Information Leakage in Machine Learning Models

By

Shakila Mahjabin Tonni

A thesis submitted to Macquarie University
for the degree of Master of Research
Department of Computing
February 2020

MACQUARIE
University

Typeset in LaTeX 2_ε.

Declaration

I certify that the work in this thesis entitled INFORMATION LEAKAGE IN MACHINE LEARNING MODELS has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree to any other university or institution other than Macquarie University. I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Shakila Mahjabin Tonni

Acknowledgements

This thesis became a reality with the kind assistance of many individuals. First of all, I would like to convey my sincerest gratitude to my supervisor, Prof Dali Kaafar, whose expertise and wisdom was invaluable in the formulating of the research topic and methodology in particular. Throughout the writing of this thesis, I have received enormous guidance and encouragement from him.

Besides my supervisor, I would also like to thank Dr Farhad Farokhi and Dr Dinusha Vatsalan, Research Scientist, Information Security and Privacy Group, Data61 for their patience, motivation, advice, and immense knowledge. Their continuous efforts guided me in the right direction to complete this thesis.

I am grateful for the inspiration and empathy that I received from my colleagues. Also, I am thankful to the Department of Computing for their relentless commitment to providing commendable research and development infrastructure.

Finally, it won't be possible for me to be focused and work with dedication without the unconditional support of my family.

Abstract

Machine Learning (ML) techniques are used by most data-driven organisations to extract insights. In addition, Machine-learning-as-a-service (MLaaS), where models are trained on potentially sensitive user data and then queried by external parties are becoming a reality. However, recently, these systems have been shown to be vulnerable to Membership Inference Attacks (MIA), where a target's data can be inferred to belong or not to the training data. While the key factors for the success of MIA have not been fully understood, existing defences mechanisms only consider the model-specific properties. In this thesis, we investigate the impact of both the data and ML model properties on the vulnerability of ML techniques to MIA. Our analysis indicates a strong relationship between the MIA success with the properties of the data in use, such as the data size and balance between the classes as well as with the model properties including the fairness in prediction and the mutual information between the data and the model's parameters. We provide recommendations on assessing the possible information leakage from a given dataset and propose new approaches to protect ML models from MIA by using several properties, e.g. the model's fairness and mutual information between data and the model's parameters as regularizers, which reduces the attack accuracy by 25% yielding a fairer and a better performing ML model.

Contents

List of Figures

List of Tables

1 Introduction

Advances in Machine learning (ML) techniques are nowadays enabling highly accurate predictions, analytics extraction, classification and recommendation tasks for a wide range of applications. However, the success of these models is largely dependent on the access to well-provisioned computationally powerful platforms as well as the availability of a substantial amount of data used for model training. Third party Machine-Learning-As-A-Service (MLaaS) providers like Google and Amazon have successfully addressed these two challenges by providing publicly accessible high-performance computing infrastructure combined with ML algorithms trained on enormous data. Working as a black-box API, MLaaS enables users to upload their own dataset to the server and get a trained ML model on it. Organisations, public and private alike, as well as data scientists and researchers, are now using the MLaaS platforms to get insights on data collected from a vast range of sources.

The availability of such services raises certain safety and privacy concerns, as for many domains, the use of sensitive data in learning is inevitable. For example, social media researchers are utilizing machine learning in analyzing human behaviours through massive social media data [25]. Similarly, medical records are analyzed by vendors who provide healthcare or insurances to verify the likelihood of certain health conditions [33]. In both the cases of social media and health, data are personal and sensitive, and therefore involves data privacy issues. In certain scenarios, biomedical data [3] and location data [32] are also deemed to be sensitive in nature. Although the structure of the learning model is generally hidden, as a user trains the model on provider's server, the centralized server presumably has access to the records it was trained on, which could potentially be misused if the presence of a record in the training data is exposed. In addition to that, some services give flexibility to the data owner to extend a pay-per-query model, where other users can query the owner's learning model [41]. Because of the publicly available querying platform, adversarial attacks are feasible by probing the outcome of the model to gain knowledge of the model structure or about some user's record [37, 43, 20, 45, 25, 17]. Especially, models deployed as deep learning (DL) neural networks, are more prone to different kinds of adversarial attacks [35, 31].

Membership Inference Attack (MIA) [37] is one of the highly regarded critical inference attacks against ML models that can reveal whether a record was used to train a model. This attack can achieve a substantial amount of successes on ML models even when the model structure is not

shared. Successful MIA can pose a severe threat to user privacy by identifying the data of a particular user contained in a dataset. For instance, knowing if a person's record was part of the data used to analyse suicidal behaviours [25] or movements of Alzheimer patients [32] reveals information about the person's suicidal tendency and health condition.
Hence, it is crucial to understand the reasons behind a successful MIA on a model and to get an idea of the possible information leakage through the attack. Although, many pieces of research [34, 30, 14] have established a link between MIA and certain ML model properties (e.g. overfitting, choice of hyperparameters), studying the effect of other data and model properties is yet to be done.
In this research, we investigate information leakage from ML models imposed by MIA. Our primary objective is to identify the reasons behind the model's susceptibility towards the attack and possible measures of reducing the severity of MIA. To build a comprehensive understanding, a rigorous analysis is performed in this research, where we investigate multiple properties related to both data and model. Explored data properties are the data size, balance in the classes and features, the number of features and entropy. Explored model properties include the selection of classifiers, hyperparameters, mutual dependency between the record and model parameters, overfitting and model's fairness. We also study MIA-indistinguishability as proposed in [47] that captures the vulnerability of a model to MIA in the term of disparity in the model's predictions towards member and non-member records. We estimate MIA's performance against variations in each of the properties and found some of the properties may accelerate the effectiveness of the attack when scaled up (e.g. data size), while others serve the opposite (e.g. model's fairness). All the experiments are performed on Artificial Neural Network (ANN) except the experiment where we have tested the effect of different model combinations. For the later experiment, we use other ML algorithms such as Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), K- Nearest Neighbour (KNN) in addition to ANN.
We further investigate approaches to enhance the model's resistance against MIA in compliance with the findings and observe that the use of certain model properties as regularizers in the ML model may reduce the attack accuracy considerably. We demonstrate the effectiveness of our proposed defence mechanism by using the model's fairness differences and mutual information between its records and model-parameters as regularizers. The used regularizers reduce attack accuracy and improve the model's performance significantly.

1.1 Background

1.1.1 Machine Learning Preliminaries

ML models outperform other Statistical Models (SM) (e.g., Bayesian regression and generalized additive models) with their extreme adaptability towards evolving and complex data features [5, 21]. In case of the supervised learning, an ML model is trained on a set of data points to capture their inherent features and map these features to a set of predefined output labels. The aim of the training phase is that, once trained, the model is capable of predicting label for a new unlabeled data point. Let's assume, $D = \{(x_i, y_i)\}$ is the set of m data points sampled from a probability distribution $P(X, y)$ of feature vectors $x_i \in X$, where X is the feature space, and class label $y_i \in y$, where y is a predefined set of class labels. An ML algorithm attempts to identify a function $f : X \rightarrow y$ that maps the input data points to different classes in the best possible way. The output is a probability vector $P(f)$ that indicates the relative association of a

data point to each of the class labels in y. Let's define the model f as:

$$y = f(X; \theta),$$

where θ represents the parameters generated by the model based on the features. A loss function $l(f(x_i; \theta), y_i)$ captures the error of the prediction by measuring the difference between actual and the model's predicted class labels. Hence, the algorithm's target is to find a function f that minimizes the below expected loss:

$$L(f) = \mathbb{E}_{(x_i, y_i) \sim P(X,y)}[l(f(x_i; \theta), y_i)]$$

The empirical loss of the model f over the training dataset D can be defined as:

$$L(f) = \frac{1}{m} \sum_{i}^{m} l(f(x_i; \theta), y_i),$$

where m is the number of data points in D. However, a model that captures the exact feature-to-label mapping, is more likely to produce erroneous prediction while encountering an unknown data point. In order to prevent the algorithm from excessive leaning towards a particular dataset D, known as overfitting, and to achieve better generalization to all the data points sampled from similar distributions, different regularizers are used in practice. A regularizer penalizes the model parameters if the model becomes complex i.e., learn too much information from the training data. Therefore, the optimization problem of the model with parameters θ is to minimize the below empirical loss:

$$\min_{\theta} \frac{1}{m} \sum_{i}^{m} l(f(x_i; \theta), y_i) + \lambda R(\theta),$$

where $R(\theta)$ is the regularizing function with a weight balancing $\lambda > 0$, also called L2-ratio. The regularizers can be the L_p norm of θ, i.e., $\|\theta\|_p^p$. For example, L_1-norm regularizer is $R(\theta) = \|\theta\|_1^1$ and L_2-norm regularizer is $R(\theta) = \|\theta\|_2^2$. The algorithm f repeatedly updates the model parameters θ to achieve the lowest possible cost function $L(f)$ by updating the decision variables in the direction of the gradients:

$$\theta^+ := \theta - \alpha \frac{\partial L}{\partial \theta},$$

where θ^+ is the updated model parameters and α is a configurable hyperparameter called the learning rate that determines how much the parameters will be updated during each epoch. Smaller learning rates require more training epochs as they make smaller changes on each update, whereas larger learning rates result in rapid changes and require fewer training epochs. However, a very high learning rate may result in a fast calculation of non-optimal parameters causing an inconsistent training.

1.1.2 Membership inference attack (MIA)

MIA [37] or tracing [8, 10], determines a record's presence in the training dataset of an ML model without knowing the structure of the model. The attack model is based on several assumptions. Firstly, the attacker has a black-box oracle access to the model and can acquire the

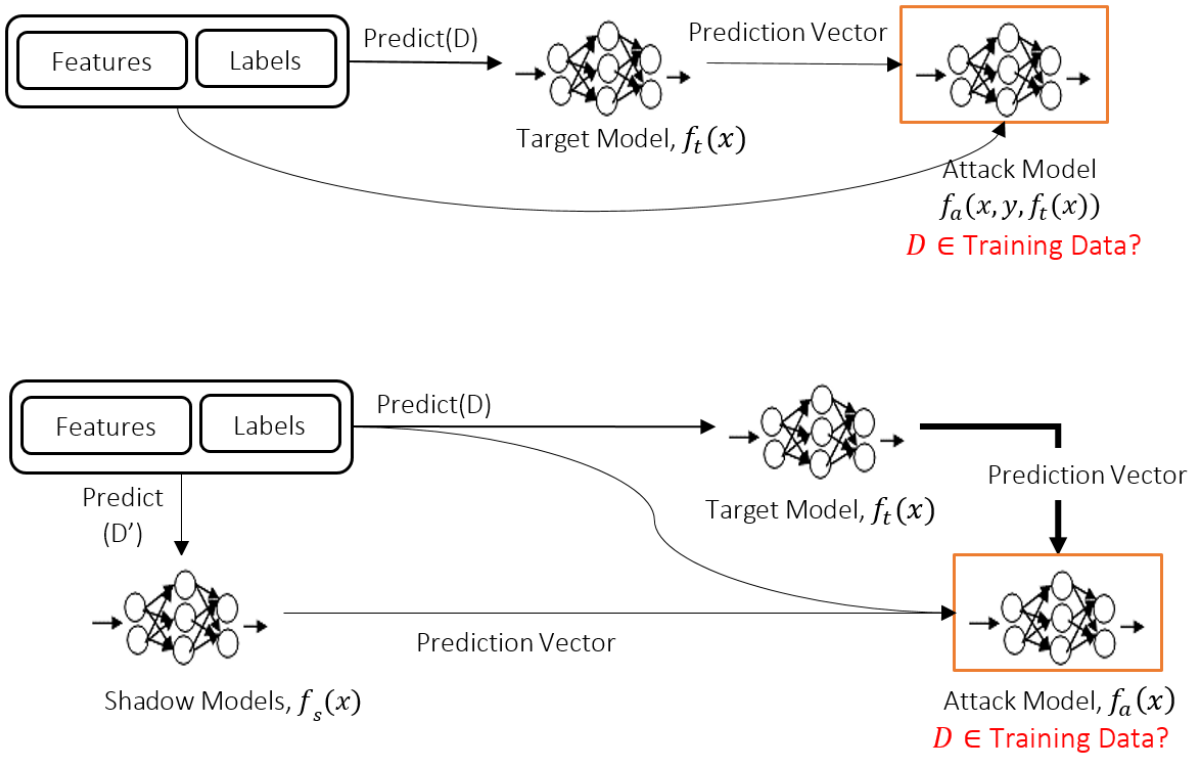

FIGURE 1.1: Membership Inference Attack (MIA) showing how the attacker acquires the prediction vector by querying the target model. Obtained prediction vector and the correctly labeled records are passed to the attack model, which determines the membership of the records. The second diagram includes the use of shadow models to train the attack model beforehand.

model's prediction vector on any data record. Secondly, the data distribution of the target model's inputs and outputs, including their number and the range of values they can take is known to the adversary. The adversary intends to distinguish training set members from non-members by observing the model's predictions. To identify a record's presence (i.e the membership), the attacker tries to generate output close to the target model on randomly sampled records from the similar data distribution and trains an ML model to classify a record's membership based on the prediction probabilities generated by the model. The attacker then uses the trained classifier model on the observed output of the target model to infer the membership information of some records.

The classic MIA [37], consists of training three different models: 1) target model, 2) shadow model and 3) attack model. The target model is the model of interest for the adversary to learn sensitive information about the individuals. The structure of this model and the used training dataset are essentially kept hidden. An adversary has access to the target model to perform queries on it and obtain some aggregated statistics on the data. The purpose of a shadow model is to imitate the target model's behaviour and generate outputs similar to it. As the target model's structure is not known, the adversary implements multiple shadow models by sampling data from similar distributions as the target model's training data. Finally, the classifier that categorizes records into member and non-member classes (i.e used in the target model's training data or not), is called the attack model. The attack model is trained on the prediction vector obtained from the shadow models and tested against the prediction vector of the target model (truth data/ground truth). An overview of the attack is illustrated in Figure 1.1. The attack model

f_a aims at increasing the number of inferred members. Hence, defences against MIA can be achieved by building a model robust against the attack model.

