USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

Disparate Privacy Vulnerability: Targeted Attribute Inference Attacks and Defenses

Ehsanul Kabir, Lucas Craig, and Shagufta Mehnaz, *Pennsylvania State University*

https://www.usenix.org/conference/usenixsecurity25/presentation/kabir

This paper is included in the Proceedings of the
34th USENIX Security Symposium.

August 13–15, 2025 • Seattle, WA, USA

978-1-939133-52-6

Open access to the Proceedings of the
34th USENIX Security Symposium is sponsored by USENIX.

Disparate Privacy Vulnerability: Targeted Attribute Inference Attacks and Defenses

Ehsanul Kabir
Pennsylvania State University

Lucas Craig
Pennsylvania State University

Shagufta Mehnaz
Pennsylvania State University

Abstract

As machine learning (ML) technologies become more prevalent in privacy-sensitive areas like healthcare and finance, eventually incorporating sensitive information in building data-driven algorithms, it is vital to scrutinize whether these data face any privacy leakage risks. One potential threat arises from an adversary querying trained models using the public, non-sensitive attributes of entities in the training data to infer their private, sensitive attributes, a technique known as the attribute inference attack. This attack is particularly deceptive because, while it may perform poorly in predicting sensitive attributes across the entire dataset, it excels at predicting the sensitive attributes of records from a few vulnerable groups, a phenomenon known as disparate vulnerability. This paper illustrates that an adversary can take advantage of this disparity to carry out a series of new attacks, showcasing a threat level beyond previous imagination. We first develop a novel inference attack called the disparity inference attack, which targets the identification of high-risk groups within the dataset. We then introduce two targeted variations of the attribute inference attack that can identify and exploit a vulnerable subset of the training data, marking the first instances of targeted attacks in this category, achieving significantly higher accuracy than untargeted versions. We are also the first to introduce a novel and effective disparity mitigation technique that simultaneously preserves model performance and prevents any risk of targeted attacks.

1 Introduction

Advancements in machine learning (ML) techniques have revolutionized the way data is analyzed and utilized, enabling the solution of complex problems and the development of a wide range of applications in many domains including privacy-sensitive ones, such as personalized healthcare [1, 25], finance [12, 16], and customer analytics [2, 11]. However, this technological leap has also launched a pivotal issue—the vulnerability of ML models to privacy attacks. Recent studies reveal that ML models are vulnerable to various privacy breaches. For instance, the models may reveal whether specific data was used in their training process [32], and even allow the deduction of confidential information from the training dataset [8, 17, 29, 38]. The second type of attack, namely, the model inversion attack, is particularly concerning as it enables adversaries to recover sensitive information from trained models. This weakness has constrained the training and application of ML models in domains sensitive to privacy, where safeguarding the privacy of the data is paramount.

Model inversion attacks can be broadly categorized into two types: class representative reconstruction [37, 39] and attribute inference [17, 29]. In class representative reconstruction, the adversary aims to construct representative data points for specific classes or categories of the training data. In attribute inference attacks, notably suited for models utilizing tabular data, the attacker aims to identify specific attribute values within the training data. It is especially disconcerting that tabular data, despite being the most widespread type of structured data, is far less investigated for vulnerabilities [15, 17, 23, 29] when compared with other kinds of data, e.g., images. This data domain faces a profound privacy challenge from attribute inference attacks, where adversaries can ascertain sensitive attributes using the model's predictions and the publicly accessible attributes of individuals whose data was leveraged for training.

Motivation. Existing attribute inference attacks struggle to achieve high performance, introducing uncertainty into their predictions and thereby reducing the perceived severity of privacy leakage through ML models. According to the evaluation by Jayaraman et al. [23], their performance is usually worse than that of an imputation attack, in which the attacker collects a small amount of auxiliary data (roughly 10% the size of the target data) and applies data imputation techniques to estimate the missing sensitive attribute values. This causes a further misjudgment of the threat level associated with attribute inference attacks. However, the presence of disparate vulnerability across various groups within the dataset [15, 29] leads us to believe that the performance assessed on the whole dataset does not truly represent the amount of privacy leakage,

as strong attack performance in some groups is offset by weak attack performance in others. Consequently, if the adversary becomes successful in identifying the subsets of data with high attack performance, they can target those subsets and predict the sensitive attribute values with greater certainty. Thus, we seek to answer the following research question: *can an adversary effectively determine the vulnerability levels of different groups within the dataset and then perform targeted attacks that are highly effective?*

Challenges. Identifying the dataset's most at-risk groups becomes a complex task under the assumption that the adversary lacks direct access to the training data and does not have a shadow dataset mirroring the training data's distribution, which is a realistic scenario in the context of model inversion attacks. However, we show that the variation in factors causing disparity such as the correlation between the sensitive and output attributes among different groups of the dataset can be leveraged to identify the most vulnerable groups of the dataset. To achieve this, the adversary needs to find a way to measure the variation in the factors causing disparity across different groups of the dataset. To this end, we introduce a novel technique that leverages the confidence score distribution of the model's predictions on various groups of the dataset to assess the variation in factors causing disparity across these groups. Our approach introduces a metric called *angular difference*, which can measure the vulnerability of a group. We also discover that angular difference is closely related to the correlation between the sensitive attribute and the output at the group level and can be used to estimate this correlation.

Proposed Attacks. We introduce a series of new attacks with the shared goal of exploiting the disparate vulnerability across groups. First, we propose an attack called *disparity inference attack*, which, to the best of our knowledge, is the first of its kind. This attack aims to rank the groups of records according to their vulnerability to attribute inference attacks. This attack can assist an adversary in launching existing attacks by assigning a degree of certainty to predictions based on group membership. Leveraging the disparity inference attack, we develop two novel targeted attribute inference attacks: single attribute-based and nested attribute-based, marking them as the first targeted attacks of their kind. Through empirical evaluation, we show that these targeted attacks can attain substantially higher performance in terms of accuracy than their untargeted counterparts [29] while necessitating far fewer queries to the target model.

The extensive and varied privacy leakage from exploiting disparate vulnerability brings forth the pressing question: *what steps can be taken to mitigate disparity?* Unfortunately, current defensive methods against attribute inference attacks are often ineffective or can even exacerbate disparity [15]. To address this, first, we explore the existing mutual information regularization (MIR [36]) defense and incorporate a disparity-aware objective into it, resulting in the Disparity-Aware Mutual Information Regularization (DAMIR) solution. We show that this disparity mitigation approach falls short in consistently achieving its goal. Consequently, we design a novel solution that mitigates disparity by balancing the contributing factors, which we term as *Balanced Correlation Defense (BCorr)*. Our evaluation shows that BCorr consistently mitigates disparity while also maintaining the original task performance of the target model.

Summary of contributions. Our work makes the following contributions:

❒ We present a novel attribute inference attack, termed the disparity inference attack, which aims to identify the most vulnerable groups in the training dataset. Our attack is the first to focus on identifying high-risk groups, and we show that our technique performs exceptionally well according to ranking similarity metrics.

❒ We are the first to explore targeted attacks within the attribute inference category, proposing two variations that focus on a vulnerable subset of the training data and achieve a significant performance boost over their untargeted counterparts.

❒ We introduce a novel disparity mitigation technique that effectively eliminates disparity between groups. At the same time, it preserves the target model's performance and prevents targeted attribute inference attacks.

2 Preliminaries

Attribute Inference Attack. Let $\mathbf{n}(x)$ denote the non-sensitive portion of a record x, and let $\mathcal{M}$ represent the target model. The objective of the attribute inference attack is to predict $s(x)$, the sensitive attribute value of x. Certain variations of the attack necessitate additional knowledge by the adversary, such as auxiliary data D_{aux}.

Confidence Score-based Model Inversion Attack (CSMIA). In this attribute inference technique introduced in [29], an adversary aims to predict the sensitive value of record x with class label y by querying the model multiple times with x_i where $n(x) = n(x_i)$ and $s(x_i) = s_i$, with s_i representing the i-th sensitive value. The model returns predictions y_i and confidence scores $conf_i$ for $i \in [1,k]$. If only one y_i matches y, the corresponding s_i is output. If multiple y_i match y, the one with the highest $conf_i$ is chosen and the corresponding s_i is output. Otherwise, the one with the lowest $conf_i$ is selected and its corresponding s_i is output.

Label Only Model Inversion Attack (LOMIA). In the attribute inference technique introduced by [29], the adversary generates predictions for x_i similar to CSMIA but does not use confidence scores. They create an attack dataset from all x that returned a true prediction for only one x_i, adding (x,y) as input and s_i as output. Subsequently, an attack model is trained and used to infer the sensitive attribute value on the remaining records.

Imputation Attack. The adversary creates an attack dataset similar to LOMIA, but using D_{aux}. Subsequently, an attack

model is trained to infer the sensitive attribute value on the records.

Neuron Importance Attack. In this whitebox attack introduced in [23], D_{aux} is utilized to identify the top 10 most correlated neurons within the MLP. For each record x, the weighted sum of the activation values of these top 10 neurons is calculated. If this sum exceeds a certain threshold, the attacker predicts that x has the sensitive value of interest.

Disparate Vulnerability of MIAI. Let, $\mathcal{M}$ denote the MIAI attack model, $\mathbb{D}$ denote the target dataset, and $\mathcal{A}$ denote the attack algorithm that the adversary aims to launch. Additionally let, $ASR(\mathcal{M},\mathbb{D},\mathcal{A})$ denote the attack success rate of launching $\mathcal{A}$ on model $\mathcal{M}$ and $\mathbb{D}$. We state that $\mathcal{A}$ is *disparate* if there exists two disjoint subsets $\mathbb{D}_1$ and $\mathbb{D}_2$ of $\mathbb{D}$ such that $|ASR(\mathcal{M},\mathbb{D}_1,\mathcal{A}) - ASR(\mathcal{M},\mathbb{D}_2,\mathcal{A})| > \varepsilon$ for some $\varepsilon > 0$. In other words, the attack success rate of $\mathcal{A}$ on $\mathbb{D}$ is not uniform across all subsets of $\mathbb{D}$. Here, ε is a threshold below which any disparity is considered negligible.

