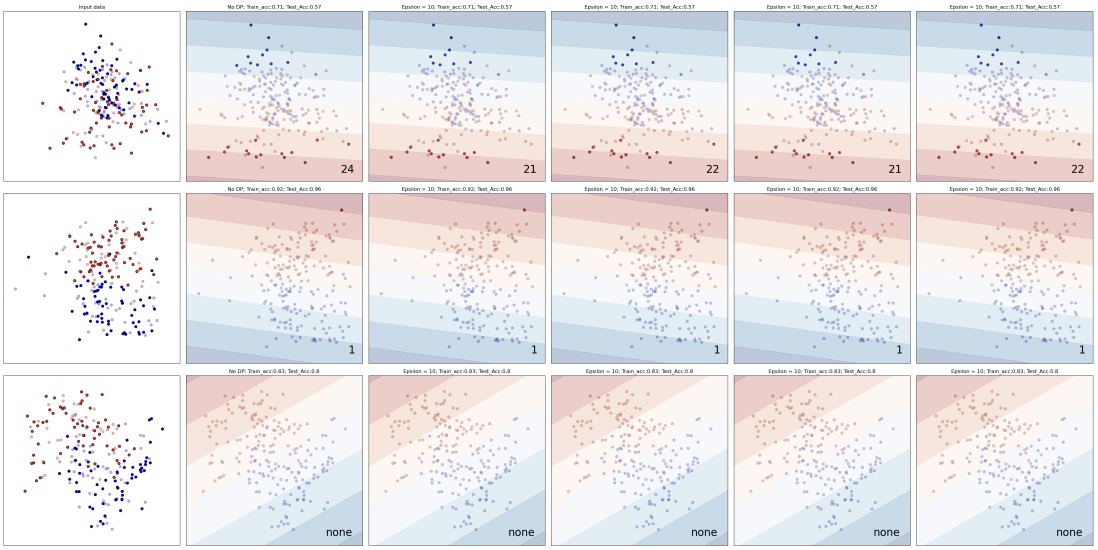

Figure A.9: Randomness in DP protection at epsilon 100 on the MIA on a model trained on the synthetic dataset: this plot is used for the same purpose as Figure A.8 but with protection at a weaker level. Compare to Figure A.8, we can see that DP at epsilon 100 protect points is less effective in terms of the number of points covered and the frequency of successful protection.