1.2 Literature Review

Adversarial attacks on ML models can be ranged from perturbing the model's prediction using adversarial examples to inferring knowledge about the model itself or the data used in it. Based on adversary's knowledge the adversarial attacks could roughly be categorized into two groups: white-box and black-box attacks [49, 36]. Adversarial attacks that are performed on a known model structure, are called white-box attacks. In contrast, if the structure of the target model is unknown to the attacker, they are called black-box attacks. The adversarial attack first proposed by [13] in a white-box setting, shows that introducing only a little amount of adversarial noise can successfully mislead the model with a higher confidence levels of predictions. Later, the attack was implemented on multiple ML models [29, 23, 19]. Recently, the explosion of Generative Adversarial Network (GAN) [7, 42, 27] exploits this idea remarkably by generating adversarial examples. In GAN-based adversarial attack prevention, adversarial examples are used to train a model to discriminate between the actual and model-generated inputs. A comprehensive study on the white-box adversarial attacks and their defenses is conducted in [24].
However, an adversarial attack that can gain insight about the records or the model without knowing the model structure (black-box attack), is allegedly more devastating [8, 41]. Membership Inference Attack (MIA) [37, 43] is one such attack, where although the model's structure is not disclosed to the adversary, the adversary successfully manages to learn whether a record is part of the private training data. MIA can be highly successful given the adversary has prior knowledge of the data distribution of the training dataset and can observe the model's outcomes in determining a record's presence in the training data (as in the MLaaS platform).
Since its inception, MIA has shown a tremendous success on other models such as GAN [14, 7, 16] and differentially private models [33] in a range of domains such as social media [25], health [3], sequence-to-sequence video captioning [17] and user mobility [32]. To attack a model by simulating its behaviour, attacker deploys multiple shadow models [37]. However, later, Salem et.al. [34] showed that an attack model that uses only one such shadow model or no model at all can still render strong membership inference.
Besides MIA, other black-box attacks include adversarial attempts inferring sensitive attributes of individual records through Attribute Inference Attack [20], and Model Inversion Attack [45, 11, 15], that allows an adversary to approximate the training data with varying accuracy level based on the confidence value generated by the model on them without any knowledge of the model.
A successful MIA exploits a model's tendency to yield higher confidence value when encountering data that they are trained on (members) than the others. The property of a model to overfit towards its training data makes it vulnerable to MIA [48]. Thus, existing defences against MIA attempt to reduce overfitting of a model by applying regularizers like L2-regularizer [37] that generalizes a model's prediction or by adding Dropout layers to the model [39] that ignore a few neurons in each iteration of training to avoid high train accuracy. Nasr et al. [30] proposed a min-max game-theoretic defence method using the highest possible attack accuracy as adversarial regularization to decrease the target model's prediction loss while increasing privacy against MIA. Differential privacy [28] is also another proposed defence method against membership inference. However, the security guarantee is limited to a certain value of the privacy budget ϵ [33, 22]. Furthermore, recent works reveal that overfitting can be necessary but not sufficient

condition for a successful MIA. [26] shows that the effect of MIA was strong even when the models were well-generalized. Therefore, it is necessary to identify other reasons for a model's vulnerability to MIA.
MIA-indistinguishability, as proposed in [47], defines a model's vulnerability based on the equality of prediction or fairness of the model on the member and non-member records. Fairness in ML [46, 12, 4] is an emerging concept, that specifies how much a model distorts from producing predictions with equal probabilities for individuals across different protected groups. Similarly, MIA-indistinguishability defines a model's vulnerability towards MIA by estimating the model's discrimination in prediction for the member and non-member records. Intuitively, besides overfitting, both MIA-indistinguishability and model's fairness in general can be possible reasons behind MIA's success.

1.3 Problem Statement

Several works have been done to improve machine learning models' resilience against adversarial MIA. Most of these solutions consider reducing the overfitting of the model to make it resistant against such adversarial attacks. However, in addition to model overfitting, there are several underlying data and model characteristics that could contribute to the success of these attacks. For example, data characteristics such as data size, balance in the classes and features and entropy, as well as model characteristics such as overfitting, group and individual fairness and MIA-indistinguishability might play an important role in determining the success rate of MIA. No comprehensive study has been conducted so far in the literature to investigate which of these factors significantly impact the attack accuracy, so that the resistance methods can be developed considering those factors. In this thesis, we aim to conduct an exploratory analysis of different data and model properties' influence in the success of MIA and provide recommendations to develop defence methods in order to improve models' resistance based on the findings of our study.

1.4 Contributions

This research aims at protecting user data privacy against MIA by establishing a relation between ML models and the attack on a black-box settings and by identifying the reasons behind information leakage from the models. To achieve this, we measure MIA effectiveness for multiple data and model properties. Explored data properties are: data sizes, balance in the classes and features, number of the features and entropy. Explored model properties are: choice of classifier and different hyperparameters, overfitting, ML fairness and MIA-indistinguishability. From the study, we find MIA to be highly affected by both the data and model properties. In this thesis, MIA is implemented according to [37]. In most of the experiments, target, shadow and attack models are implemented as Artificial Neural Network (ANN) except in one experiment where we also investigate other ML models such as Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest(RF). Furthermore, this research illustrates ways to improve robustness of the model against MIA by using the model properties as regularizers in the model. We verify our proposition by investigating models with mutual information between records and parameters and three different fairnesses as regularizers against MIA. In summary, our contributions are:

- **Identifying the correlation of different data and model properties with MIA's success and their potential impact:** Some of the properties show strong positive correlation

with MIA (for example, fairness of the model), while others show negative correlations (for example, balance in the classes). However, for a few of the properties, such as, number of the features and entropy, we could not find any straightforward correlations with MIA, which we intend to explore further in future.;

- **Minimizing information leakage in ML models by reducing MIA accuracy based on these findings:** we propose to use influential model properties such as model's fairness and mutual information between the records and the model parameters as regularizers in the model for improved defense against MIA;
- **Studying the effectiveness of the recommended defense methods:** We demonstrate the models implemented with the above mentioned custom regularizers, reduce MIA accuracy by a higher rate and increase the model's performance compared to the models without any regularizer and with the L1 or L2 regularizer.

1.5 Outline of the Report

This thesis is structured as follows. Chapter 1 provides a brief introduction including background, literature review and contributions. The details of the methodology used in this research are described in Chapter 2. Chapter 3 illustrates the experimental set up and Chapter 4 presents the results obtained from the experiments. Based on the obtained results, our proposed defence methods to build MIA-resilient model are presented in Chapter 5. Chapter 6 provides conclusion and remarks on future works.

2

Research Methodology

This chapter provides an outline of the research methodology followed in this thesis.In addition, we also describe the selected data and model properties in this chapter. We use the classic MIA structure described in [37]. We study the effect of the data size, class and feature balances, number of features in the dataset, and entropy on the success of MIA. Furthermore, we also explore the impact of the model's hyperparameter selection, fairness, and overfitting on MIA. Besides, we capture the relationship between the amount of information about the data available in the ML model and the success of MIA using mutual information. Also, we study multiple ML algorithms arranged in several target and shadow combinations to determine the MIA-resilience of a target model against different classifiers.

Algorithm 1 Membership Inference Attack (MIA) (Shokri'17 [37])

$D(X, y)$ is the total population. $m \in \{0, 1\}$ is the membership label.

1: Sample, $D_t, D_s \leftarrow D$, where $D_t \neq D_s$
2: **On target model f_t:**
3: Sample, $D_{t-train}, D_{t-test} \leftarrow D_t$, where $D_{t-train} \neq D_{t-test}$
4: Get, $P(f_t) \impliedby P[\hat{y} = y|(X, y) \in D_{t-train}]$
5: **On shadow model f_s:**
6: Sample, $D_{s-train}, D_{s-test} \leftarrow D_s$, where $D_{s-train} \neq D_{s-test}$
7: Get, $P(f_s) \impliedby P[\hat{y} = y|(X, y) \in D_{s-train}]$
8: Assign membership label m to $(X, y) \in D_s$
9: **On attack model** $f_a(x, y, P(f)$:
10: Train on $(X, y) \in D_s$ and $P(f_s)$
11: Test on $(X, y) \in D_t$ and $P(f_t)$

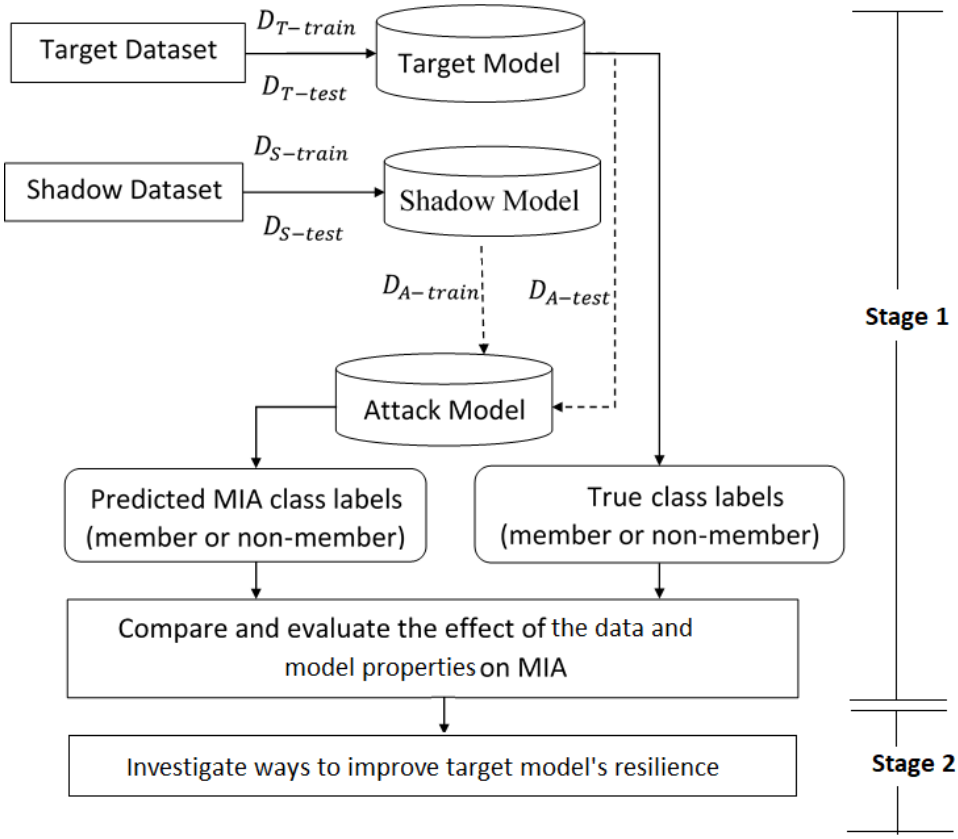

FIGURE 2.1: An overview of the proposed 2-stage methodology. In the first stage, customized datasets are used as an input to the MIA model and the effects of different data and model properties are observed comparing the predicted membership and true membership of the records. In the second stage, based on the observations, further exploration is done to improve the model's resilience against MIA.

2.1 Exploratory Analysis of Different Properties on MIA

The overall research is based on an exploratory research method that consists of two major stages. In the first stage, MIA is implemented on customized datasets and multiple models to systematically evaluate MIA attack accuracy, attack precision and attack recall against different data and model properties. Details of the explored data and model properties are discussed in Sections 2.2 and 2.3 and Algorithm 1 outlines the steps to implement MIA. Based on the results obtained from the first stage, we study the correlation between different data and model properties with the success of the attack. In the second stage, we study the effectiveness of using multiple model-based properties to improve the model's resilience against MIA by applying them as regularizers. We evaluate the model's performance in terms of both producing the correct prediction and resistance against information leakage through MIA. We also compare our results with model without any regularizer and with the two standard $L1$ and $L2$-norm regularizers. Figure 2.1 illustrates an overview of the methodology.

2.2 Explored Data Properties

In an ML setting, a dataset $D(X, y)$ can be considered as a collection of records or data points (X, y) with a feature vector X and a set of class labels y. Each feature $x_i \in X$ contains different level of balances in the feature values and in the entropy. Also, the data points of each dataset

may have different balances between the class labels. Our first objective is to identify whether properties of a dataset have any impact on MIA. A brief description of the investigated properties and their evaluation metrics are given below:

1. Data sizes: Number of the records sampled from the population for training the target model has an effect on the model's prediction. More records often enhance the prediction of an ML model. On the other hand, availability of many records may influence the vulnerability of the model to MIA.
2. Balance in the classes: Balance in the classes can be measured as the frequency ratio between different classes in the dataset. For instance, a dataset $D(X, y)$ would have a proper balance between the binary class labels $y \in [0, 1]$ if:

$$P[y = 0|\forall x_i, x_i \in X] = P[y = 1|\forall x_i, x_i \in X] \tag{2.1}$$

   In case of the multiple classes, we consider the ratio between one selected class against all the other classes. Also, for simplicity, throughout the thesis, the class balances are denoted as the percentage of one of the class labels. For example, 10% class balance refers to the dataset having 10% records labeled as the class $y = 1$.
   Balance among the class labels has a tremendous impact on model's prediction and fairness [12]. High imbalance in the classes makes the model producing biased prediction towards the class that has more occurrences. Thus, this property can be considered as one of the factors impacting MIA.
3. Balance in the features: Similar to the balance in the classes, balance in the features can have a huge impact on model's prediction. Let's assume, X is the set of features and A is the set of records, we denote $a_j.x_i$ for the feature $x_i \in X$ of the record $a_j \in A$. The set of possible feature values for a single feature $x_i \in X$ is $C = \forall_{a_j \in A}\{a_j.x_i\}$.
   In the case of multiple features, we concatenate all the feature values of a record into one and consider as a single feature. The set of all possible combinations of feature values for all features X is: $C = \forall_{a_j \in A}\{\forall_{x_i \in X} a_j.x_i\}$. The feature balance is calculated as the ratio between one feature value c in C and all the other feature values in C. So, a dataset with balanced feature would have following equality:

$$P[a_j.x_i = c|\forall_{a_j \in A}\forall_{x_i \in X}, c \in C] = P[a_j.x_i \neq c|\forall_{a_j \in A}\forall_{x_i \in X}, c \in C] \tag{2.2}$$

   For example, a properly balanced single feature x_i with only two feature values $\{0, 1\}$, has an equal probabilities of the feature values:

$$P[a_j.x_i = 0|\forall_{a_j \in A}] = P[a_j.x_i = 1|\forall_{a_j \in A}] \tag{2.3}$$

4. Entropy: Entropy is a measure of randomness or uncertainty of a variable. A dataset containing more random features effectively distorts the prediction [38]. Thus, this property has the potentiality to affect MIA. Following Shannon's entropy formula, if $a_j.x_i$ are the feature value of the feature $x_i \in X$ and of the record $a_j \in A$, the entropy of each feature $x_i \in X$ is:

$$H[P(x_i)] = -\sum_{a_j.x_i \in x_i} P_{a_j.x_i} \log P_{a_j.x_i}, \tag{2.4}$$

   Thus the entropy of overall dataset can be calculated by taking the mean entropy over n number of features:

$$H[P(X)] = \frac{1}{n}\sum_{x_i \in X} H[P(x_i)] \tag{2.5}$$

5. Number of features: Besides the above-mentioned features, we also observe the effect of the number of features in a dataset on MIA.