3 Attack Threat Model

We assume the following adversary capabilities:

❏ Access to the black-box target model, i.e., the adversary can query the model with x and obtain the output label y and the corresponding confidence scores.

❏ Full knowledge of the non-sensitive attributes.

❏ Knowledge of every possible value of the sensitive attribute and any non-sensitive attributes the adversary regards as group attributes.

These capabilities are considered realistic in the context of model inversion attacks, with most current model inversion attacks assuming at least these capabilities. Notably, the attacks introduced in Yeom et al [38], Fredrikson et al [17] and CSMIA require full non-sensitive attribute knowledge for the specific target record x, whereas LOMIA needs complete non-sensitive attribute information for the entire target dataset. Unlike in previous works [17, 38], the adversary in this case does not need to be aware of the marginal priors, defined as the relative frequencies of the sensitive attribute values, to conduct the attack. In addition, the adversary can perform the attack without needing an auxiliary dataset. This differs from most existing attribute inference attacks [23, 38], which typically require an auxiliary dataset that matches the distribution of the target dataset, except for CSMIA and LOMIA [29]. The attacker is considered to have complete knowledge of all possible non-sensitive attribute values. This assumption is realistic because publicly queryable ML models often reveal all possible values of query attributes. Leveraging this information and our proposed targeted attribute inference attacks, the attacker can identify groups based on different non-sensitive attributes. When performing attacks, we assume the adversary has fewer capabilities to investigate the extent of privacy leakage under practical constraints. Conversely, during the evaluation of our defense, we consider an adversary with greater capabilities to rigorously evaluate the defense's strength, as outlined in the threat model in section 7.2.

4 Uncovering High-Risk Groups

4.1 Key Factor Contributing to Vulnerability

. To identify groups with a high risk of privacy leakage, it is essential to understand the factors contributing to the differing vulnerability levels among records from high-risk groups and those from low-risk groups. The core factor contributing to the vulnerability of ML models to attribute inference attacks is the association between input and output data. For a model to accurately predict outputs during inference, it must learn the associations present in the training data. Therefore, a strong association between input and output data in the training set increases the likelihood of model inversion attacks, where an adversary uses the model to infer sensitive attribute values in the training data. One basic way to measure the association between two variables is through correlation, leading to the natural assumption that the correlation between the sensitive attribute and the output plays a crucial role in the vulnerability to attribute inference attacks. By 'correlation', we mean Pearson's correlation, which is often used interchangeably in this domain. We conduct a simple experiment to provide evidence supporting this hypothesis.

Experiment Setup. We use the Census19 and Texas-100X datasets (detailed in section 6.1) for this experiment and apply a sampling technique (described in section 6.1) to create 19 training sets from each dataset, each with varying levels of correlation. We then train target models on each dataset and perform CSMIA [29], LOMIA [29], Imputation [23], and NeuronImportance [23] attacks on these models. For the latter two attacks, we assume the adversary utilizes a single auxiliary dataset across all scenarios, with a distribution identical to that of the original datasets.

Results. Figure 1 reports the accuracy and F1 scores of the attacks. The plot offers several compelling observations. In the Census-19 dataset, both CSMIA and LOMIA exhibit a monotonically increasing trend within the correlation range of -0.2 to -0.9, a pattern not observed in the Imputation and NeuronImportance attacks. A similar trend is seen in the Texas-100X dataset for correlation values ranging from 0.3 to 0.9. The main factor behind the poor performance of Imputation and NeuronImportance at higher correlation magnitudes is that their auxiliary data does not have the same high correlation as the training data. The performance of an imputation attack varies greatly depending on whether the auxiliary data shares the same distribution as the original data, an unrealistic scenario, or has a different distribution, which is more likely in practice. This distinction is discussed thoroughly in Section 6.2. Interestingly, correlation's effect is sign-agnostic; whether negative or positive, high correlations lead to increased vulnerability. This occurs because different

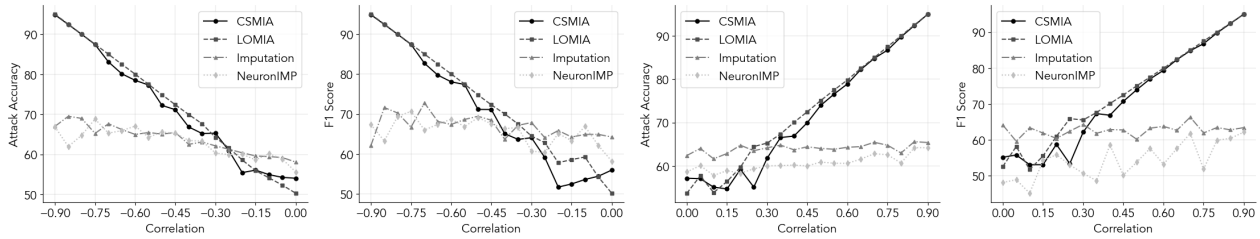


Figure 1: Comparative evaluation of CSMIA, LOMIA, Imputation Attack, and NeuronImportance Attack across scenarios where dataset have varying level of correlation ranging from 0 to -0.9 for Census19 (a-b) and 0.9 for Texas-100X (c-d)

labeling methods can alter the correlation sign. An alternative labeling method (e.g., switching positive and negative outputs) can convert a previously negative correlation to a positive one. The key takeaway from the results is that correlation is a significant factor in vulnerability to attribute inference attacks, prompting the question: does correlation also influence disparate vulnerabilities among groups? We explore this with a brief experiment. Throughout the rest of this paper, we describe correlation as high/low to refer to its magnitude, avoiding redundancy and improving readability. Similarly, 'correlation' is used as shorthand to specifically refer to the correlation between the sensitive attribute and the output.

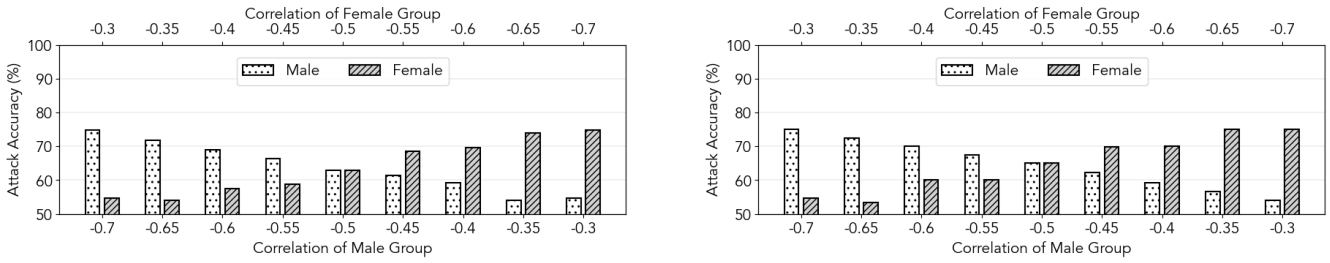
Impact of Correlation on Disparate Vulnerability. We conduct another brief experiment to assess the impact of correlation on disparate vulnerability. This experiment is carried out on the Census-19 dataset, using the `SEX` attribute to divide the dataset into Male and Female groups. We explore 9 scenarios, progressively increasing the correlation in Female records and decreasing it in Male records. The experimental results are depicted in Figure 2. The results clearly show that the impact of correlation on vulnerability is significant at a group level. When the correlation is high in Male records and low in Female records, the attack performance is high in the Male group and low in the Female group, and vice versa. Both CSMIA and LOMIA performances exhibit this trend. To emphasize, the experimental results demonstrate that correlation is a key factor influencing the varying vulnerabilities among groups within a dataset.

Given this observation, the key question is—*Can an adversary having only black-box access to the model compare the correlation among these groups and thus identify the groups with high privacy risk?* Note that precise measurement of correlation is not mandatory; being able to compare correlation between groups is sufficient. However, comparing correlation among groups poses a challenge since the adversary in our threat model lacks access to an auxiliary dataset that matches the distribution of the training data. To the best of our knowledge, no method exists to estimate correlation in the training data, let alone compare correlation among groups.

4.2 Comparing Correlation between Groups

A high degree of correlation in the dataset indicates that certain values of the sensitive attribute, or certain ranges of values if the sensitive attribute is non-discrete, are more frequently associated with a specific class label compared to other values of the sensitive attribute. Conversely, a low correlation implies that the relative difference in the occurrence of sensitive attributes for a particular class label is minor or insignificant. Our primary intuition is that during training, the target model is influenced by the relative frequency of sensitive attributes within a class label, leading it to predict with higher confidence for a record from that label when the sensitive attribute is set to the most frequent value, rather than to less common values. In other words, the confidence score gap, defined as the difference in confidence scores generated by querying the same record with different sensitive attribute values, depends on the relative frequency of those values. Since the level of correlation determines relative frequency, the distribution of the aforementioned confidence gap serves as an indicator of the correlation level, which could be used by an adversary. First, we will consider a dataset with a binary sensitive attribute and a binary output class to present our argument more clearly. Afterwards, we will discuss how the argument can be extended to multi-category or non-discrete sensitive attributes and multi-class outputs. Let's assume the sensitive attribute values are 'yes' and 'no', and the output class values are 'True' and 'False'. If the dataset has a strong correlation, the 'True' class will contain significantly more 'yes' records than 'no' records. As a result of this imbalance, the model will develop a bias towards assigning higher confidence to the 'True' class for queries with the sensitive attribute set to 'yes' compared to 'no' for the same record. If the records in a group exhibit a high degree of imbalance, the expected confidence gap will be large; conversely, a low degree of imbalance will result in a smaller expected confidence gap. Thus, if an adversary analyzes the distribution of the confidence gap across multiple groups, they can predict which groups are more vulnerable.