2.3 Explored Model Properties

We consider below model properties as possible candidates to have an impact on MIA:

1. Choice of model's hyperparameters: In case of Neural Network (NN) based ML models, performance of the model is highly case-specific for distinct prediction problems. It requires exploring enormous possibilities to design a highly specific model. The structure of the model and the choice of hyperparameter affect the prediction [40], which may also affect MIA's success. Relying on the best practice recommendation we select below hyperparameters and observe their impact on MIA:
   - Number of layers in the model: Number of layers in an NN increases the learnability of the model. However, too many layers may also cause the model to overfit by allowing the model to learn excessive information about the training data, which is one of the established reasons behind a successful membership inference [48]. In contrast, fewer number of layers may limit the model's performance, yielding underfitting.
   - Number of nodes per layer: The nodes in each layer hold the network generated weights or parameters based on the input features and information for the transformation that a layer performs. For the first layer, this is driven by the number of features. In subsequent layers, the number of units depend on the choice of expanding or contracting the representation from the previous layer. The result of the entire NN solely relies on the computation of the outputs done by the nodes at each layer. Hence, varying the number of nodes supposedly affect MIA as well.
   - Rate of regularization: Regularizers are used to generalize a model. Regularization rate (λ) controls the amount of penalty that would be applied on the model's loss if the model becomes complicated. Use of regularizers is one of the effective defences against MIA [37]. Hence, MIA's effectiveness would also be influenced by the variations in the rate of regularizations.
   - Learning rate: The rate of changes in the network parameters between iterations/epochs is defined by the learning rate (α) of the model. Large learning rate causes bigger fluctuations in the parameters that result in non-optimal prediction. On the other hand, lower learning rate takes more iterations to converge. Either way, learning rates affect MIA by altering the model's learning capacity.
2. Target-shadow model combination: As the target model structure is unknown to the attacker, the attacker applies different shadow models against the target model to infer membership information. Hence, evaluating a target model against only one type of ML classifier may not reveal the overall scenario. It is also necessary to compare different shadow models in combinations to identify the maximum amount of information leakage from the model.
3. Mutual information of the data and model parameters: The focus of an ML algorithm is to identify a function $f(X;\theta)$ that maps the input data points to different classes. The model's parameters θ are generated based on all the features of the training set. The mutual dependence between the model parameters and the features holds the amount of information obtained by the model after observing the features and can be represented as the mutual information between them $I(X;\theta)$ which can be calculated as:

$$I(X;\theta) = H(X) - H(X|\theta), \tag{2.6}$$

where $H(X)$ is the marginal entropy of the features and $H(X|\theta)$ is the conditional entropy that quantifies the amount of information needed to explain X when the value of θ are known.

4. Fairness of the model: Within the past few years, fairness has become one of the most popular topics in ML. Many definitions of ML fairnesses have been proposed in the literature [12, 46, 6]. Fairness is one of the critical properties of a model that measures a model's behavior of prediction towards different individuals grouped based on a particular protected feature. Fairness of a model can be defined in many ways [6, 4]. In order to evaluate a model's fairness against the severity of MIA, we consider three types of fairness measures: group or statistical fairness, predictive fairness and individual fairness. Let's assume, 'Gender' is a protected feature in a dataset. It contains two groups of individuals: 'male' and 'female'. Fairness of an ML model on this dataset would essentially mean whether the model treats both 'male' and 'female' records equally without giving benefit to one group more than the other. Different fairnesses can be defined as below:

   - Group Fairness: A model is considered fair, if it predicts a particular outcome for individuals across the protected subgroups with almost equal probabilities [12]. A predictor $f(X) \implies y$ achieves group fairness with respect to the two groups of records $g_i, g_j \in X$ iff,

   $$P[\hat{y}_i = y | x_i \in g_i] = P[\hat{y}_j = y | x_j \in g_j], \tag{2.7}$$

   where $\hat{y}_i$ and $\hat{y}_j$ are the predicted outcomes for the records in group g_i and g_j. For instance, in our previous example of the 'Gender' feature, the model is said to have perfect group fairness if the number of the predicted class (for instance, class 1) is the same for both 'male' and 'female' records.
   To determine the group fairness of a model, we estimate the fairness difference, δ_g, between two subgroups of the records as below:

   $$\delta_g(f) := \frac{1}{2} \sum_{\hat{y}_i, \hat{y}_j \in y} |P(\hat{y}_i) - P(\hat{y}_j)| \tag{2.8}$$

   - Predictive Fairness: A classifier satisfies this definition if the subgroups have equal probability to truly belong to the positive class, in other words, has similar rate of precisions [46]. That is,

   $$P[\hat{y}_i = y | y_i = y, x_i \in g_i] = P[\hat{y}_j = y | y_j = y, x_j \in g_j] \tag{2.9}$$

   Based on Equation 2.9, we measure the difference in predictive fairnesses δ_p of a model as below:

   $$\delta_p(f) := \frac{1}{2} \sum_{\substack{\hat{y}_i, \hat{y}_j \in y \\ y}} |P(\hat{y}_i) - P(\hat{y}_j)| \tag{2.10}$$

   - Individual Fairness: The concept of individual fairness was introduced in [9] that ascertains a predictor is fair if it produces similar outputs for similar individuals. That is, if two records, x_i and x_j are similar, then the prediction on them would be similar by the model [12]. However, equal distance between two records and between their predictions are difficult to achieve. According to [9], a realistic definition of individual fairness would be:

   $$\Delta(\hat{y}_{x_i}, \hat{y}_{x_j}) \leq d(x_i, x_j), \tag{2.11}$$

where, d is the distance between the records x_i and x_j and Δ is the distance between the model's prediction for x_i and x_j denoted as $\hat{y}_{x_i}$ and $\hat{y}_{x_j}$, respectively.
Both Δ and d can be measured in different ways. In [9], a statistical distance metric is proposed for Δ that measures the total variation norm between two probabilities for the outcome of the classifier for the two records:

$$\Delta_{tv}(\hat{y}_{x_i}, \hat{y}_{x_j}) = \frac{1}{2} \sum_{\hat{y}_{x_i}, \hat{y}_{x_j} \in y} |P(\hat{y}_{x_i}) - P(\hat{y}_{x_j})| \tag{2.12}$$

This metric assumes that the distance metric selected for d will scale the measured distance within 0 to 1 range, where distance between two similar records would be nearly zero and distance between two dissimilar records would be close to 1. They also suggest a better choice for Δ using the relative l_∞ norm metric:

$$\Delta_{\infty}(\hat{y}_{x_i}, \hat{y}_{x_j}) = \sup_{\hat{y}_{x_i}, \hat{y}_{x_j} \in y} \log(\max\{\frac{P(\hat{y}_{x_i})}{P(\hat{y}_{x_j})}, \frac{P(\hat{y}_{x_j})}{P(\hat{y}_{x_i})}\}), \tag{2.13}$$

which would allow the use of a metric for d that considers two records to be similar if $d << 1$ and dissimilar if $d >> 1$. The authors discuss a few futuristic insights in identifying a proper metric for d considering scenarios such as distance between two records from the same protected group and different protected groups. They also propose to build a user-specific distance metric for calculating d by allowing users to specify the set of attributes to consider in the calculation (refer to [9] for details).
Beside the above measures, according to [46], a possible distance metric for d is the distance between two records x_i and x_j divided based on one certain feature which would be 0 if all the other feature values are identical and 1 if some are different, i.e, $d = x_i \oplus x_j$. Similarly, the authors of [46] proposed to simply define Δ as 0 if the classifier resulted in the same prediction and 1 otherwise.
In our experiments, to measure the individual fairness δ_i, we consider the differences between two different distance measures d and Δ derived from Equation 2.11:

$$\delta_i(f) := \Delta(\hat{y}_{x_i}, \hat{y}_{x_j}) - d(x_i, x_j) \tag{2.14}$$

For simplicity of the calculation, we use the statistical distances between the predictions (Equation 2.12) and between the records to compute Δ and d respectively as below:

$$\delta_i(f) := |P(x_i) - P(x_j)| - \frac{1}{2} \sum_{\hat{y}_{x_i}, \hat{y}_{x_j} \in y} |P(\hat{y}_{x_i}) - P(\hat{y}_{x_j})| \tag{2.15}$$

A lower value of the estimated fairness differences represents a model producing fairer outputs. For instance, we can assume a model is fair on the record subgroups, if δ_g is very small (≈ 0) and unfair if large (≈ 1). Also, throughout the thesis, the terms fairness and difference in the fairnesses are used synonymously.

5. MIA-Indistingushability: As proposed by [47], MIA-indistinguishablity represents a model's strength to withstand the MIA. Intuitively, a model can be said MIA-indistinguishable if the probability of a record's presence in the training dataset of the model is the same as the probability of its presence in the test dataset. As an intruder has access to the outcome of the model, according to [47], the target model f_t satisfies perfect MIA-indistinguishability if for any prediction over the class labels $y \in \{0, 1\}$:

$$P[m = 1|\hat{y} = y, y] = P[m = 0|\hat{y} = y, y], \tag{2.16}$$

where $m \in \{0, 1\}$ is the membership value denoting whether the record is a member of the train data ($m = 1$) or test data ($m = 0$). Another equivalent form of this formula assuming the adversary's zero prior knowledge on a record's membership and unbiased classes sampled in the dataset, f_t satisfies the perfect MIA-indistinguishability iff for any $\hat{y} \in y$:

$$P[\hat{y} = y|y, m = 1] = P[\hat{y} = y|y, m = 0] \tag{2.17}$$

In the worst case scenario, how much a model deviates from being indistinguishable can be measured using the l_∞-relative metric between the considered probabilities and maximum divergence across the classes of y:

$$\delta_{mi}(f) := D(P(\hat{y}_{m1}), P(\hat{y}_{m0})) = \sup_{\substack{\hat{y} \in y \\ y}} |\log \frac{P(\hat{y}_{m1})}{P(\hat{y}_{m0})}|, \tag{2.18}$$

where $\hat{y}_{m1}, \hat{y}_{m0} \in y$ are the prediction for the records with the membership value $m = 1$ and $m = 0$, respectively. Thus, a model is presumably less vulnerable to MIA if δ_{mi} is close to zero. We estimate the MIA-indistinguishability by computing the statistical difference between prediction of the train $(m = 1)$ and test data $(m = 0)$ of the target model using the below equation equivalent to Equation 2.18:

$$\delta_{mi}(f) := \sup_{\substack{\hat{y} \in y \\ y}} |\log\{\max\left(\frac{P(\hat{y}_{m1})}{P(\hat{y}_{m0})}, \frac{P(\hat{y}_{m0})}{P(\hat{y}_{m1})}\right)\}| \tag{2.19}$$

We evaluate MIA-indistinguishability for different member and non-member ratios. To simplify, we denote the ratios as the member rates, representing the percentage of member records sampled from the target dataset. For instance, member rate 10% refers to 10% train records and 90% test records split randomly from the target dataset. Similarly, 50% member rate represents equal proportion of the member and non-member records.

6. Model Overfitting: As indicated in [30, 37], overfitting is one of the main contributors to the machine's vulnerability towards MIA. ML models tend to predict correctly with higher probability when the record is a member of the training set. This tendency to lean towards the training data, makes the model vulnerable to MIA, as the adversary can make assumption about a record's membership based on the prediction outcome. However, [26] shows that, membership inference is significant even though the models are well generalized. To explore this property, according to the previous works [37, 34, 26], we measure overfitting as the difference between the training and testing accuracy of the target model.

3 Experimental Setup

The investigations conducted in this work are two-folded: we apply and evaluate MIA in multiple experimental setup by varying the properties as described in Section 2.2 and 2.3. In this chapter we explain how the data are customized along with the model setup for different experiments.

3.1 Property Measures

In our experiments we estimate different properties (listed in Section 2.2 and 2.3) as below:

- We measure the balance in the classes by calculating the ratio of the two (binary) classes $(0, 1)$ available in the dataset, where 0% being the total absence of the class label 1 and 50% being the equal frequency of both the class labels 0 and 1. For simplicity, the balances are denoted based on the proportion of the records with class label of 1. For example, when sampling records from the Purchase dataset with the ratio of 1's and 0's is $1:9$ as the class labels, we simply refer the dataset to have a 10% balance in the classes.

- We select one value from the range of each feature and measure the feature balance as the ratio between the selected feature value and all the other feature values. Similar to the class balance, we also denote the feature balance as the percentage of the chosen feature value for each feature. For example, for the feature 'Repeater' that has the values $\{0, 1\}$, a 10% balance in its feature means 10% of the records have 'Repeater'$= 1$ as the feature value. Throughout the thesis, the terms balanced classes and balanced features refer to the datasets having an equal proportion of the classes and feature values.

- The entropy of a dataset is measured by taking the mean entropy calculated using Equation 2.5 over all the features.

- The mutual information between the features and the model-generated parameters $I(X;\theta)$ is calculated according to Equation 2.6 by taking the mean value over θ produced in multiple layers for each feature.

- Model's fairnesses are measured as the difference of fairness values according to Equation 2.8,2.10 and 2.15 between one selected group of records and all the other records.

- We measure MIA-indistinguishability of the target model for multiple member rates (defined in Section 3.3) following Equation 2.19.

3.2 Dataset Preparation

We select three extensively used datasets in studying MIA [37, 20, 35] and create multiple modified versions of them to perform different experiments. The datasets are pre-processed as follows:

- UCI Adult dataset [44]: This dataset contains individuals' records classified into two groups based on whether a person makes over $50k per year. Total number of records is $48,842$ with 14 census features such as age, gender, education, marital status, occupation and working hours. In different experiments we use $1000 - 10,000$ randomly sampled records from the dataset.

- Purchase dataset [2]: This is an unlabeled dataset containing records of the customers and their purchasing history. To prepare the dataset, we combine two datasets from the data source [2]. The "transactions" dataset contains the customer's buying history and The "history" dataset contains the incentives offered to them. We obtain 16 features by joining them including chain, category, purchase quantity, purchase amount, offer, market and repeater. We prepare the primary dataset by randomly sampling 400,000 records. Later in different experiments we use 10,000-100,000 records randomly sampled from the primary dataset. Labels of the dataset are assigned using K-means clustering. In earlier works[37], the customers were clustered into $2, 10, 20, 50, \ and \ 100$ classes based on their buying patterns. However, we cluster the records into 2 classes $\{0, 1\}$ to measure the balance between them during different experiments.

- Texas hospital dataset[18]: This dataset is based on publicly available Texas Hospital Discharge Data with information on inpatient stays in several health facilities, released by the Texas Department of State Health Services. We use $400,000$ records randomly sampled from years 2006 to 2009 and 16 features such as patient's gender, country, race, principal surgical procedure code & day, risk mortality and illness severity. We label the records into two classes $(0, 1)$ denoting whether a patient got immediate response by calculating the difference between the date of admission and the date of principal surgery. In different experiments we use $10,000 - 100,000$ randomly sampled records from the primary dataset.

3.3 Dataset Customization

Customizing the datasets based on different data properties is a crucial part of our study. During an experiment to capture the effect of a certain property, we keep consistency in the other property values as far as possible to understand that particular property's sole impact. For example, while evaluating the effect of variable balances in the features on MIA, the data size and level of balance in the classes are kept consistent. A consolidated details of the property values used in different experiments are summarised in Table 3.1.
Details of the dataset preparation for different experiments are as follows:

Evaluated Property	Controlled Properties	Other Properties
Data size	10000 to 100000 (10000 interval) 1000 to 10000 (1000 interval)	Balance in the classes 10% and 50% Number of features 16 (14 for Adult)
Balance in the classes	0% to 50% (10% interval)	Data size 100,000 (10,000 for Adult) Number of features 16 (14 for Adult)
Balance in the features	0% to 50% (10% interval)	Data size 100,000 (10,000 for Adult) Balance in the classes 10% and 50% Number of features 5
Number of features	1 to 16 features (1 feature interval) 1 to 14 for Adult dataset	Data size 100,000 (10,000 for Adult) Balance in the classes 10% and 50%
Entropy	Property not controlled; Measured on different datasets	Data size 100,000 (10,000 for Adult) Balance in the classes 10% and 50% Number of features 16 (14 for Adult
Mutual information between record and model parameters $I(X;\theta)$		
Model overfitting		
MIA-indistinguishability	Member-rate 10% to 90% (10% interval)	
Target-shadow model combinations	Implemented 5 ML models (Table 3.3) 1-target vs. 1-shadow model 1-target vs. 5-shadow models (different models)	
Choice of model hyperparameters	ANN with multiple set of hyperparameters (Table 3.2)	
Model fairness $(\delta_g, \delta_p, \delta_i)$	Property not controlled; Measured on different datasets	Data size 100,000 (10,000 for Adult) Balance in the classes 10% to 50% (10% interval) Balance in the features 10% to 50% (10% interval) Number of features 5
For all experiments a default ANN is used as the ML model (ANN hyperparameters are chosen as described in Table 3.3) except "Target-shadow Model Combination" experiment		

TABLE 3.1: Details of the property values used in different experiments

- We create datasets containing $10,000$ to $100,000$ records with $10,000$ interval for Purchase and Texas datasets and $1,000$ to $10,000$ records with $1,000$ interval for Adult to evaluate the effect of the data sizes on MIA.
- To evaluate the impact of the balance in the classes on MIA we sample records keeping the rate balance rate of the class label '1' from 0% to 50% with 10% interval.
- For estimating how the balance in the features influences the MIA, we create the datasets with only 5 selected features and tuned their feature balances between 0% and 50% with 10% interval by sampling records accordingly.
- For analysing the effect of the number of features on MIA, we generate multiple datasets having 1 to 16 features (14, in case of Adult dataset).
- To understand the impact of the model fairness properly, we consider datasets with 5 features having both the classes and the features balanced from 10% to 50% with 10% interval.
- For all the experiments we always use 25% member rate while splitting the target dataset into train and test records except for the experiment on MIA-indistinguishability, where we measure MIA-indistinguishability for the different member rates from 10% to 90%, with a 10% interval.

We use $100,000$ records ($10,000$ in case of Adult dataset) with all the features and with two levels of balances (10% and 50%) in the classes as default setting in all the other experiments.

hyperparameters	Values
Hidden layers	1-layer to 5-layers
Number of nodes	5, 50, 100, 500
Learning rate, α	0.00001, 0.0001, 0.001,0.01, 0.1
L2-ratios, λ	0, .001, .01, .1, 1, 2, 3

TABLE 3.2: Different hyperparameter's values used in the ANN model to study their impact on the MIA attack accuracy.

Model	hyperparameters
Logistic Regression (LR)	C=0.01, solver= LBGFS
Support Vectopr Machine (SVM)	C=0.01, kernel= RBF
Random Forest (RF)	n-estimators=100, criterion=gini, max-depth=2
K-Nearest Neighbour (KNN)	p=2, neighbors=3
Artificial Neural Network (ANN)	α = 0.001, solver= sgd, epochs=50

TABLE 3.3: Hyperparameters selected for the different models used while exploring target-shadow model combinations against MIA.

3.4 Model Setup

We implement MIA according to Algorithm 1, where both the target and shadow datasets D_T and D_S are sampled distinctly from the total population D. The default ANN model used as the target, shadow and attack model for every experiment is strutured as a one hidden layer network and 50-nodes with other hyperparameters selected as $\alpha = .001$, solver=sgd and epochs=50 (similar to the ANN setting in Table 3.3). Furthermore, to reduce the computational complexity, we use only one shadow model following the work of Salem et.al. in [34], as their experiments show that similar attack accuracy were observed when using one shadow model instead of multiple shadow models as in [37].
We also investigate how choosing different hyperparameters for a neural network may affect the MIA. For this experiment, we probe ANN models containing 1 to 5-hidden layers with different number of nodes, learning rates α and L2-ratios λ using L2-regularizer. The detailed ranges of the hyperparameters probed for this experiment are given in Table 3.2.
Five models are chosen to study the effect of using different classifiers as shadow models on the attack accuracy. The models are Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), K- Nearest Neighbour (KNN) and ANN. Choice of the hyperparameters for all the models are listed in Table 3.3. We experiment on target and shadow model combinations in two settings: one-to-one and one-versus-all. In one-to-one setting, the models are examined against each other using only one shadow model. On the other hand, in the one-versus-all setting, each model is tested against five shadow models each structured as one of the considered models.

3.5 Evaluation Metrics

We use the classification accuracy score of the attack model to evaluate MIA's performance and denoted as the attack accuracy throughout the thesis. The score reflects the number of correctly

	True Member ($m = 1$)	True Non-member ($m = 0$)
Predicted Member ($f_a(X) = 1$)	True Positive (TP)	False Positive (FP)
Predicted Non-member ($f_a(X) = 0$)	False Negative (FN)	True Negative (TN)

TABLE 3.4: Confusion matrix for evaluating the performance of MIA

inferred membership information about the desired records by the attack model and can be defined as:

$$Attack\ accuracy = \frac{Number\ of\ Correctly\ Predicted\ Memberships}{Total\ Number\ of\ Predictions} \tag{3.1}$$

As a binary classifier, MIA's attack accuracy can be written in terms of the positive and negative predictions (Table 3.4) as below:

$$Attack\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

The attack precision and attack recall values can be calculated as below:

$$Attack\ Precision = \frac{TP}{TP + FP} \tag{3.3}$$

and,

$$Attack\ Recall = \frac{TP}{TP + FN} \tag{3.4}$$

We compute Pearson's correlation coefficients for each of the properties against attack accuracy for better understanding of how they correlate to each other. Pearson's correlation coefficient formula is given below:

$$\rho = \frac{\text{COV}(p, a)}{\sigma_p \sigma_a}, \tag{3.5}$$

where ρ is the correlation coefficient, p represents the different property values and a is the accuracy of membership inference obtained on them. $\text{COV}(p, a)$ is essentially the covariance between the variables, and σ_p and σ_a are the standard deviations of p and a, respectively. A positive value of the coefficient suggests that increase in the property value boosts the attack accuracy, while the negative value refers to the opposite. The higher the value for ρ is, the higher the correlation between the evaluated property and MIA accuracy.

4

Exploratory Analysis of the Impact of Different Properties on MIA

This chapter illustrates all the experimental results showing the performance of MIA in correspondence to the investigated data and model properties. We describe MIA's performance in terms of the attack accuracy, precision and recall with respect to the changes in the properties. The acquired results reveal that, most of all the different properties affects MIA's performance indicating strong relationships among them. Moreover, the Pearson's correlation coefficients computed between the properties and the attack accuracy also support our findings. Table 4.1 shows the correlation coefficients for some of the evident properties with the attack accuracies. Among all the properties, data size, balance in the classes, group fairness and mutual information between records and model generated parameters show a strong correlation with MIA.

	Adult	Purchase	Texas
Datasize	0.280	0.327	0.337
Balance in the classes	-0.940	-0.181	-0.015
No of features	-0.146	-0.181	-0.187
Balance in the features	0.011	0.045	0.009
Entropy	-0.221	-0.018	-0.006
Individual fairness	0.067	0.041	0.038
Group fairness	0.313	0.475	0.291
Predictive fairness	0.028	0.043	0.020
$I(X;\theta)$	0.246	0.303	0.140
MIA-indistinguishability	-.03.495	-.008623	-.047929

TABLE 4.1: Person's correlation coefficients calculated between different properties and MIA attack accuracy.

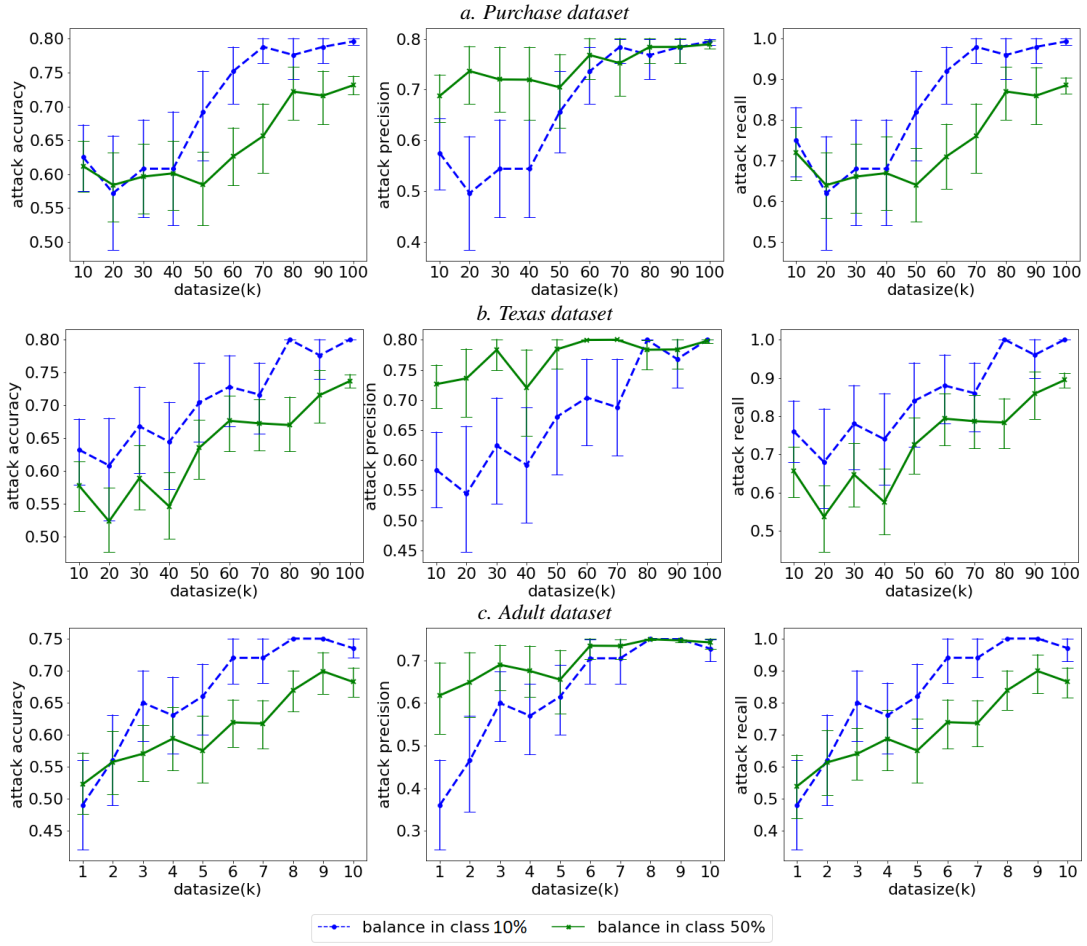

FIGURE 4.1: Data size vs attack accuracy, attack precision and attack recall: results obtained on different datasets (confidence Interval=95%). Figure shows that increasing the data size increases the severity of MIA in the case of both 10% and 50% class balances.