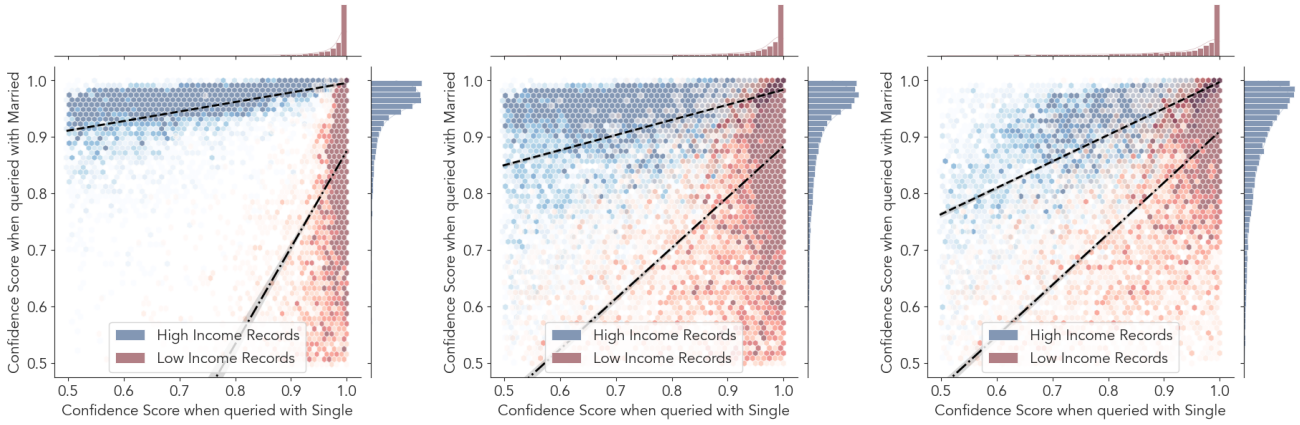
To validate our intuition, we use the scenario from the previous section where the correlation for the Male group



(a) CSMIA strategy

(b) LOMIA strategy

Figure 2: Correlation vs. Attack performance for Male and Female group for 9 different scenarios using Census-19 Dataset.



(a) Group Correlation = -0.6

(b) Group Correlation = -0.5

(c) Group Correlation = -0.4

Figure 3: Histograms (bivariate and univariate) of confidence scores generated by querying records with different sensitive values, taken from groups with varying correlation levels in the Census19 dataset.

is -0.6 and for the Female group is -0.4, and then we plot the confidence scores generated by querying the same record with different sensitive attribute values and present it in Figure 3. We plot the confidence scores for the records where all predictions from various queries matched with the correct label. In particular, we use hexagon bins to plot bivariate histograms on a 2D plane and univariate histograms on the axis margins to study the distributions of the confidence scores. The Census19 dataset features a binary output class with labels 'High Income' and 'Low Income.' The sensitive attribute `MAR`, representing the marital status of the individual in the record, has two possible values: 'Married' and 'Single.' In addition to 'Male' and 'Female,' we include the 'White' group, which consists of records where the `RACE` attribute is set to 'White,' with a correlation of -0.5. High Income and Low Income records are plotted in different colors to illustrate the differences in their confidence score distributions. The plot presents several intriguing insights. Each distribution forms a comet-like shape with a concentrated head at the top right and a tail extending away from that point. The direction of the tail's trajectory differs between High Income records and Low Income records. The trajectory features support our hypothesis; High Income records show more confidence with the sensitive attribute set to 'Married' than 'Single', resulting in a horizontal slant of the trajectory. The tail trajectories of the distributions for High Income records across groups are tilted at different angles. This occurs due to the differences in the relative frequency of sensitive attribute values, as discussed in the previous paragraph. A similar characteristic is observed in Low Income records, except that their tail trajectories are vertically slanted. After drawing regression lines on each set of confidence scores, the gradual shift in angle toward the center from high correlation to low correlation groups becomes even more pronounced. Note that the regression line is slightly slanted relative to the tail's trajectory, which occurs due to the high density of points in the top right region for High (Low) Income records. The difference in angles between the regression lines for High Income and Low Income records also diminishes gradually from high to low correlation groups. We term this the *angular difference*, which we argue an adversary could exploit to understand and compare the correlation between groups.

The histograms clearly show that the level of correlation affects the distribution of confidence scores when records are queried with different sensitive attribute values. An adversary could attempt to compare correlations between groups by computing the difference in confidence score distributions for records from different class labels. This method, however, is prone to errors because the high density shared between the distributions within the 0.95-1 confidence score range skews the calculation of differences. We propose that computing the angular difference for each group and using it to compare correlations would be more effective, as the regression lines reflect the tail points in the distributions, which are the key differentiators.

Extending to Multi-Valued/Non-Discrete Sensitive Attributes and Multi-Class Outputs. We draw regression lines in an n-dimensional space for multi-valued sensitive attributes, where n denotes the number of possible sensitive attribute values. For multi-class outputs, we generate regression lines for records from each class label and determine the average angular difference between all pairs of lines. We propose following Mehnaz et al. [29]'s approach for non-discrete sensitive attributes by binning the value ranges and using the set of bin means as a substitute for the sensitive attribute values.

5 Attack Methodology

In this section, we outline the details of the two newly proposed classes of attacks: disparity inference attack and targeted attribute inference attack. In the disparity inference attack, the adversary uses angular difference to rank groups based on attack vulnerability. In the targeted attribute inference attack, the adversary uses angular difference to strategically identify target subsets with high vulnerability and then launches attribute inference attacks on them. Since both types of attacks rely on computing angular difference on groups of records, we will begin by detailing this technique.

Definition 5.1 (Confidence Matrix)**.** *For a set of training data $\mathbb{D}$ containing n records, a trained model $\mathcal{M}$, the confidence matrix $\mathcal{C}$ is of dimension $n \times |\mathcal{S}|$ and defined as follows -*

$$\mathcal{C} = \Big[\big[Pr(\mathcal{M}(x')) : x' \in \mathcal{T}(x)\big]^T \quad \forall\, x \in \mathbb{D}\Big]^T$$

where $\mathcal{S}$ denotes the set of all sensitive attribute values and $\mathcal{T}(x)$ denotes the set of records generated by varying the sensitive attribute value of record x.

Definition 5.2 (Angular Difference)**.** *Given a training dataset $\mathbb{D}$ with n records and a confidence matrix $\mathcal{C}$ generated by querying target model $\mathcal{M}$ with $\mathbb{D}$, and with $\mathbb{Y}$ representing the set of output labels, the angular difference is defined as the average difference in angle between all pairs of lines L_{y_1} and L_{y_2}, where $y_1, y_2 \in \mathbb{Y}$. L_y denotes the regression line fitted through $|\mathcal{S}|$-dimensional points from $\mathcal{C}_y$, the submatrix of $\mathcal{C}$ that contains rows corresponding to records with label y.*

Algorithm 1 Confidence Matrix Generation

Input: $\mathcal{M}, \mathcal{N}(\mathbb{D})$ ▷ $\mathcal{M}$ is the target model and $\mathcal{N}(\mathbb{D})$ is the non-sensitive portion of dataset $\mathbb{D}$
Output: $\mathcal{C}, \mathbf{t}$ ▷ $\mathcal{C}$ is the confidence matrix of the dataset $\mathbb{D}$ and $\mathbf{t}$ is a boolean vector

1: **for** each $(\mathbf{n}(x), y)$ in $\mathcal{N}(\mathbb{D})$ **do**
2: $\quad\mathcal{X} \leftarrow \{x' : \mathbf{n}(x) = \mathbf{n}(x') \wedge s(x') \in \mathcal{S}\}$
3: $\quad\mathcal{C}[x] \leftarrow (Pr(\mathcal{M}(x')) : x' \in \mathcal{X})$
4: $\quad\mathbb{Y}' \leftarrow \{M(x') : x' \in \mathcal{X}\}$ ▷ Set of output predicted on $\mathcal{X}$
5: $\quad$**if** $\mathbb{Y}' = \{y\}$ **then**
6: $\quad\quad\mathbf{t}[x] \leftarrow$ `true` ▷ If all predictions were correct
7: $\quad$**else**
8: $\quad\quad\mathbf{t}[x] \leftarrow$ `false`
9: $\quad$**end if**
10: **end for**
11: **return** $\mathcal{C}, \mathbf{t}$

Algorithm 2 Angular Difference Computation

Input: $\mathcal{M}, \mathcal{N}(\mathbb{D}), \mathcal{C}$,t ▷ $\mathcal{M}$ is the target model, $\mathcal{N}(\mathbb{D})$ is the non-sensitive portion of dataset $\mathbb{D}$, $\mathcal{C}$ is the confidence matrix, and $\mathbf{t}$ is the prediction correctness vector
Output: Δ is the angular difference of the subset of data corresponding to group i

1: $C_y \leftarrow \varnothing \quad \forall y \in \mathbb{Y}$
2: **for** each $(\mathbf{n}(x), y)$ in $\mathcal{N}(\mathbb{D})$ **do**
3: $\quad$**if** $\mathbf{t}[x] =$ `true` **then**
4: $\quad\quad C_y \leftarrow C_y \cup \{\mathcal{C}[x]\}$
5: $\quad$**end if**
6: **end for**
7: $\mathcal{L} \leftarrow \{regression\_fit(C_y) : y \in \mathbb{Y}\}$
8: $\Delta \leftarrow \underset{\substack{L_1, L_2 \in \mathcal{L} \\ L_1 \neq L_2}}{\text{Mean}} (angle(L_1, L_2))$
9: **return** Δ

5.1 Computing Angular Difference

The process of computing angular difference involves generating confidence scores from records by querying the target model while altering the sensitive attribute values. We call this collection of confidence scores the *confidence matrix*, which is formally defined in Definition 5.1. Angular difference is formally defined in Definition 7.1. Algorithm 1 outlines the steps for generating the confidence matrix and Algorithm 2 outlines the steps for computing angular difference from the confidence matrix. Initially, a set of records is generated by varying the sensitive attribute value for each original record, and the target model is queried with these sets (Lines 1-3, Algorithm 1). Subsequently, the predictions are recorded, and a boolean vector $\mathbf{t}$ tracks whether all predictions returned were correct for that record (Lines 4-8, Algorithm 1). For computing angular difference, we consider records where predictions are correct for any sensitive attribute value. We hypothesize that for these records, the differences in confidence scores with varying sensitive values are highly indicative of the correlation level. We refer to $\mathbf{t}$ as the prediction correctness vector.