1-hidden layer, no regularizer (λ=0)

Datasets	Train Acc	Test Acc	Attack Acc
Adult	80.956	81.232	58.893
Purchase	71.512	71.405	54.121
Texas	78.465	78.226	57.716

a

5-hidden layers, no regularizer (λ=0)

Datasets	Train Acc	Test Acc	Attack Acc
Adult	89.806	89.782	71.080
Purchase	88.802	88.780	71.736
Texas	90.993	90.862	71.257

b

1-hidden layer, with regularizer (λ=.01)}

Datasets	Train Acc	Test Acc	Attack Acc
Adult	74.972	75.076	55.977
Purchase	64.829	65.532	56.36
Texas	75.974	76.074	49.891

c

5-hidden layers, with regularizer (λ=.01)

Datasets	Train Acc	Test Acc	Attack Acc
Adult	88.364	88.355	66.955
Purchase	88.33	88.358	70.211
Texas	90.498	90.473	72.664

d

TABLE 4.2: Target model's train and test accuracies (%) versus obtained MIA attack accuracies (%) for the models having a) 1 and b) 5-hidden layers respectively, without using any regularizer, whereas, c) and d) illustrate the results with regularizer. Data size is 100,000 (10,000 for Adult) and class balance is 10%. Selected hyperparameters: α: .001, λ: .01 and number of nodes in each hidden layer: 5.

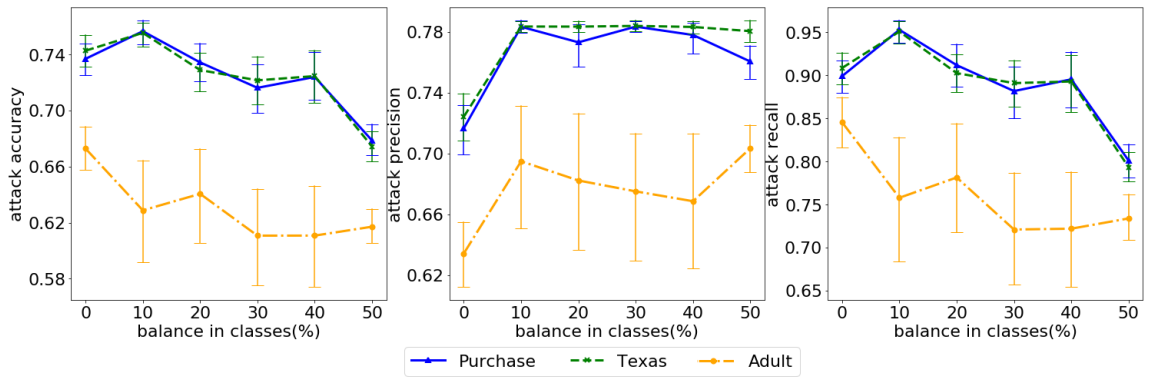

FIGURE 4.2: Different levels of balance in the classes vs attack accuracy, attack precision and attack recall for Purchase, Texas and Adult dataset ($CI = 95\%$). Datasets with properly balanced classes result in low attack accuracy.

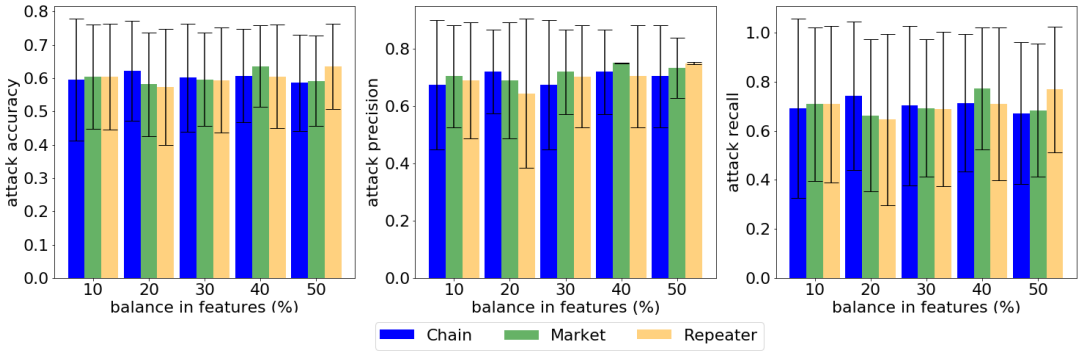

FIGURE 4.3: Attack accuracy, attack precision and attack recall for the different levels of feature balances in the three selected features from Purchase dataset ($CI = 95\%$). Datasize is 100,000 and class balance is 10%.

4.1 Effect of Different Data Properties

1. Data Sizes: From the experimental results obtained on all three datasets, it is evident that the size of the dataset has a huge impact on the MIA's success. The adversary has more advantage in attacking rather large datasets. On the same model structure, regardless of the highly balanced or imbalanced class labels, an increase in the data size enhances the attack accuracy. Figure 4.1 illustrates the results acquired on datasets with different data sizes. For all the datasets, the shadow model's data sizes are twice larger than the target model's data sizes $(1 : 2)$. Note that, a similar trend in the MIA's success is obtained when the data size of the target and shadow models are equal $(1 : 1)$. Table 4.1 also indicates a strong positive correlation between data sizes and attack accuracy for all three datasets. The attack accuracy and recall are higher for the datasets with imbalanced class labels than the balanced ones, while attack precision is lower for the datasets with imbalanced class labels. That means while more TPs are predicted more FPs are also included in the prediction when classes are imbalanced. However, TPs have more cost than FPs in this problem settings. We obtain the similar trend of attack precision and attack recall in most of the experiments.
2. Balance in the Classes: Figure 4.2 shows the attack accuracy for different balances in the class labels for all three of the datasets. When a dataset has a proper balance in the class labels(50%, i.e. $1 : 1$), the MIA is less successful. In terms of an ML model's

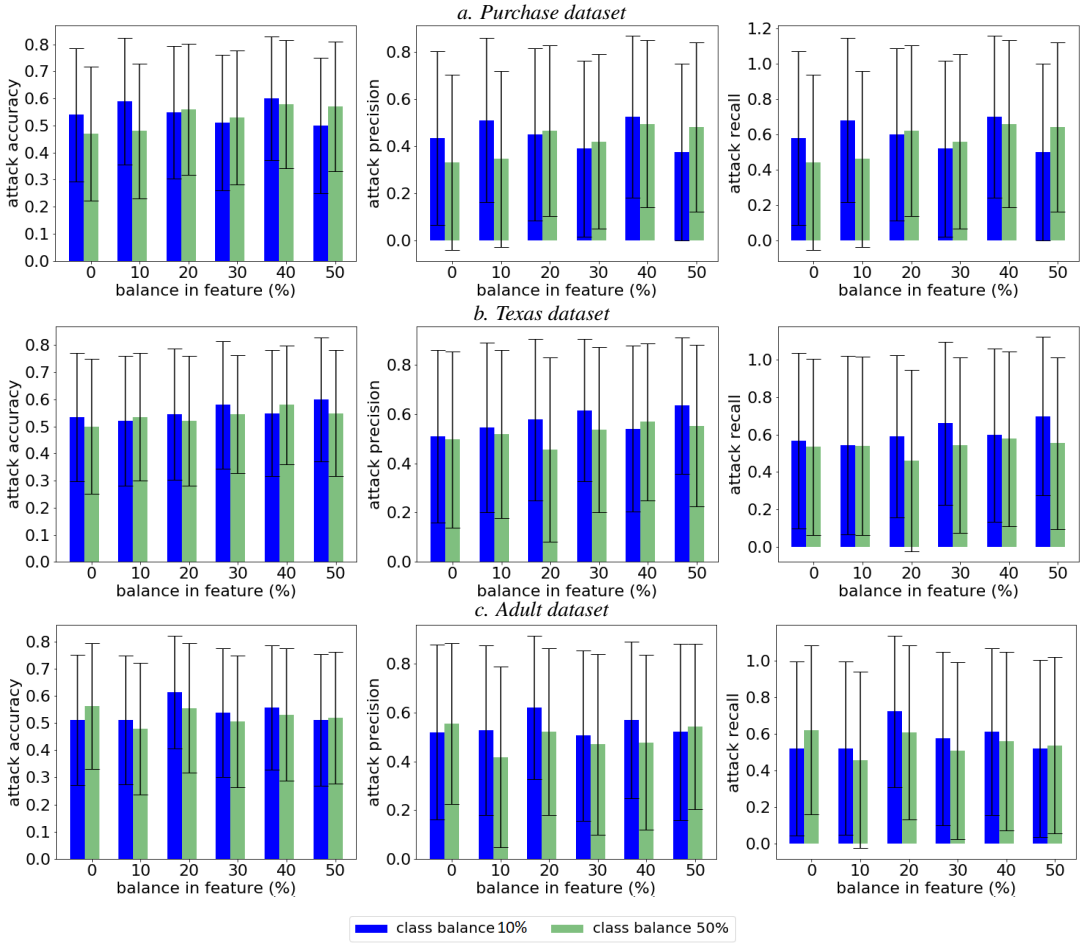

FIGURE 4.4: Attack accuracy, attack precision and attack recall obtained for the different levels of feature balances in the cases of two class balances- 10% and 50% ($CI = 95\%$). Number of features in each of the datasets is 5. Data size is 100,000 (10,000 for Adult dataset)

prediction, this behaviour explains that the model produces less biased prediction towards one particular class when it is trained on a dataset that has properly balanced classes, hence giving less benefit to the adversary. From the Table 4.1, we can also observe a negative correlation between balance in the classes and the attack accuracy. Similar trend with attack accuracy, precision and recall are achieved as described above for data size results. In addition to this, from the figure, it can be seen that, although the attack accuracy is lower for 50% class balance than 0% class balance, the lowest attack accuracy value is not always achievable from a perfectly balanced class. For example, in case of the Adult dataset, the lowest attack accuracy value is obtained with 30% balance level in class labels.

3. Balance in the Features: In case of the balance in the features, the effect of each feature is different when they are experimented individually against MIA. Figure 4.3 shows the performance of the attack on three selected features from the Purchase dataset investigated separately. Similar patterns are found after carrying out the experiments on 5 features selected from Purchase, Texas and Adult datasets that show no consistency in the increase or decrease of the attack accuracy with respect to the increase in the feature balances. The results are depicted in Figure 4.4. However, attack accuracy is lower as expected for the class balance 50% compared to the class balance 10%. Although, according to the Table

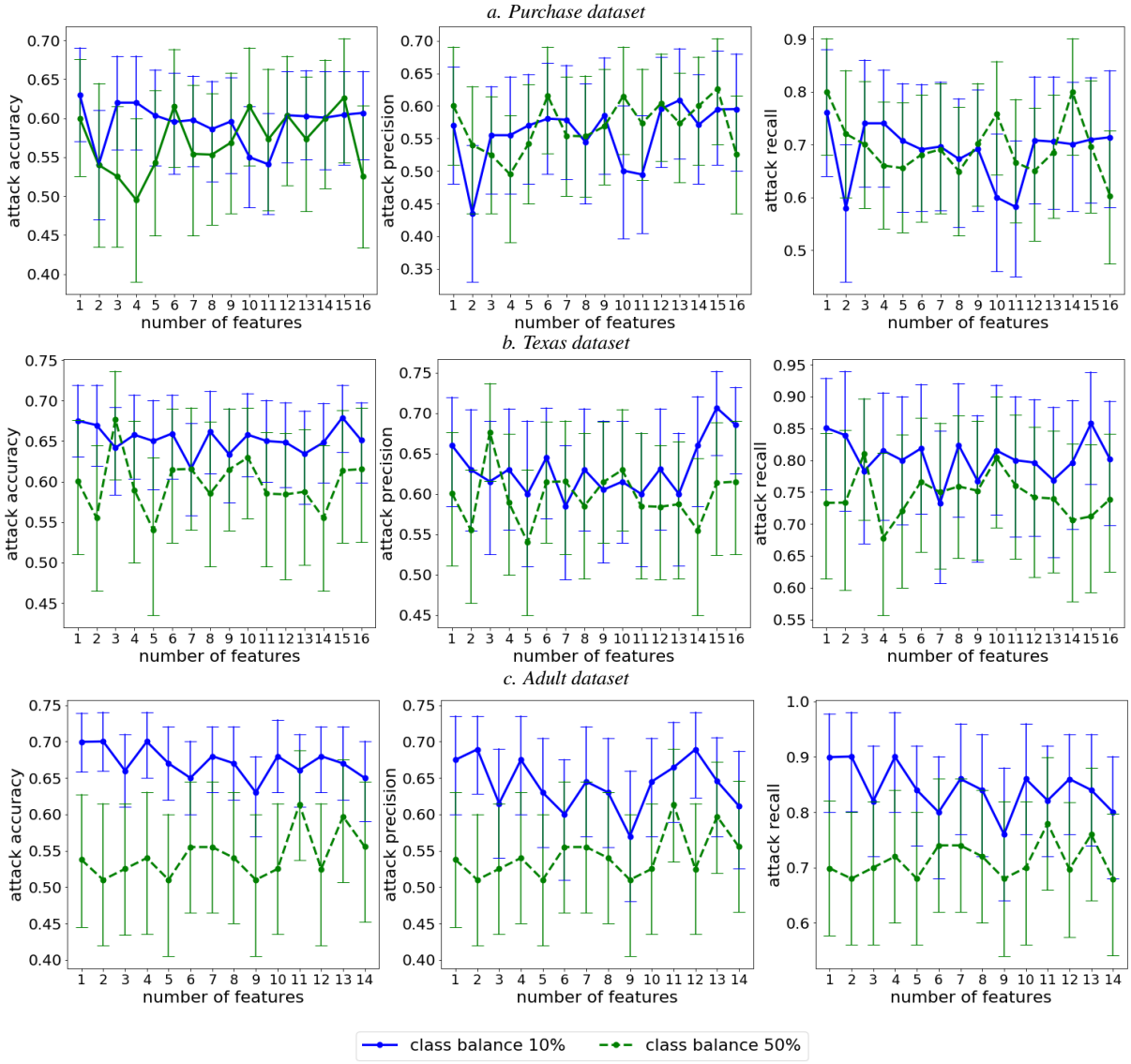

FIGURE 4.5: Attack accuracy for the datasets containing 1 to 16 features (14 featuers in the case of Adult dataset; $CI = 95\%$). Balance in the classes are 10% and 50%. Data size is 100,000 (10,000 for Adult dataset).