Afterward, for each class label, a collection of confidence scores is created from records in that class label with a prediction correctness value of `true` (Lines 1-6, Algorithm 2). A regression line is then fitted to each set of confidence scores, using the dimension associated with the positive sensitive attribute value as the output. The angular difference is defined as the mean distance in angles between these lines (Lines 7-9, Algorithm 2).

To enhance readability and simplify notation, $\mathbb{D}$ will refer to the non-sensitive portion of the dataset for the remainder of this section, except when defining attack objectives, where $\mathcal{N}(\mathbb{D})$ will be used for correctness.

5.2 Disparity Inference Attack

Objective. Let $\mathcal{M}$ be a target model trained on dataset $\mathcal{N}(\mathbb{D})$, which can be divided into k non-overlapping subsets $\mathcal{N}(\mathbb{D}_1), \mathcal{N}(\mathbb{D}_2), \ldots, \mathcal{N}(\mathbb{D}_k)$. The success rate of an attack $\mathcal{A}$ on target model $\mathcal{M}$ using dataset $\mathbb{D}$ is denoted by $ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}), \mathcal{A})$, with $\mathcal{N}(\mathbb{D})$ indicating the non-sensitive part of the data accessible to the adversary. The attack vulnerability ranking $\mathcal{R} = (r_1, r_2, \ldots, r_k)$ of the groups corresponding to subsets $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_k$ is defined such that:

$$\{r_1, r_2, \ldots, r_k\} = [1,k]$$
$$ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}_{r_i}), \mathcal{A}) \geq ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}_{r_j}), \mathcal{A})$$
$$\forall\, 1 \leq i < j \leq k$$

The attacker's goal is to find $\mathcal{R}$ or a ranking very close to $\mathcal{R}$. It is important to consider that the attacker does not know the true sensitive attribute values and therefore cannot determine the exact $ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}_i), \mathcal{A})$ for all $i \in [1,k]$. Otherwise, the attacker's goal would be trivial.

Attack Steps. Query $\mathcal{M}$ using Algorithm 1 to generate the confidence matrix $\mathcal{C}$ for $\mathbb{D}$. Next, use Algorithm 2 to calculate the angular difference Δ_i for each subset $\mathbb{D}_i$. Rank the indices $1, 2, \ldots, k$ by decreasing Δ_i values and output this as the desired ranking.

5.3 Targeted Attribute Inference Attack

Objective. Let $\mathcal{M}$ be a target model trained on dataset $\mathbb{D}$. The attack success rate of an attack $\mathcal{A}$ on target model $\mathcal{M}$ using dataset $\mathbb{D}$ is denoted by $ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}), \mathcal{A})$ where $\mathcal{N}(\mathbb{D})$ denotes the non-sensitive portion of the data available to the adversary. The objective of the adversary is to find a dataset $\mathbb{D}_{target} \subset \mathbb{D}$ satisfying the following conditions:

$$\left| \frac{|\mathbb{D}_{target}|}{|\mathbb{D}|} - \kappa \right| < \varepsilon \tag{1}$$

$$ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}_{\text{target}}), \mathcal{A}) \geq ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}'), \mathcal{A})$$
$$\forall\, \mathbb{D}' \in \{\mathbb{D}' \subset \mathbb{D} \mid |\mathbb{D}'| > |\mathbb{D}_{\text{target}}|\} \tag{2}$$

where κ is the attack budget controlling the target dataset size within 0 to 0.5. The value of ε is set to a very small amount to ensure that the size of the target dataset meets the attack budget. Condition 2 ensures that the adversary performs better in the target subset than in any subset of equal or greater size. The degree of disparate vulnerability determines how much performance improvement the adversary can obtain through a targeted attack versus an untargeted attack.

Key Insights. It is computationally intractable to evaluate attack success rate on all possible $\mathbb{D}'$ to find a $\mathbb{D}_{target}$ that satisfies Condition 2. However, we can constrain our exploration to the subsets of $\mathbb{D}'$ defined by restricting one or more of non-sensitive attributes of the records to a subset of their respective possible values. Our next challenge is finding a method to compare attack success rates between two subsets without using an auxiliary dataset, since our threat model assumes the adversary lacks access to any additional data. Similar to our previous attacks, we propose utilizing the strong connection between angular difference and attack success rate on a subset. Hence, the adversary's goal shifts to finding the subset that has the greatest angular difference. Although constraining the number of subsets to explore makes the adversary's objective feasible, it is still exponentially costly to naively compute angular difference on the constrained set of subsets. Therefore, we propose two innovative strategies to optimally explore the subset space which are outlined in detail in the sections that follow.

Single Attribute-based Targeted Attack

1. Randomly sample $\mathbb{D}^q$ from $\mathbb{D}$ with the condition: $|\mathbb{D}^q| = q \times |\mathbb{D}|$ where q is the query budget to limit the number of queries made to the model.

2. Use Algorithm 1 to query $\mathcal{M}$ and generate confidence matrix $\mathcal{C}$ on $\mathbb{D}^q$.

3. For each non-sensitive attribute a, split $\mathbb{D}^q$ into subsets $\mathbb{D}^q_1, \mathbb{D}^q_2, \ldots, \mathbb{D}^q_k$ such that $\mathbb{D}^q_i = \{x \in \mathbb{D}^q \mid a(x) = \mathcal{A}_i\}$ for all $i \in [1,k]$ with $\mathcal{A}$ representing a set of size k that contains possible values of a and $a(x)$ denoting the value assigned to attribute a in record x. Compute the angular difference on each subset $\mathbb{D}^q_i$ and put them in the vector Δ^q_a. Define, $\text{range}(\Delta^q_a) = \max(\Delta^q_a) - \min(\Delta^q_a)$.

4. Find the attribute a with the highest $\text{range}(\Delta^q_a)$. Let, $\mathbb{D}_i = \{x \in \mathbb{D} \mid a(x) = \mathcal{A}_i\}$ and Δ_i be the angular difference on subset $\mathbb{D}^q_i$ that we computed on the previous step for all $i \in [1,k]$.

5. Rank indices $1, 2, \ldots, k$ based on the value of Δ_i in increasing order. Let, $\mathbb{I}$ denote this ordered set of indices and $\mathbb{I}_{[m,n]}$ denote the ordered subset of $\mathbb{I}$ starting at index m and ending at index n.

6. Let, $\mathbb{D}_{[1,m]} = \bigcup_{i \in \mathbb{I}_{[m,n]}} \mathbb{D}_i$ for all $m \in [1,k]$. Output $\mathbb{D}_{[1,m]}$ that satisfies Condition 1.

The initial four steps identify the most suitable single attribute for exploring the subset space. Intuitively, the attribute with the widest range of angular differences in its subsets is likely to contain the most vulnerable subsets among those defined by a single attribute. Then the attacker can rank the vulnerability of subsets defined by the selected attribute by ranking corresponding angular difference values (step 5) and aggregating the most vulnerable subsets in decreasing order of angular difference until the attack budget is reached (step 6).

Nested Attribute-based Targeted Attack

1. Set the depth or the number of nested attributes d to $\lceil \log_2(\kappa) \rceil$.

2. Follow steps 1 to 3 exactly as specified in the single attribute-based targeted attack (section 5.3) to find the range of angular differences for each attribute.

3. Select the attributes $a_1, a_2, .., a_d$ that are among the top-d in terms of the largest ranges of angular differences. Let $\mathbb{D}^i_j$ denote the set of records with the attribute a_i set to the j-th value from $\mathcal{A}_i$.

4. For each attribute $a_i \in \{a_1, a_2, .., a_d\}$, define $\mathbb{I}^i$ as the ordered set of indices ranked on angular differences of groups defined by attribute a_i. Let $\mathbb{I}^i_m$ denote the top-m groups defined by attribute a_i in terms of their angular difference.

5. Let the *above-average-risk segment* denote the set of the most vulnerable groups based on their angular differences, ensuring they comprise close to half of the total records in all groups defined by the attribute a_i. For each attribute $a_i \in \{a_1, a_2, .., a_d\}$, define $\mathbb{I}^i_{1/2}$ as the indices within $\mathbb{I}^I$ that identify the groups constituting the above-average-risk segment. Formally, $\mathbb{I}^i_{1/2} = \underset{m \in [1, |\mathcal{A}_i|]}{argmin} \left| \left| \cup_{j \in \mathbb{I}^i_m} \mathbb{D}^i_j \right| / |\mathbb{D}| - 0.5 \right|$. Find above-average-risk segment, $\mathbb{D}^i_{1/2} = \cup_{j \in \mathbb{I}^i_{1/2}} \mathbb{D}^i_j$, for each attribute $a_i \in \{a_1, a_2, .., a_{d-1}\}$.

6. Let, $\mathbb{D}_m = \mathbb{D}^1_{1/2} \bigcap \ldots \bigcap \mathbb{D}^{d-1}_{1/2} \bigcap \left(\bigcup_{i \in \mathbb{I}^d_m} \mathbb{D}^d_i \right)$ for m in $[1, |\mathcal{A}_d|]$. Output $\mathbb{D}_m$ with the lowest value of m that satisfies Condition 1.

In this strategy, we examine the subset space made up of nested groups, which are intersections of groups defined by different attributes. We achieve this by combining the above-average-risk segment from each attribute and forming nested groups through their intersections. To comply with the attack budget, we cap the number of attributes considered at d. This greedy approach is adopted due to the exponential computational complexity of considering every combination of nested groups. Steps 1 to 3 involve identifying the best d attributes in terms of the range of angular differences. The attack budget may not permit selecting the full above-average-risk segment from the last chosen attribute; therefore, for that attribute, we select as many groups as possible while satisfying Condition 1 as shown in step 6.

6 Experiments

In this section, we explain our experimental arrangement, datasets, machine learning models, and performance metrics. We then examine the performance of our proposed attacks.

6.1 Experimental Setup

Datasets. We use the following three datasets in our experiments: Census19 [7], Texas-100X [30], and Adult [5]. More details of the datasets are presented in Appendix C.1.

Sampling Technique. Both Census19 and Texas-100X datasets contain around a million records each, making them ideal candidates for selective sampling to manage dataset-specific variables, including the relationship between sensitive and output attributes. In contrast, earlier studies [17, 29] utilizing datasets like Adult [5] and GSS [3] struggled with limited data sizes, barely sufficient for training a model that could achieve reliable accuracy on a test set and could not conduct detailed sampling. In our experimental evaluation, we incorporate a sampling technique that allows us to set a predefined correlation between the sensitive and output attributes not only for the full training data but also for specific groups. The specifics of this sampling technique are detailed step-by-step below.

1. Let n denote the number of desired samples for the training data or a specific group. Additionally, let m denote the ratio between the number of samples with negative sensitive values and the number of samples with positive sensitive values, and let c denote the desired correlation between the sensitive attribute and the output.

2. Let n^+_+ and n^-_+ denote the number of samples to be picked that have positive sensitive values with positive and negative output values, respectively. Similarly let, n^+_- and n^-_- denote the number of samples to be picked that have negative sensitive values with positive and negative output values, respectively. Simply put, the subscript denotes the sensitive value and the superscript denotes the output value. $n^+_- = \lfloor \frac{\sqrt{m} \times (\sqrt{m} - c) \times n}{2 \times (m+1)} \rfloor$, $n^-_- = \lfloor \frac{\sqrt{m} \times (\sqrt{m} + c) \times n}{2 \times (m+1)} \rfloor$ $n^-_+ = \lceil \frac{n}{2} - n^-_- \rceil$, $n^+_+ = \lceil \frac{n}{2} - n^+_- \rceil$

The sampling method used above ensures that the desired correlation is achieved while maintaining a balanced number of positive and negative output samples. The correctness proof of our claim is provided in Appendix B.1. For all experiments on Census19 and Texas-100X, the value of m is set to 1 to align with the sensitive attribute distribution of the original dataset. This sampling method enables precise control

over attribute correlations, which enhances our capacity to identify significant trends or effects on attack efficacy, while still reflecting real-life data—something that prior research methodologies have not been able to achieve. In addition, we experiment with the Adult dataset without using controlled sampling to create an even more realistic setting.

Model Training. For our experiments with both Texas-100X and Census19 datasets, we select 50,000 records to create the training set for training the neural network model. Additionally, we randomly sample another 50,000 records from the remaining data to form the test set, ensuring that the training and test sets are mutually exclusive. For the Adult dataset, we utilize the full dataset, splitting it into training and test data as done in [29]. We use Scikit-learn's [31] implementation of Multi-layer Perceptron (MLP) as our default class of machine learning model. The architecture and training hyperparameter details are put into Appendix C.2.

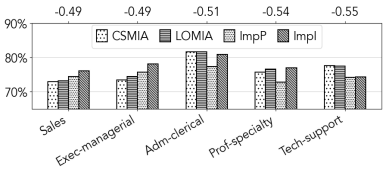
6.2 Ideal vs. Practical Imputation Attacks

An ideal imputation attack is characterized by the adversary having an auxiliary dataset that precisely matches the distribution of the target data. This includes having similar distribution properties, such as correlation and marginal priors, for any group of records. However, it is impractical to assume that an adversary, who is typically external to the organization owning the private training data, would be able to acquire an auxiliary dataset that accurately mirrors the distribution at a granular level. Thus, any dataset obtained by a realistic adversary would differ in distributional properties, either at a macro-level (calculated across the whole dataset) or at a micro-level (calculated for specific subsets within the dataset). An imputation attack conducted with such auxiliary data is termed a practical imputation attack. In this section, we perform two experiments on Adult dataset to evaluate the performance of practical imputation attacks against ideal imputation attacks, highlighting the substantial gap between them. The first experiment explores dataset-level distributional drift in the auxiliary dataset by altering the marginal prior relative to the training dataset. The second experiment investigates group-level distributional drift in the auxiliary dataset, maintaining the same dataset-level correlation as the training data but with different group-level correlations.

Dataset-level Distributional Drift. In this experiment, we vary the marginal prior η of the auxiliary dataset, which represents the fraction of positive samples, from 0.5 to 0.1, thus deviating it further from the training data's marginal prior of 0.52. We also consider auxiliary datasets of various sizes, ranging from 5000 to 100. Figure 4(a) presents the performance of imputation attacks for every combination of auxiliary datasets across the two analyzed dimensions. The results demonstrate a significant performance decline with increased deviation in η. Notably, for η values of 0.1 and 0.2, the imputation attack's performance falls below that of CSMIA (69.97)

η	$\lvert D_{aux}\rvert$ 100	500	1000	5000
0.1	47.96	54.41	60.15	62.54
0.2	57.79	68.95	66.93	67.68
0.3	64.16	71.50	70.58	71.57
0.4	71.47	71.98	72.89	73.31
0.5	68.84	73.43	71.82	73.46

(a)



(b)

Figure 4: (a) Imputation attack performance across 2 dimensions: Auxiliary Dataset Size (row) and Marginal Prior, η (column). (b) Imputation Attacks (Practical – ImpP, Ideal – ImpI) versus AI attacks (CSMIA, LOMIA) in Occupation Groups with high correlation. Bottom x-axis labels indicate group Names; Top x-axis labels indicate Group Correlation Values; Y-axis denotes Attack Accuracy.

and LOMIA (70.61) regardless of auxiliary dataset size. Imputation attacks outperform CSMIA and LOMIA only when the auxiliary dataset's marginal prior is close to the original. This implies that *practical imputation attacks may only achieve high performance when conditions are nearly ideal. Otherwise, they are likely to perform worse than existing AI attacks.*

Group-level Distributional Drift. In this experiment, we use an auxiliary dataset where the correlation for each occupation group is -0.44, which matches the overall correlation of -0.4412 in the training data. However, correlations within the groups of the original training data vary significantly, ranging from -0.17 to -0.55, suggesting a group-level distributional drift in the auxiliary data. We perform an imputation attack with this auxiliary dataset, an ideal imputation attack, LOMIA, and CSMIA, and display the attack success rate of all attacks for groups with higher-than-overall correlation in Figure 4(b). According to the results, CSMIA and LOMIA outperform the practical imputation attack in the top 3 out of 5 vulnerable groups. By contrast, ideal imputation attack performance is very similar to CSMIA and LOMIA in 2 out of these 3 groups. Nevertheless, the results indicate that *a practical imputation attack with group-level distributional drift is likely to perform poorly in highly vulnerable groups compared to attribute inference (AI) attacks.*

The results of the experiments above suggest that practical imputation attacks, the only type that a realistic adversary can execute, are likely to underperform relative to AI attacks despite having access to an auxiliary dataset that the AI attacks do not assume. This implies that AI attacks on ML models enable adversaries to more accurately predict sensitive attributes compared to practical imputation attacks, signaling privacy leakage from these models. In short, even if AI attacks do not surpass ideal imputation attacks, their superiority over practical imputation attacks demonstrates privacy leakage from ML models. Thus, practical imputation attacks should be viewed as a baseline for AI attack evaluation, whereas ideal impu-

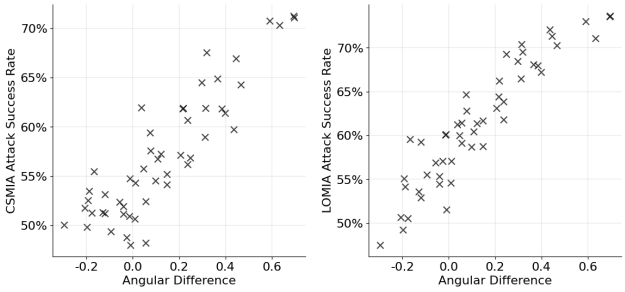


Figure 5: Angular difference vs. attack performance (CSMIA - left, LOMIA - right) of 51 states from Census19 dataset. Accuracy is used as attack performance metric.

tation attacks can provide a useful benchmark for assessing privacy leakage. In the sections that follow, we first evaluate the effectiveness of our Disparity Inference Attack and then we investigate the extent of privacy leakage resulting from the two targeted AI attack types.

6.3 Disparity Inference Attack Performance

Sensitive Attribute and Group Attribute Selection. We select the `MAR` column, denoting marital status, as the sensitive attribute for the Census-19 dataset, and for the Texas-100X dataset, we pick `SEX_CODE` as the sensitive attribute to be inferred. For the Census-19 dataset, we employ the `State` attribute to divide the training data into 51 groups and for the Texas-100X dataset, we use the `PAT_STATUS` attribute to divide the training data into 10 groups. Since attribute inference attack performance is significantly influenced by the correlation between the sensitive attribute and output, we assign different correlation values to each of the 51 groups to achieve varying levels of vulnerability. Each state is indexed from 0 to 50, and the desired correlation between the sensitive and output attributes is set to $-0.01 \times i$ for records from the state with index i. Whenever possible, 1,000 records are sampled per group, or fewer if fewer are available, while preserving the specified correlation. Similarly for Texas-100X, each group, defined by a specific value of PAT_STATUS, is assigned a desired correlation value from $0, 0.05, 0.10, \ldots, 0.45$. Afterward, we train target models on each subset of records.