4.1, the balance in the features and the attack accuracy is positively correlated, it is hard to make any straightforward conclusion on how the balance in the features affects MIA's success. The results demonstrate that several features combined together by keeping a certain balance among them may prevent membership inference better, which needs further exploration.

4. Number of the Features: Introducing more features in a dataset increases dimensionality in the dataset, that makes it difficult to perform MIA by the adversary. Figure 4.5, represents the observed attack accuracy for each dataset with a gradual increase in the number of features. Although there is no straightforward trend in the attack accuracy, from the figure, it can be understood that a certain combination of features in a dataset shows better defence against MIA. For instance, in case of the Purchase dataset, mean attack accuracy for a 2-features dataset is 64%, while for the 16-features the accuracy decreases to 57.5%. Further exploration may reveal the actual reason behind this behavior of the dataset.
5. Entropy: The entropy of a dataset determines the level of randomness among the features.

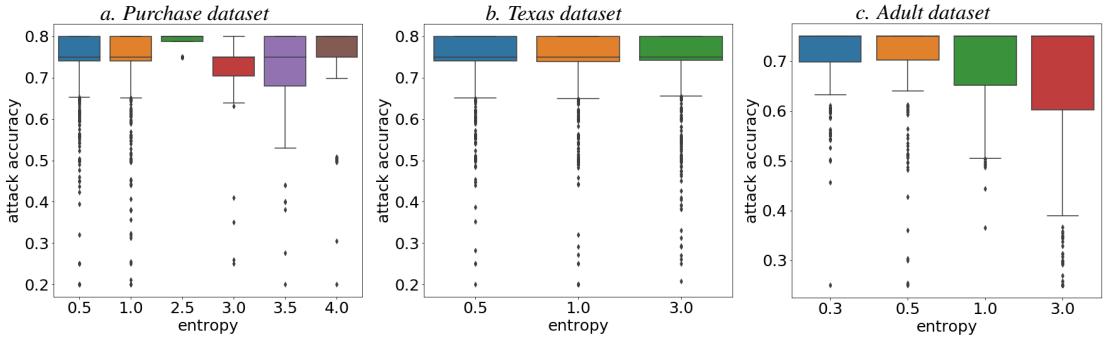

FIGURE 4.6: Relationship between the entropy of the datasets vs. attack accuracy for 10% balance in the classes. Data size is 100,000 (10,000 for Adult dataset).

Large datasets with many features render higher entropy value. For a limited number of features used in the dataset, the result still shows high attack accuracies for the small entropy values. Figure 4.6 shows the results of this experiment for the datasets with 10% balance in the classes. However, from the result, it is hard to derive any conclusion about the relationship between the attack accuracy and entropy except the Adult dataset. In the case of Adult dataset, attack accuracy range is lower for low entropy value. We followed a similar trend for the 50% balance in the classes as well. Although the correlation between the entropy and the attack accuracy is negative according to the Table 4.1, the experimental results suggest the necessity of further investigation on this property.

4.2 Effect of Different Model Properties

1. Selection of Model's Hyperparameters: Properly tuned hyperparameters has a tremendous contribution to a model's performance. To understand the effectiveness of a proper selection of the hyperparameter values against MIA, we experiment on multiple hyperparameter settings based on the number of layers, number of nodes in each layer, the learning rate and L2-ratio. Obtained results for the models having 1 to 5-hidden layers are presented in Figure 4.7. From the figure it can be observed that, having a higher number of nodes on each layer increases attack accuracy. The figure also demonstrates a significant increase in the attack accuracy for a slight increase in the learning rate (α) of the target model. Also, the selection of the L2-ratio (λ) to control the amount of regularization impacts MIA significantly. More regularization decreases attack accuracy. However, after a certain point ($\lambda > .1$), increasing the regularization value seems to behave the opposite.
2. Target-Shadow Model Combination: In one-to-one setting we find that, in general, shadow models structured as ANN and RF may yield higher attack accuracy against most of the models (Figure 4.8). We also observe that, ANN shows maximum vulnerability against shadow model built as ANN. However, for lower data size in Adult dataset (10,000) ANN also proves to be vulnerable against SVM. Furthermore, it is evident that, shadow models built as similar to the target do not guarantee maximum attack accuracy. However, when each model encounters all five shadow models (right-side graphs in Figure 4.8), the attack accuracy surges higher in case of all the target models. From the experimental results it can be realized that, the success of MIA is highly classifier dependent and combined

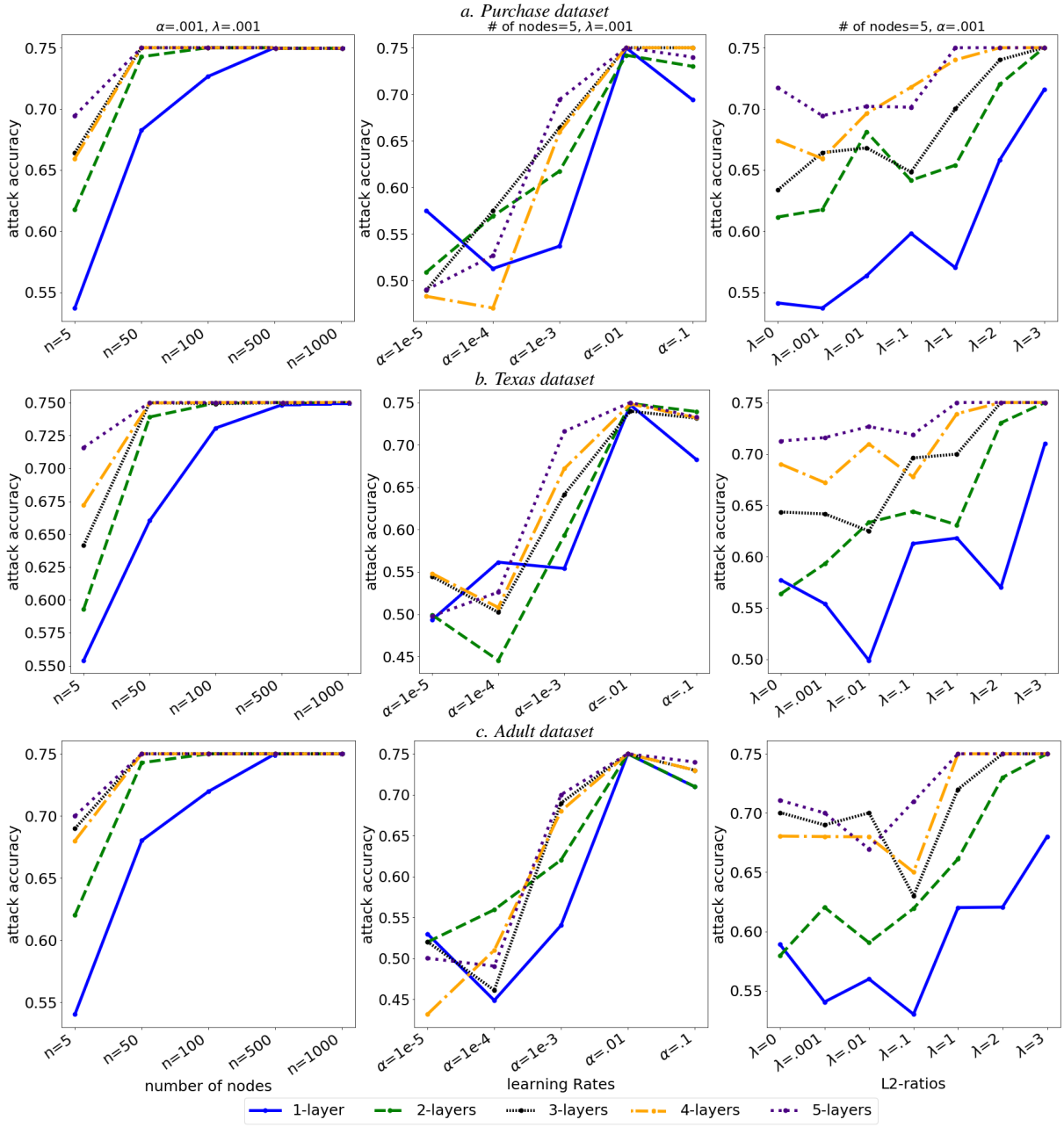

FIGURE 4.7: MIA accuracy for models with different combinations of hyper parameters. Left-most graphs show the effect of increasing the number of nodes on each layer on MIA. The graphs in the middle show the effect of changing the learning rate and the right-most graphs show the effect of changing the L2-ratio. Data size is 100,000 (10,000 for Adult dataset) and class balance is 10%.

attack of multiple models against one model is more severe.

3. Mutual Information between Records and Model Parameters: Mutual information between model-generated parameters and the records $I(X;\theta)$ can capture the information learned by the model over a dataset. Higher mutual information between them indicates that the model captures more information from the records, which in turn could result in vulnerability towards membership inference. The resulting figure (Figure 4.9) of the experiments and the correlation coefficient (Table 4.1) also supports this interpretation. For all the datasets, attack accuracy values soar up for a small increase in the values of I.

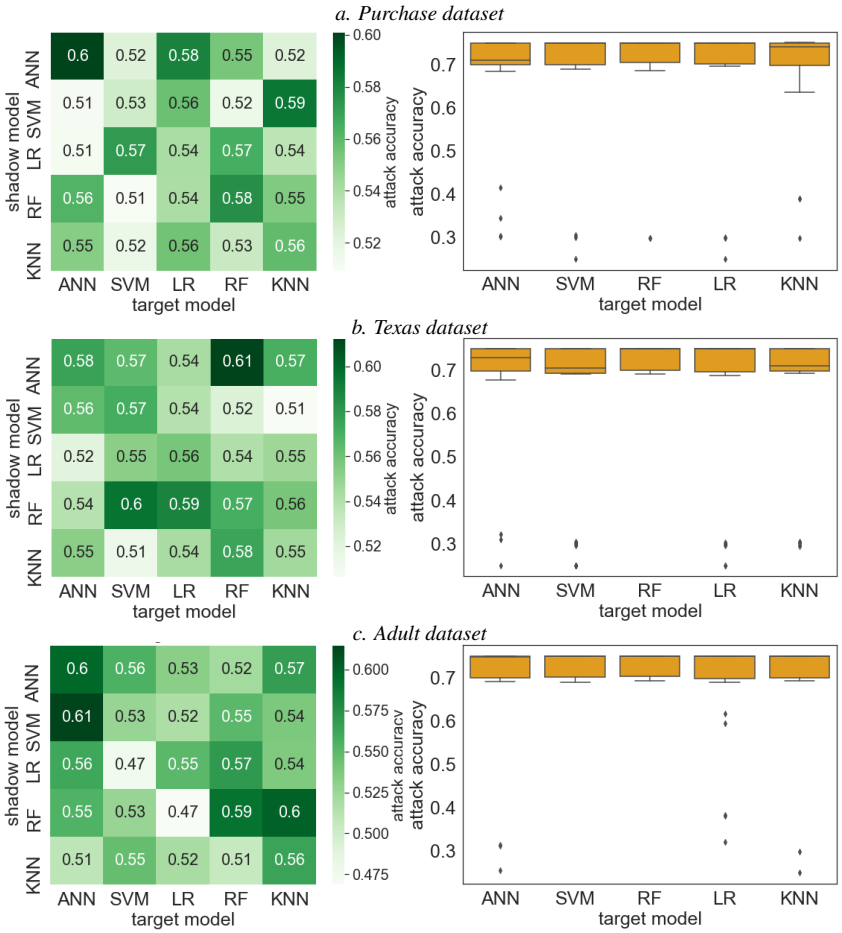

FIGURE 4.8: Attack accuracy obtained using different ML algorithms as target and shadow models. Left-side figures show the level of attack accuracy achieved from one-to-one setting between the target and shadow model combinations. The right-side figures show attack accuracy for each of the target models against five shadow models using all the five examined algorithms.Data size is 100,000 (10,000 for Adult dataset) and balance in the classes is 10%

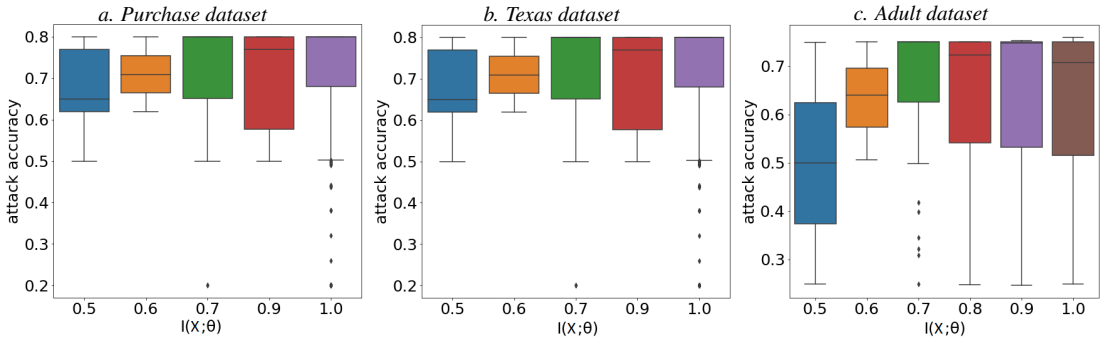

FIGURE 4.9: Relationship between $I(X;\theta)$ and attack accuracy. Here, I is the mutual information between the features of a dataset and the model parameters estimated on them. Class balance is 10% and data size is 100,000 (10,000 for Adult dataset)