Evaluation. We launch our proposed disparity inference attack, which computes angular differences for each group and ranks them based on that. Figure 5 plots the angular difference in X-axis and the actual attack performance in Y-axis for both CSMIA (left plot) and LOMIA (right plot). The results clearly demonstrate a strong association between the angular difference and the success rate of the attack. Specifically, a higher angular difference for a group indicates that both attacks will likely perform well, whereas a lower angular difference suggests that the performance will be poor. *Therefore, an adversary can compute angular differences on multiple groups and use that as an indicator of the quality of attack predictions made on the records of the groups.*

To evaluate the ranking quality of our disparity inference attack, we use two statistical metrics: Kendall's Tau [26] and Spearman's Rank Correlation [33]. These two metrics, known for their robustness against outliers and ability to minimize the impact of extreme values, have been extensively applied across numerous scientific domains. The metrics range from -1 to 1, however, ranking performing close to 0 is considered poor while ranking performance further from 0 (close to either -1 or 1) is considered good. That is because a ranking performing negatively implies the order is in reverse and may not necessarily be lacking in utility. Furthermore, we consider the following baseline vulnerability ranking attack for comparison – we assume that the attacker possesses an auxiliary dataset that is the same size as the training dataset in which the sensitive attribute values of all records are known. To ensure that the auxiliary dataset and the original dataset have different distributions, we randomly sample from the full versions of Census19 and Texas-100X to create the auxiliary dataset. The attacker launches CSMIA on the auxiliary dataset and evaluates attack performance on all the groups and ranks them based on the attack performance on the auxiliary data. Table 1 presents the comparative evaluation of ranking quality between our proposed disparity inference attack and the baseline attack we considered. The results show that the vulnerability ranking by our disparity inference attack is far superior to that of the baseline attack. *This not only establishes that the task of disparity inference is not trivial but also our proposed approach is very effective in ranking the groups in terms of their vulnerability.* The extremely low p-values corresponding to the null hypothesis indicate that the closeness between the rankings is not due to chance and the null hypothesis can be rejected.

6.4 Targeted Attribute Inference Attack

We adopt the same configuration as the disparity inference attack for our evaluation. Alongside Census19 and Texas-100X, we also perform targeted attribute inference (AI) experiments on the full Adult dataset to illustrate the real-world effects of the attacks. The adversary follows the steps outlined in Section 5.3, using either CSMIA or LOMIA as the underlying attribute inference algorithm. As baselines, we consider two variations of the imputation attack: ImpI and ImpP. ImpI uses an auxiliary set that matches the distribution of the original data, while ImpP uses an auxiliary set with a different distribution. For both methods, the adversary applies an imputation attack on the auxiliary data to determine the attack success rate across various groups. Using the success rate as a substitute for angular difference, the adversary then mimics the steps of our proposed targeted attacks to initiate a targeted imputation attack. Once the target subset is chosen, the

Census19				
	Kendall Tau		Spearman R	
	CSMIA	LOMIA	CSMIA	LOMIA
Disparity Inference Attack	0.6914 (1.39e-12)	0.7579 (8.30e-15)	0.8767 (7.23e-17)	0.9104 (5.06e-20)
Baseline	-0.0759 (4.37e-01)	-0.0931 (3.40e-01)	-0.1225 (3.97e-01)	-0.1275 (3.77e-01)

Texas-100X				
	Kendall Tau		Spearman R	
	CSMIA	LOMIA	CSMIA	LOMIA
Disparity Inference Attack	-0.7778 (9.46e-04)	-0.7778 (9.46e-04)	-0.9273 (1.12e-04)	-0.9152 (2.04e-04)
Baseline	-0.2444 (3.81e-01)	-0.2444 (3.81e-01)	-0.3576 (3.10e-01)	-0.3939 (2.60e-01)

Table 1: Comparative evaluation of the ranking quality of disparity inference attacks versus baseline attacks employing an auxiliary dataset. The values inside the parentheses are p-values corresponding to the null hypothesis.

Census19							
κ	1	0.75	0.5	0.375	0.25	0.1	0.05
CSMIA	62.56	64.73	67.43	69.02	70.42	72.54	73.27
LOMIA	61.24	63.71	67.56	69.45	70.82	72.92	73.78
ImpI	65.38	65.65	66.10	65.91	65.71	67.30	66.85
ImpP	62.99	63.30	63.57	62.96	64.12	64.17	64.25

Texas-100X					
κ	1	0.75	0.5	0.25	0.1
CSMIA	60.95	62.82	64.39	61.00	62.82
LOMIA	61.50	63.90	66.72	64.68	69.68
ImpI	59.33	63.95	65.01	65.57	66.75
ImpP	51.64	47.55	46.73	46.26	48.97

Adult					
κ	1	0.75	0.5	0.25	0.1
CSMIA	69.96	69.14	72.37	74.83	81.61
LOMIA	70.61	69.79	73.36	74.86	81.68
ImpI	74.46	76.96	77.72	79.82	81.52
ImpP	65.05	68.31	69.70	63.94	65.21

Table 2: Attack success rate of single attribute-based targeted inference attack compared with targeted imputation baselines. Accuracy is used as the metric for attack success rate.

adversary carries out an imputation attack on it.

Single Attribute-based Targeted AI. Table 2 presents the evaluation results of the single attribute-based targeted inference attack. To evaluate the attack, we use various κ values from 1 to 0.1. For Census19 specifically, we set κ as low as 0.05, which is feasible because the grouping attribute has 51 unique values. $\kappa = 1$ denotes the untargeted version. For Census19, Texas-100X, and Adult, we choose `STATE`, `PAT_STATUS`, and `Occupation` respectively as the grouping attributes, since these exhibit the greatest range in angular difference. The results show that targeted attacks using both CSMIA and LOMIA consistently improve in accuracy as κ is reduced, demonstrating a clear trend across nearly all cases. The CSMIA variant targeted attack results in performance increases of 17.12% for Census19, 5.65% for Texas-100X, and 16.66% for Adult, compared to the untargeted counterpart. For the LOMIA variant, the performance gains are 20.48%, 13.31%, and 15.68%, respectively.

Nested Attribute-based Targeted AI. Table 3 presents the evaluation results of the nested attribute-based targeted inference attack. The top-d attribute ordered set selected for Census19, Texas-100X, and Adult are {`STATE`, `SCHOOL`, `RACE`, `SEX`}, {`PAT_STATUS`, `RACE`, `ADMITTING_DIAGNOSIS`, `TYPE_OF_ADMISSION`, `SOURCE_OF_ADMISSION`}, {`Occupation`, `Work`, `Race`, `Sex`} respectively. Similar to single attribute-based attacks, nested attribute-based targeted AI shows increased attack performance with a smaller attack budget, i.e., increased depth of attributes. For the Adult dataset, the attack success rate increases up to 86.77%, and for the Texas-100X dataset, it reaches 100% at a depth of 5 for both CSMIA and LOMIA variants.

Comparison with Baselines. The pattern of improved performance as κ decreases is not observed for ImpP, implying that when an adversary's auxiliary dataset differs in distribution, a targeted imputation attack will not outperform an untargeted attack. ImpI demonstrates improved performance with decreasing κ values, though at a slower rate than our proposed attack, suggesting a correlation between group-level attack success rates on ImpI's auxiliary dataset and the original data. Given that obtaining auxiliary dataset with the same distribution as the private training data is impractical, our proposed targeted AI attacks remain the most feasible option for adversary with minimal knowledge and capabilities.

Effect of MLP Architecture. Targeted AI attacks are launched on MLPs with varying depths (with hidden layers ranging from 2 to 4) trained on Census19 to examine whether the complexity of ML models influences their susceptibility to AI attacks. Table 4 displays the experimental results. The findings indicate that both single and nested variations show similar performance across MLPs with different numbers of layers, suggesting that the architecture of MLPs does not influence vulnerability to AI attacks.

7 Potential Mitigation Strategies

In this section, we examine potential strategies to mitigate disparate vulnerabilities that an adversary could exploit to launch various types of attacks, as demonstrated in the previous section. Existing literature [23, 36] provides techniques to defend against attribute inference attacks, but none propose methods to mitigate disparity between groups. Dibbo et al. [15] apply defense techniques tailored to mitigate disparity in other domains [27] or to defend against attribute inference attacks without attempting to address disparity [36], finding limited success in mitigating disparity against attribute inference attacks. We investigate two defensive strategies: the first adapts the mutual information regularization (MIR) technique [36] to address disparity, referred to as disparity-aware mutual information regularization (DAMIR), while the second is a novel method focusing on balancing correlation across the dataset, termed as the balanced correlation defense (BCorr).

7.1 Disparity-Aware Mutual Information Regularization (DAMIR)

The mutual information regularization method [36] incorporates a secondary loss minimization objective during training to reduce the mutual information between sensitive attribute

Census19					
κ	1 (0)	0.5 (1)	0.375 (2)	0.25 (3)	0.1 (4)
CSMIA	62.56	67.43	67.43	69.70	69.36
LOMIA	61.24	67.56	66.96	68.20	70.26
ImpI	65.38	66.06	65.98	66.17	66.85
ImpP	62.99	63.57	63.42	63.60	60.86

Texas-100X						
κ	1 (0)	0.5 (1)	0.25 (2)	0.1 (3)	0.05 (4)	0.01 (5)
CSMIA	60.95	64.39	61.77	66.12	67.77	100.00
LOMIA	61.50	66.72	63.64	64.85	66.19	100.00
ImpI	59.33	63.97	63.78	72.02	72.67	78.42
ImpP	51.64	46.73	48.11	45.38	48.40	43.15

Adult					
κ	1 (0)	0.5 (1)	0.375 (2)	0.25 (3)	0.1 (4)
CSMIA	69.97	71.18	73.72	77.57	86.74
LOMIA	70.61	72.19	73.90	73.90	86.77
ImpI	74.46	77.72	76.47	78.16	77.86
ImpP	65.05	69.70	73.59	58.89	68.65

Table 3: Attack performance of nested attribute-based targeted inference attacks compared with targeted imputation baselines. The values inside parentheses denote the number of nested attributes considered. Accuracy is used as the attack performance metric.