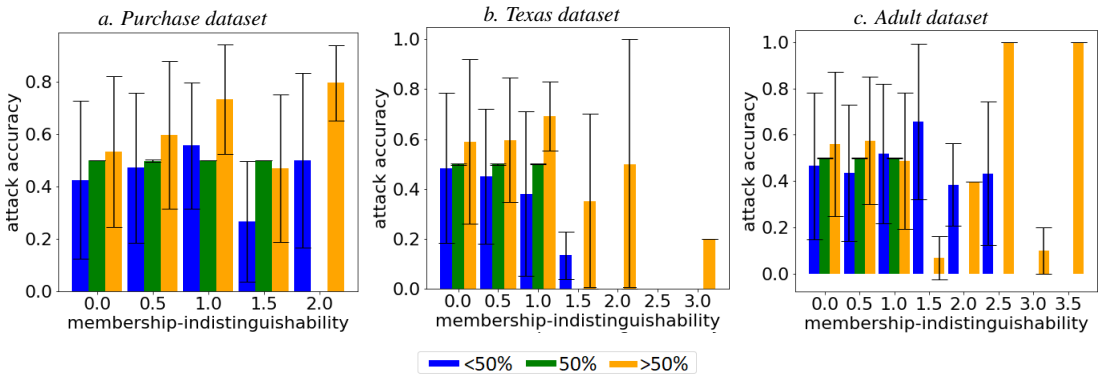

FIGURE 4.10: Relationship between the MIA-indistinguishability and attack accuracy for the member rates of <50%, 50% and >50% ($CI = 95\%$). Results obtained on 100,000 data (10,000 for Adult dataset) with 10% balance in the classes.

4. Model's Fairness: Fairness is one of the high-impact model properties on MIA. Table 4.1 shows a strong correlation between the fairness and attack accuracy. We measure fairnesses as the difference between two groups of records. Thus, lower fairness difference represents a fairer model. Figure 4.11 depicts the results of our studies on three different fairnesses- group, predictive and individual fairness for multiple balances in the classes and features. The figure clearly shows that, for a model with a lower fairness difference, the attack is comparatively weaker. Also, there is a significant positive relationship between balance in the features and classes with model's fairness. For instance, a model depicts lower fairness difference for the datasets with properly balanced features and classes.
5. MIA-indistinguishability: The vulnerability of a model towards MIA in the form of MIA-indistinguishability is explored for different member rates used in the target model. MIA-indistinguishability is measured as the difference between the member and non-member predictions. Hence, a lower value of MIA-indistinguishability suggests a model with similar member and non-member predictions. Figure 4.10 illustrates results obtained from the experiment. We observe that, when the member rate is 50%, the range of MIA-indistinguishability is very low and attack accuracy is always close to the random guess (50%). For both the cases where the member and non-member ratio is unequal (member rates >50% and <50%), increase in the MIA-indistinguishability boosts the attack accuracy, except for the Texas dataset. This results require a further assessment in order to verify the relationship between MIA and MIA-indistinguishability. The correlation coefficients between MIA-indistinguishability and attack accuracy indicate a negative correlation as expected (Table 4.1).
6. Model's Overfitting: The significance of a model's overfitting tendency is well-explored by multiple pieces of research [48, 26, 34]. In our experiments, in order to observe the contribution of the overfitting to MIA's success, we use multiple hyperparameter settings that depict very low overfitting on all the datasets. Acquired test accuracy, train accuracy and attack accuracy values from these models are illustrated in Table 4.2. In the table we compare ANNs consisting of 1 and 5-hidden layers in the cases of both with and without using a regularizer. It is evident from the results that even though the models are not overfitted, increasing the number of hidden layers shows an increase in the attack accuracy. Also, model's with regularizer (L2) yields lower attack accuracy compared to the model with the same structure but no regularization.

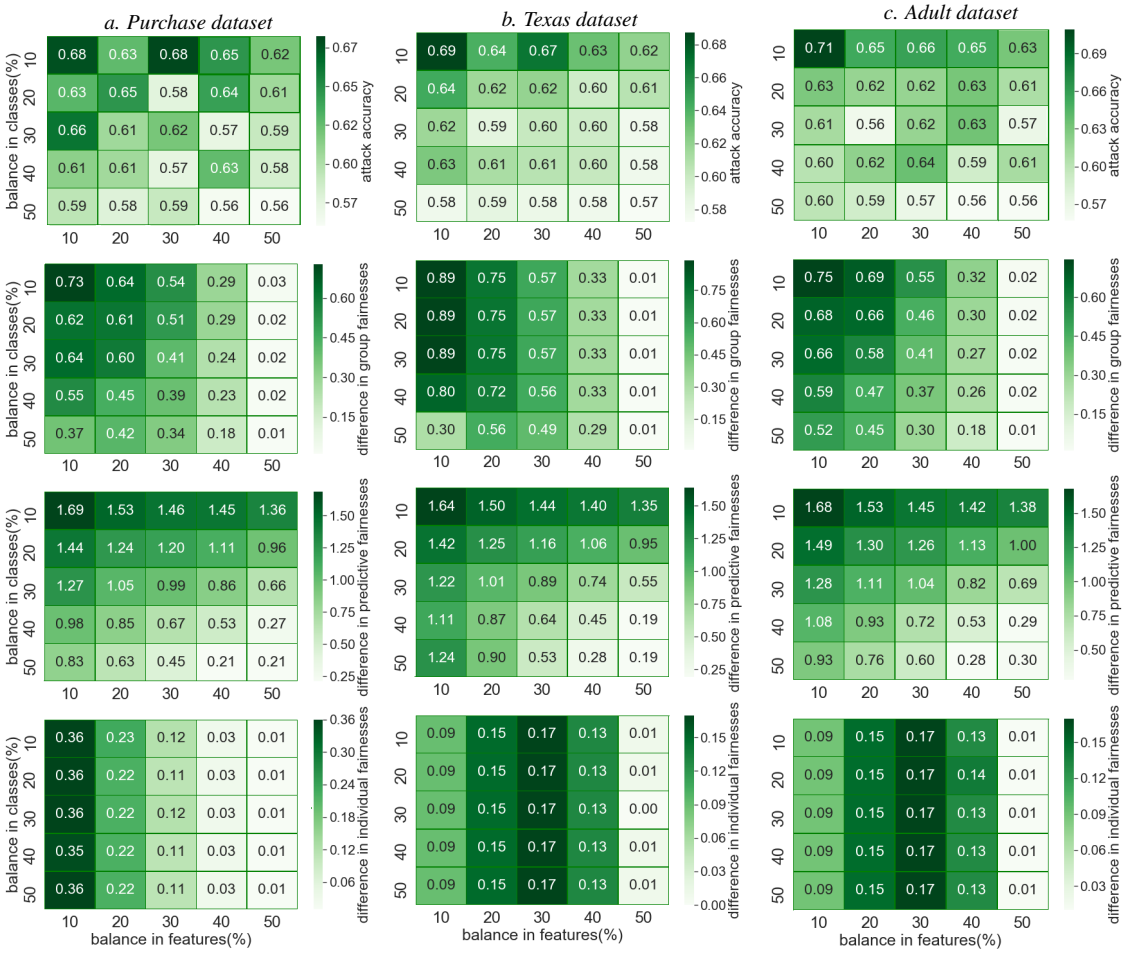

FIGURE 4.11: Relationship between fairnesses and attack accuracy for the datasets with 10% to 50% class and feature balances. The graphs in each column show the attack accuracy, differences in group, predictive and individual fairness respectively for each of the datasets. Data size is 100,000 (10,000 for Adult dataset) and the number of used data features are 5.

In summary, to prevent information leakage, the major concerning properties regarding the data are the data size and balance in the classes. Although using smaller datasets limit the fruitfulness of the training, the adversarial advantage over the huge data should be estimated before deploying them in MLaaS platform. In addition to that, sampling records with the class labels proportionately balanced would also reduce the impact of MIA and may foster fairness in the model. Besides the data size and the class balances, features used in the dataset may also be tested against MIA in several combinations with different balances among them to identify an optimal set of features that may expose less information. In the case of a model, hyperparameter selection, mutual information between the records and model's parameters learned on them and model's fairness has more impact on MIA. However, in reality, it would be extremely difficult to control the properties to minimize information leakage, especially the data properties. Thus, we attempt to elevate the model's resilience against MIA according to the above observations. We proceed further to develop an optimal MIA-resistant model by utilising the values of the dominant model properties. Chapter 5 describes our method and findings on the experiments of the MIA resilience in detail.

5 Towards MIA-resilient ML models

In the previous chapter, we present the results of our empirical study to identify factors that contribute to MIA's success. Our findings inspired us to further look into preventing ML models from leaking membership information. In this chapter, we propose a new technique by introducing custom regularizer in the model that aims at strengthening the model against MIA with a successful reduction in the attack accuracy. To validate our proposed technique, we implement multiple model properties as regularizers in the model and observe the impact on both MIA and model's prediction. The results demostrate that, the regularizers reduce MIA accruacy and at the same time, improve model's predictability by inducing better generalization of the parameters.

5.1 Tuning the Model for Better Resilience

Previous research showed that using a well-generalized model with lower overfitting that produces similar number of correct predictions for both training and testing records is one of the ways to prevent MIA [48]. However, from the previous chapter, we learn that models with marginal overfitting can also result in high MIA accuracy (results acquired on model's overfitting are discussed in Section 4.2). In addition, compared to the overfitting, model properties such as fairness difference between two record groups, and the mutual information between the records and model parameters ($I(X;\theta)$) show greater influence on MIA accuracy. As both of these properties are negatively correlated to the attack accuracy, decrementing them towards an optimal value would essentially improve the model's performance. Also, one of the existing defences proposed in [30], suggests a game-theoretic approach to minimize classification loss of the model by using the maximum gain of MIA as a regularizer.
Based on the above analysis, we investigate a novel technique to improve the underlying model's performance as well as the resilience against MIA by using property values as regularizers in the model. We explore following four model properties that showed high correlation with MIA accuracy in our experimental results presented in the previous chapter:

- Difference in the group fairness, δ_g,
- Difference in the predictive fairness, δ_p,

- Difference in the individual fairness, δ_i and,
- Mutual information between the records and the model parameters, $I(X;\theta)$

Algorithm 2 Group Fairness as a Regularizer

$D : X \times y$ is the total population where $g_1, g_0 \in D$ are the data points resembling two groups of individuals with $g_1 \cap g_0 = \phi$

1: **for** k number of epochs **do**
2: **for** l steps **do**
3: Randomly sample m data points from g_1 and n data points from g_0
4: Use the records to train the target model f_t and measure the group fairness δ_g as:
▷ According to Equation 2.8

$$\delta_g = \frac{1}{2}\sum_{\hat{y}\in y} |P_{g_0}(\hat{y}) - P_{g_1}(\hat{y})|$$

5: Set $r \leftarrow \max_l \delta_g$
6: Update f_t by descending its stochastic gradients over its parameters θ using below equation:

$$\nabla_\theta \frac{1}{m+n} \sum_{i=1}^{m+n} l(\hat{y}, y_i) + \lambda \sum_j |\theta.e^r|$$

Algorithm 2 gives an outline of how a model's fairness difference in the groups can be used as a regularizer in that model. The algorithm starts by sampling records from the two groups of individuals and obtaining the group fairness difference δ_g according to Equation 2.8 by training the target model f_t. After repeating the steps for l times (steps 2-4 in Algorithm 2), the maximum gourp fairness difference r of the model is estimated (step 5). Finally below regularizer function is used to update the gradients (step 6):

$$R(\theta) = \sum_j |\theta.e^r|,$$

Where j is the number of parameters θ generated by the model in one epoch.

	a			b			c		
Regularizers	Train Acc	Test Acc	Attack Acc	Train Acc	Test Acc	Attack Acc	Train Acc	Test Acc	Attack Acc
None	73.04	72.80	73.89	76.17	75.87	74.04	76.54	76.52	73.36
δ_g	85.00	84.72	50.03	84.75	84.71	53.27	84.49	84.52	52.33
δ_p	84.72	84.59	48.84	84.73	84.80	58.53	84.57	84.15	51.84
δ_i	84.89	84.93	51.99	84.71	84.72	58.15	84.41	84.60	50.25
$I(X;\theta)$	73.14	73.35	51.15	70.93	70.77	59.40	70.14	70.31	53.34
L1	76.78	76.81	55.29	79.93	80.34	55.76	76.42	76.72	52.41
L2	76.69	76.94	59.13	78.63	78.74	55.72	69.64	69.83	53.95

TABLE 5.1: Train, test and attack accuracy (%) for different regularizers applied in the models for a) Purchase, b) Texas and c) Adult datasets with 5 features and 10% balance both in the classes and features.

To evaluate the model's performance, we repeat the steps for 50 epochs and measure the train

loss and difference in the group fairness in each epoch. We have also estimated the train and test accuracy of the model and attack accuracy after performing MIA on the model. We follow a similar algorithm in implementing models with other properties as regularizers. We use datasets with 100,000 records (10,000 records for Adult dataset), 5 features and 10% balance in the classes in these experiments.

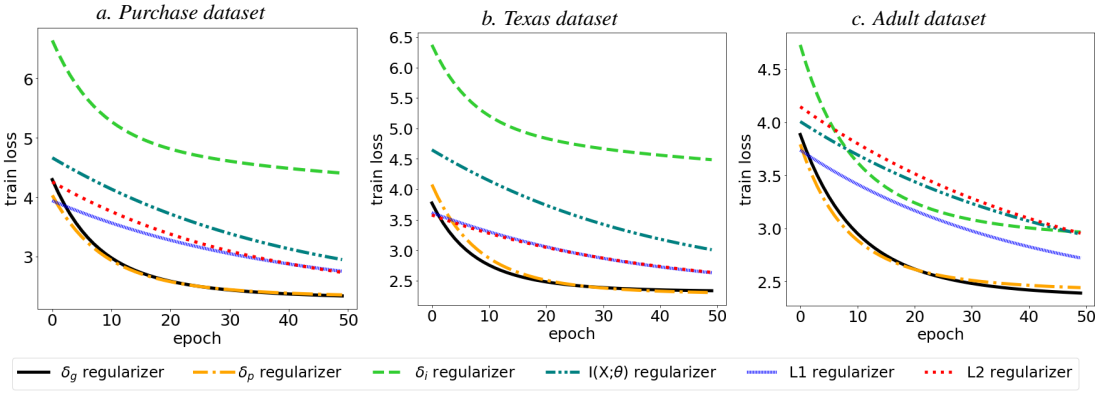

FIGURE 5.1: Comparison of the model's training loss on each epoch using δ_g, δ_p, δ_i, $I(X, \theta)$, L1 and L2 regularizers. The number of used features is 5, data size is 100,000 (10,000 for Adult dataset) and class balance is 10%.

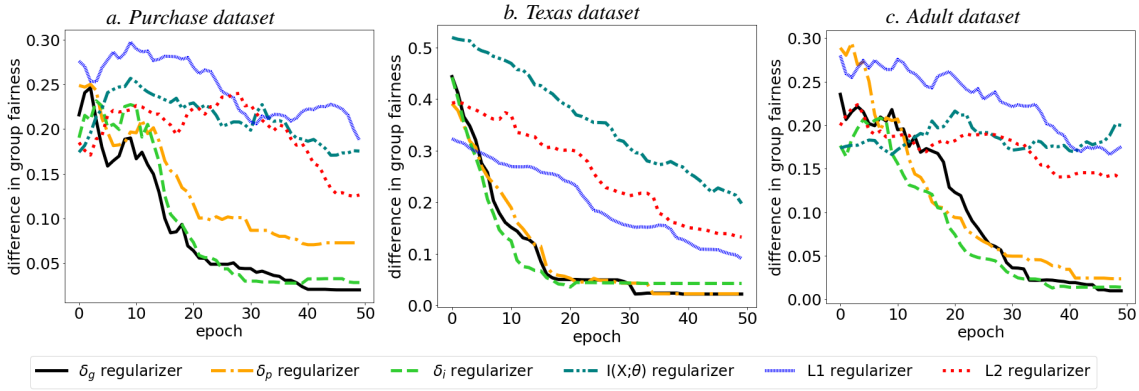

FIGURE 5.2: The decrease in group fairness differences in each epoch for multiple models applying δ_g, δ_p, δ_i, $I(X, \theta)$, L1 and L2 regularizers.

5.2 Experimental Evaluation of the Proposed Regularizers

For all the four studied regularizers, the obtained results show success both in terms of improving the model's performance and reducing MIA accuracy. The attack accuracy values as reported in Table 5.1 show that the lowest attack accuracy was achieved by using one of the proposed regularizers compared to the model without any regularizer or using L1 and L2 regularizers. As expected, all the regularizers perform significantly better in terms of reducing attack accuracy compared to using no regularizer. The train and test accuracy are also improved substantially (Table 5.1). The improvement in the train accuracy can be observed from Figure 5.1 that exhibits

the decrease in the model's loss in each epoch. Also, for all the datasets, group and predictive fairness differences achieve the minimum loss of prediction compared to the other regularizers. Furthermore, in term of reducing fairness difference, from Figure 5.2 it is evident that models with the group, predictive and individual fairness regularizers ensure a higher level of fairness towards the record groups compared to L1 and L2 regularizers.

6
Conclusion & Future Work

Complete prevention against the inevitable disclosure of information from an ML model through the adversarial attacks is not achievable yet. But, an optimal model that leaks minimum information can be ensured by carefully adopting both the model and data properties. In this thesis, several selective data and model properties are analysed against MIA, to monitor how the attack accuracy is influenced by different property values. Considered data properties in our study are- the data size, balance in the classes, balance in the feature values, number of the features and data entropy. On the other hand, we explore model properties like the choice of the target model and model's hyperparameters, mutual information between the records and model parameters and overfitting. In addition to that, we also examine the impact of a model's fairness which is a reasonably new concepts in ML researches that shows model's integrity of prediction towards different groups of individuals. Furthermore, MIA-indistinguishability, referring to the equal predictability of a model for both member and non-member records, is also investigated against MIA accuracy. We explore whether models with higher fairness and MIA-indistinguishability offer better resistance against MIA.

The outcome of the investigation promisingly shows that both the data and model properties may considerably affect the membership inference from the model by the intruder. Our investigation shows that larger datasets with unbalanced classes or features are the most vulnerable towards MIA. On the other hand, choosing the right model with proper hyperparameter setting can reduce the vulnerability. Furthermore, an ML model with high learnability produces higher mutual information between the records and model's parameters which may increase the attack accuracy. Additionally, if a model gives more benefit to one group of individuals than the other in terms of correct prediction, the chances of a record to be exposed increase tremendously. Similarly, if a record is predicted differently by the model because of its presence in the training or test dataset, the probability of a successful attack capturing the difference surges. However, a few of the explored properties such as balance in the feature values, entropy and MIA-indistinguishability need further exploration to understand the obtained experimental results and realize their contribution to MIA's success.

In addition to assessing different data and model properties' contribution to the information leakage from an ML model through MIA, we further study how the observations can be utilised to strengthen a model, before allowing public access to it. We apply model properties such as

group, predictive and individual fairnesses and mutual information between the records and the parameters as regularizers in the model to control the model's ability to extract information from the features. All the four regularizers deemed to reduce the attack accuracy as well as model's training loss according to our experimental results. The results also demonstrate improved fairness of the model compared to the model without any regularizer and with other standard regularizations (L1, L2- norm).

6.1 Future Works

The outcomes of our study in the thesis motivate further exploration of the data and model properties to build an optimal ML model with marginal information disclosure. In this research, we only consider MIA as the adversarial attack, while other variants of black-box attacks such as model inversion attack [45] and attribute inference attacks [20] are yet to be studied. In addition, it is necessary to study the impact of other model specific properties for other variants of ML algorithms, such as re-inforced ML and agent-based modelling. Besides, existing defence mechanisms of MIA including the use of dropout layer [39], differential privacy based noise addition to the data and model parameters [1] and adversarial regularization [30] are not explored in our research. In addition, as the defences are model-specific, a further attempt can be made towards formulating a comprehensive defence against MIA that is not bounded by the type of the target model. Hence, the futuristic notion of this research could be described as a three-folded study:

- Further exploration of other data and model properties and their impact on different black-box and white-box adversarial attacks;
- Evaluating the effectiveness of adversarial attacks on the range of techniques for supervised and unsupervised learning in the case of both centralised and federated settings;
- Investigation on the practicality of the existing defences for the above-mentioned attacks in order to develop an optimal attack-resistant ML model and preferably model-independent defence mechanism.

Bibliography

[1] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 308–318.

[2] *Acquire Valued Shoppers Challenge*. https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data. Accessed: 2019-08-30. 2014.

[3] Michael Backes et al. “Membership privacy in MicroRNA-based studies”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 319–330.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. “Fairness in machine learning”. In: *NIPS Tutorial* (2017).

[5] Marco Barreno et al. “The security of machine learning”. In: *Machine Learning* 81.2 (2010), pp. 121–148.

[6] Reuben Binns. “Fairness in machine learning: Lessons from political philosophy”. In: *arXiv preprint arXiv:1712.03586* (2017).

[7] Dingfan Chen et al. “GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs”. In: *arXiv preprint arXiv:1909.03935* (2019).

[8] Cynthia Dwork et al. “Exposed! a survey of attacks on private data”. In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 61–84.

[9] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM. 2012, pp. 214–226.

[10] Cynthia Dwork et al. “Robust traceability from trace amounts”. In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 650–669.

[11] 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*. ACM. 2015, pp. 1322–1333.

[12] Pratik Gajane and Mykola Pechenizkiy. “On formalizing fairness in prediction with machine learning”. In: *arXiv preprint arXiv:1710.03184* (2017).

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[14] Jamie Hayes et al. “LOGAN: Membership inference attacks against generative models”. In: *Proceedings on Privacy Enhancing Technologies* 2019.1 (2019), pp. 133–152.

[15] Seira Hidano et al. “Model Inversion Attacks for Online Prediction Systems: Without Knowledge of Non-Sensitive Attributes”. In: *IEICE Transactions on Information and Systems* 101.11 (2018), pp. 2665–2676.

[16] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models". In: *Proceedings on Privacy Enhancing Technologies* 2019.4 (2019), pp. 232–249.

[17] Sorami Hisamoto, Matt Post, and Kevin Duh. "Membership Inference Attacks on Sequence-to-Sequence Models". In: *arXiv preprint arXiv:1904.05506* (2019).

[18] *Hospital Discharge Data Public Use Data File*. https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm. Accessed: 2019-08-30. 2019.

[19] Sandy Huang et al. "Adversarial attacks on neural network policies". In: *arXiv preprint arXiv:1702.02284* (2017).

[20] Jinyuan Jia and Neil Zhenqiang Gong. "AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning". In: *CoRR* abs/1805.04810 (2018). arXiv: 1805.04810. URL: http://arxiv.org/abs/1805.04810.

[21] Paul B. de Laat. "Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" In: *Philosophy & Technology* 31.4 (Dec. 2018), pp. 525–541. ISSN: 2210-5441. DOI: 10.1007/s13347-017-0293-z. URL: https://doi.org/10.1007/s13347-017-0293-z.

[22] Klas Leino and Matt Fredrikson. "Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference". In: *arXiv preprint arXiv:1906.11798* (2019).

[23] Yen-Chen Lin et al. "Tactics of adversarial attack on deep reinforcement learning agents". In: *arXiv preprint arXiv:1703.06748* (2017).

[24] Xiang Ling et al. "Deepsec: A uniform platform for security analysis of deep learning model". In: *IEEE S&P*. 2019.

[25] Gaoyang Liu et al. "SocInf: Membership Inference Attacks on Social Media Health Data With Machine Learning". In: *IEEE Transactions on Computational Social Systems* (2019).

[26] Yunhui Long et al. "Understanding membership inferences on well-generalized learning models". In: *arXiv preprint arXiv:1802.04889* (2018).

[27] Aleksander Madry et al. "Towards deep learning models resistant to adversarial attacks". In: *arXiv preprint arXiv:1706.06083* (2017).

[28] Frank McSherry and Ilya Mironov. "Differentially private recommender systems: Building privacy into the netflix prize contenders". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 627–636.

[29] Jan Hendrik Metzen et al. "On detecting adversarial perturbations". In: *arXiv preprint arXiv:1702.04267* (2017).

[30] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Machine learning with membership privacy using adversarial regularization". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2018, pp. 634–646.

[31] Nicolas Papernot et al. "The limitations of deep learning in adversarial settings". In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2016, pp. 372–387.

[32] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. “Knock knock, who’s there? Membership inference on aggregate location data”. In: *arXiv preprint arXiv:1708.06145* (2017).

[33] Md Atiqur Rahman et al. “Membership Inference Attack against Differentially Private Deep Learning Model.” In: *Transactions on Data Privacy* 11.1 (2018), pp. 61–79.

[34] Ahmed Salem et al. “Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models”. In: *arXiv preprint arXiv:1806.01246* (2018).

[35] Lea Schonherr et al. “Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding”. In: *arXiv preprint arXiv:1808.05665* (2018).

[36] Reza Shokri, Martin Strobel, and Yair Zick. “Privacy Risks of Explaining Machine Learning Models”. In: *arXiv preprint arXiv:1907.00164* (2019).

[37] Reza Shokri et al. “Membership inference attacks against machine learning models”. In: *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE. 2017, pp. 3–18.

[38] Lewis Smith and Yarin Gal. “Understanding measures of uncertainty for adversarial example detection”. In: *arXiv preprint arXiv:1803.08533* (2018).

[39] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[40] Yoshiki Takahashi, Masato Asahara, and Kazuyuki Shudo. “A framework for searching a predictive model”. In: *SysML Conference*. Vol. 2018. 2018.

[41] Florian Tramer et al. “Stealing machine learning models via prediction apis”. In: *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 2016, pp. 601–618.

[42] Florian Tramèr et al. “Ensemble adversarial training: Attacks and defenses”. In: *arXiv preprint arXiv:1705.07204* (2017).

[43] Stacey Truex et al. “Towards demystifying membership inference attacks”. In: *arXiv preprint arXiv:1807.09173* (2018).

[44] *UCI Machine Learning Repository: adult data set*. https://archive.ics.uci.edu/ml/datasets/Adult. Accessed: 2019-08-30. 1996.

[45] Michael Veale, Reuben Binns, and Lilian Edwards. “Algorithms that remember: model inversion attacks and data protection law”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018), p. 20180083.

[46] Sahil Verma and Julia Rubin. “Fairness definitions explained”. In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE. 2018, pp. 1–7.

[47] Mohammad Yaghini, Bogdan Kulynych, and Carmela Troncoso. “Disparate Vulnerability: on the Unfairness of Privacy Attacks Against Machine Learning”. In: *arXiv preprint arXiv:1906.00389* (2019).

[48] Samuel Yeom et al. “Privacy risk in machine learning: Analyzing the connection to overfitting”. In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE. 2018, pp. 268–282.

[49] Xiaoyong Yuan et al. “Adversarial examples: Attacks and defenses for deep learning”. In: *IEEE transactions on neural networks and learning systems* (2019).