Number of MLP Layers →		4		3		2	
	κ →	1	0.05	1	0.05	1	0.05
Single Attribute-based	CSMIA	63.82	72.42	62.56	73.27	60.73	71.85
	LOMIA	61.63	73.02	61.24	73.78	60.79	73.66
	κ →	1	0.1	1	0.1	1	0.1
Nested Attribute-based	CSMIA	63.82	68.66	62.56	69.36	60.73	68.96
	LOMIA	61.63	70.75	61.24	70.26	60.79	70.56

Table 4: Attack success rate of targeted inference attacks across MLP models of varying depths.

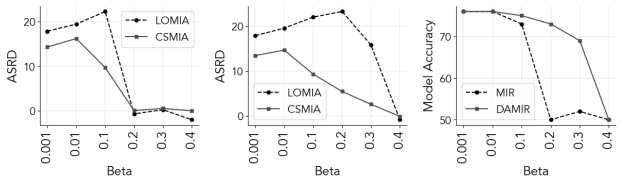


Figure 6: ASRD of Mutual Information Regularization defense (MIR - left, DAMIR - middle) under CSMIA and LOMIA and target model accuracy of MIR and DAMIR trained models (right).

and output. A hyperparameter β adjusts the weight of the secondary loss compared to the primary loss, to adjust the strength of the mutual information regularization. Although it successfully decreases the performance of untargeted attribute inference attacks, it does not mitigate disparity and may even worsen it, as shown in the evaluation results in Figure 7 of [15]. To overcome this limitation, we make a key adjustment: mutual information loss is calculated solely on records from the vulnerable group, ensuring that mutual information between sensitive attribute and output is minimized only for that group's records. We refer to this revised defense strategy as disparity-aware mutual information regularization (DAMIR).

Evaluation. We apply both MIR and DAMIR during the training of a subset of the Census-19 dataset where the correlation of Male and Female group is -0.4 and -0.1 respectively. The range of β, from 0.001 to 0.4, is used to alter the intensity of the mutual information regularization. Figure 6 shows the attack success rate difference (ASRD) (defined in section 7.2) between the Male and Female groups and the model accuracy of both MIR and DAMIR trained models. The plots indicate that MIR can only reduce disparity when β is set to a high value, resulting in a significant loss of utility in the target model. DAMIR, in contrast, exhibits slightly better performance in CSMIA, being able to reduce disparity to a degree without substantial degradation of model utility. Nevertheless, it can only achieve full disparity mitigation with a considerable loss of model utility. Furthermore, in LOMIA, the performance is worse, where reducing disparity appears to be impossible without a significant loss of model utility. Therefore, mutual information regularization is ineffective in reducing disparity.

7.2 Balanced Correlation Defense (BCorr)

This defense strategy focuses on addressing disparities by ensuring similar correlation levels across all relevant groups in the dataset, thereby eliminating group-specific differences contributing to the disparity. The main rationale for this defense is that disparity stems from differences in correlation between groups, and thus, mitigating disparity requires eliminating these differences.

Objective. The objective of BCorr is to mitigate disparate vulnerability by lowering attack performance on more vulnerable groups. In particular, BCorr aims to reduce the metric ASRD (Attack Success Rate Difference) as defined below.

Definition 7.1 (ASRD)**.** *Given a target model $\mathcal{M}$ trained with $\mathbb{D}$ and $\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_k$ denoting the set of records belonging to groups defined by a non-sensitive grouping attribute a, ASRD is defined as: $ASRD(\mathcal{M}, \mathbb{D}, a, \mathcal{A}) = \max_{i \in [1,k]} ASR_i - \min_{j \in [1,k]} ASR_j$ where $\mathcal{A}$ denotes the attack algorithm and $ASR_i = ASR(\mathcal{M}, \mathcal{N}(\mathbb{D}_i), \mathcal{A})$.*

Defense Threat Model. To achieve the objective of the defense, we make the following assumptions: the defender has access to the full dataset and the trained target model, and operates as a single entity with complete control over the dataset and model training. Additionally, the defender is aware of which groups are more vulnerable to attacks. This assumption is practical because the defender can simulate attack scenarios using the full dataset and target model to identify vulnerable subgroups. Moreover, the defender can compute correlations between sensitive attributes and outputs for each subgroup, leveraging these correlations as strong indicators of attack vulnerability. These capabilities enable the defender to effectively identify and mitigate disparities in vulnerability.

Design. BCorr comprises of the following steps.

- Given non-sensitive grouping attribute *a* such that some groups are more vulnerable than others, the first step

is to rank groups defined by a in terms of their correlation, $correlation(\mathcal{S}(\mathbb{D}_i)), \mathbb{Y}(\mathbb{D}_i))$, where $\mathbb{D}_i \subset \mathbb{D}$ denotes records from group i. Suppose, m is the index of the group with the least correlation which is c_m.

- Sample records from each $\mathbb{D}_i$ to form $\mathbb{D}'_i$ such that the correlation of $\mathbb{D}'_i$ is equal to c_m. Note that, $\mathbb{D}_m$ does not need to be sampled and therefore, $\mathbb{D}'_m = \mathbb{D}_m$. Let, $\mathbb{D}' = \mathbb{D}'_1 \cup \mathbb{D}'_2 \cup \cdots \cup \mathbb{D}'_k$. Train model $\mathcal{M}'$ on $\mathbb{D}'$.

Evaluation. We examine the effectiveness of the balanced correlation defense with the Census-19 and Texas-100X datasets, specifically targeting the binary attributes `SEX` and `SEX_CODE`, and the multi-valued attributes `STATE` and `PAT_STATUS` for each dataset, respectively. We also apply a Fairness Constraint-based defense (FC), as used in [27], to mitigate disparate vulnerability in Membership Inference attacks. For this baseline, we use the Exponentiated Gradient algorithm with the Equalized Odds [22] constraint. The results of this experiment are presented in Table 5. We measure group vulnerability by the difference in attack success rates between the most and least vulnerable groups which is referred to as ASRD. Accuracy is used as the metric for ASRD. To evaluate group fairness, we use Equalized Odds Difference (EOD) [10] and Demographic Parity Difference (DPD) [10]. Model accuracy (MA) is reported at the group level for binary attribute scenarios and across the entire test dataset for multi-valued attribute scenarios. BCorr can completely mitigate disparities between Male and Female groups in binary attribute scenarios for both datasets, achieving this without sacrificing model utility or group fairness despite using 66.67% of the original training data in the balanced correlation set. FC, in contrast, fails to effectively reduce disparity in either binary attribute scenarios, although it preserves group fairness. For multi-valued attribute scenarios, BCorr can substantially reduce disparity and lower the ASR of the most vulnerable group from 73.8% to 62.92% for Census19 and from 72.59% to 59.07% for Texas-100X under CSMIA. It is worth noting that reducing ASRD between 51 groups is much harder than between 2 groups, yet BCorr still achieves a substantial reduction. The FC evaluation for the multi-valued attribute case was omitted because it failed in the binary case, and the computational cost was deemed too high. BCorr effectively mitigates disparity regardless of MLP depth used in training, as it addresses the root cause of vulnerability at the dataset level, as shown in Table 6.

How/Why Bcorr Works. Suppose, occupation=nurse and gender=female might constitute the most vulnerable groups within their respective categories in a particular scenario. Female individuals may be more vulnerable than males because a large proportion of them belong to the highly vulnerable 'nurse' occupation compared to other occupations like teacher or salesperson. In contrast, there are significantly fewer males in the 'nurse' occupation, resulting in a lower overall vulnerability for males. The assumed scenario is similar to that of the UC Berkeley Admissions Rate Bias study [6]. Their study found that women faced higher rejection rates overall but were often favored at the department level. This paradox stemmed from women applying disproportionately to competitive departments. In our case, the higher privacy vulnerability of female individuals is driven by their significant representation in 'nurse' occupation, a highly vulnerable group. Through our targeted attack approach, an attacker could determine that 'nurse' is the most vulnerable occupation and 'female' is the most vulnerable gender. In our proposed defense, the defender can similarly identify privacy disparities in gender and occupation groups. Depending on the severity of the observed disparities or the defender's priorities, they can apply BCorr to either attribute (gender/occupation) to mitigate the disparate vulnerabilities effectively. The severity of disparity can be measured through ASRD. Suppose the disparity across the occupation groups is more severe than that of gender groups and the defender applies BCorr on the occupation attribute. While lowering the attack vulnerability of occupation-based groups, our sampling-based defense approach is unlikely to increase the attack vulnerability of the gender=female group. Specifically, reducing the correlation between the sensitive attribute and model output within the occupation=nurse group—the core mechanism of BCorr—is also likely to reduce the correlation within the female records of that group, resulting in lower attack vulnerability for the gender=female group.

Comparison with Fairness. The existing work in fairness [10, 13, 14] primarily focuses on ensuring equal model performance across groups. In contrast, BCorr aims to equalize the success rates of attribute inference attacks across groups while simultaneously preserving model performance fairness. Due to this distinction, fairness metrics like equalized opportunity or equalized odds, often used to measure model fairness, are not directly applicable to evaluate BCorr's ability to mitigate attack disparity. To address this, we introduce ASRD, which parallels fairness metrics like equalized odds difference but measures disparities in attacker success rates across groups rather than disparities in model performance. Theoretical guarantees for attack disparity mitigation are inherently challenging due to the use of non-linear DNNs and the varying strategies employed by the attacker. Any bound on DNNs must be derived through approximations of their non-linear behavior, which often fail to capture intricate interactions between layers and parameters, resulting in imprecise bounds. Moreover, the variability in attribute inference attack strategies adds another layer of complexity to establishing reliable bounds on attack performance disparity. Nevertheless, we provide empirical evidence that BCorr effectively mitigates disparate vulnerability.

8 Related Works

Foundational Works. Fredrikson et al. initially introduced model inversion attacks for linear regression models in [18]

Census19

Defense	SEX						STATE				
	ASRD		EOD	DPD	MA		ASRD		EOD	DPD	MA
	CSMIA	LOMIA			Male	Female	CSMIA	LOMIA			
None	12.52	14.97	0.0674	0.0178	73.21	78.16	22.94	26.75	0.4	0.2401	73.19
BCorr	2.06	3.59	0.0416	0.0015	73.89	77.90	10.93	7.58	0.4	0.1923	74.02
FC	11.45	11.90	0.0221	0.0144	70.72	70.30	-	-	-	-	-

Texas-100X

Defense	SEX_CODE						PAT_STATUS				
	ASRD		EOD	DPD	MA		ASRD		EOD	DPD	MA
	CSMIA	LOMIA			Male	Female	CSMIA	LOMIA			
None	12.12	14.45	0.0332	0.0177	72.53	74.08	17.27	20.54	0.3707	0.2176	75.11
BCorr	0.94	2.48	0.0172	0.0186	74.70	74.35	7.34	4.26	0.3519	0.1982	75.19
FC	11.45	13.28	0.02176	0.0020	71.788	68.108	-	-	-	-	-

Table 5: Comparison of group vulnerability (ASRD), group fairness (EOD, DPD), and group-level model utility (MA) between models trained with and without BCorr Defense.

Census19

	MLP Depth	ASRD		EOD	DPD	MA	
		CSMIA	LOMIA			Male	Female
BCorr	2-layer	0.45	3.05	0.0494	0.0171	75.312	78.532
	3-layer	2.06	3.59	0.0416	0.0015	73.89	77.90
	4-layer	1.45	4.54	0.0353	0.0030	72.39	75.62

Table 6: BCorr performance across varying MLP models.

and extended these attacks to non-linear ML models in [17]. This subsequent work defined two major types of model inversion attacks: attribute inference, in which the adversary tries to uncover the sensitive attributes in the dataset used to train the model, and class representative reconstruction, where the goal is to reconstruct instances resembling those in the training data. Mehnaz et al. provide the first evidence of disparity in attribute inference attacks [29].

Notable Works in Attribute Inference. Existing works on attribute inference attacks have considered a wide range of threat models, with most assuming strong attacker capabilities, rendering their proposed attacks impractical. In a series of works [20, 21, 24], Gong et al. and Jia et al. applied attribute inference attacks in social media scenarios, where an adversary infers private attributes of a user based on their public information, but these attacks depend on users who also disclose their private attributes publicly, limiting their applicability to cases where private-public attribute pairs can be collected for an auxiliary set of data matching the distribution of the target data. Yeom et al. [38] investigate how the influence, defined as the extent to which changes in a sensitive attribute affect predictions, impacts vulnerability to attribute inference attacks. Their analysis reveals that while attacker advantage grows with initial increases in influence, it actually decreases as the influence becomes more significant. Conversely, our research identifies the correlation between the sensitive attribute and the output as a consistent factor contributing to the vulnerability of ML models. Wang et al. [36] propose a mutual information regularization method to defend against model inversion attacks, including attribute inference attacks. Mehnaz et al. [29] introduce CSMIA and LOMIA, with the latter being the first to show that attribute inference attacks can be launched on models without confidence scores, proving the inadequacy of defenses that mask these scores. Jayaraman et al. [23] provide evidence that existing attribute inference attacks perform worse compared to an imputation attack using auxiliary data without even querying the target model. Our evaluation in section 4, however, demonstrates that when there is a high correlation in the private training data, existing attacks surpass the imputation attack in performance. The study by Tramer et al. [35] reveals that attribute inference attacks perform better if the adversary can poison the training data. However, the assumption that training data can be poisoned is applicable in crowdsourced or collaborative scenarios; the former is unrealistic for private training data, while the latter falls outside the scope of our work.

Other Related Works. Dibbo et al. [15] investigate various potential contributors to disparity but find no consistent factors. While their work focuses on understanding disparity in attribute inference attacks, it does not identify specific contributing factors. Our work goes further by identifying correlation as a key factor driving this disparity. Kulynych et al. [27] proposes a defense to mitigate disparate vulnerability in membership inference attacks. In contrast, we tackle attribute inference attacks. We also include [24]'s fairness constraint-based (FC) defense as a baseline in our experiments and demonstrate its ineffectiveness in mitigating disparate vulnerability for attribute inference attacks. Zhong et al. [40] propose a defense against disparate vulnerability in link inference attacks within GNNs, but their technique cannot be transferred to attribute inference attacks in tabular data. Property inference attacks, closely related to correlation estimation attacks, were first introduced in [4] and later applied to deep learning models in [19]. However, existing attacks [9, 28, 34] focus on inferring properties such as group size rather than the correlation between sensitive attributes and outputs, which is our focus.

9 Conclusion

In this paper, we present a series of novel attacks that expose significant privacy leakage in ML models trained on tabular data, a risk that has often been underestimated and overlooked. Our comprehensive evaluation reveals that adversaries can identify high-risk groups within records, posing an alarming threat to individuals in those groups. Our findings underscore that privacy leakage through targeted attribute inference attacks is far from trivial, with adversaries capable of making highly accurate predictions. To address these significant privacy concerns, we introduce a novel defense that have been demonstrated to be effective through evaluation. Looking forward, we are committed to incorporating advanced privacy-preserving techniques into our methods to provide a comprehensive toolkit for creating safe and secure data-driven algorithms.

Ethics Consideration

In evaluating the ethical implications of our research, we considered both deontological and consequentialist perspectives. From a deontological standpoint, our work respects individuals' rights to privacy by avoiding the use of identifiable data and adhering to the principle of minimizing harm. Our research utilizes publicly available datasets (Census19 [7], Texas-100X [30], and Adult [5]) to evaluate the vulnerability of machine learning models to attribute inference attacks. All data used in our experiments is de-identified, meaning that no direct identifiers are associated with any individual's private information. We have adhered strictly to the data use policies outlined by the data providers and ensured that all experimental results are presented in aggregate form to avoid re-identification risks.

From a consequentialist perspective, our research seeks to maximize benefits by improving the understanding of machine learning vulnerabilities and contributing to the development of stronger defenses against privacy attacks. We have weighed the potential harms of publicizing new attacks against the benefits of increased awareness and preparedness within the security community. We believe that the positive outcomes, particularly the advancement of privacy-preserving technologies and methodologies, outweigh the risks associated with our findings. Moreover, our inclusion of mitigation strategies provides actionable insights to help practitioners secure their models against these new forms of attack.

Compliance with the Open Science Policy

We fully support the principles of open science as outlined in the USENIX Security 2025 Open Science Policy. We have released the full codebase, including all scripts and binaries necessary for replicating our experiments, via a public repository: https://zenodo.org/records/14732956. However, due to licensing restrictions associated with the Texas-100X dataset, we are unable to provide direct access to this specific dataset. For the other datasets used in our research, we have shared them freely alongside our codebase. For the Texas-100X dataset, we provide detailed instructions on how to access it directly from the original provider, along with scripts to preprocess the data to match the format used in our experiments. This approach ensures compliance with the dataset's licensing agreements while still promoting the principles of reproducibility and transparency.

References

[1] Farhad Ahamed and Farnaz Farid. Applying internet of things and machine-learning for personalized healthcare: Issues and challenges. In *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, pages 19–21. IEEE, 2018.

[2] Omer Artun and Dominique Levin. *Predictive marketing: Easy ways every marketer can use customer analytics and big data*. John Wiley & Sons, 2015.

[3] NORC at the University of Chicago. Gss general social survey | norc — gss.norc.org. https://gss.norc.org/, 2024. [Accessed 29-05-2024].

[4] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

[5] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[6] Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

[7] US Census Bureau. Pums data— census.gov. https://www.census.gov/programs-surveys/acs/microdata/access.2019.html#list-tab-735824205, 2019. [Accessed 29-05-2024].

[8] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[9] Harsh Chaudhari, John Abascal, Alina Oprea, Matthew Jagielski, Florian Tramer, and Jonathan Ullman. Snap: Efficient extraction of private properties with poisoning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 400–417. IEEE, 2023.

[10] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *Advances in neural information processing systems*, 33:15088–15099, 2020.

[11] Ping Chou, Howard Hao-Chun Chuang, Yen-Chun Chou, and Ting-Peng Liang. Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. *European Journal of Operational Research*, 296(2):635–651, 2022.

[12] Robert Culkin and Sanjiv R Das. Machine learning in finance: the case of deep learning for option pricing. *Journal of Investment Management*, 15(4):92–100, 2017.

[13] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315, 2019.

[14] Saswat Das, Marco Romanelli, and Ferdinando Fioretto. Disparate impact on group accuracy of linearization for private inference. *arXiv preprint arXiv:2402.03629*, 2024.

[15] Sayanton V Dibbo, Dae Lim Chung, and Shagufta Mehnaz. Model inversion attack with least information and an in-depth analysis of its disparate vulnerability. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 119–135. IEEE, 2023.

[16] Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in finance*, volume 1170. Springer, 2020.

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

[18] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 17–32, San Diego, CA, August 2014. USENIX Association.

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

[20] Neil Zhenqiang Gong and Bin Liu. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 979–995, Austin, TX, August 2016. USENIX Association.

[21] Neil Zhenqiang Gong and Bin Liu. Attribute inference attacks in online social networks. *ACM Trans. Priv. Secur.*, 21(1), January 2018.

[22] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[23] Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? *arXiv preprint arXiv:2209.01292*, 2022.

[24] Jinyuan Jia, Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. Attriinfer: Inferring user attributes in online social networks using markov random fields. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1561–1569, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[25] Balaram Yadav Kasula. Harnessing machine learning for personalized patient care. *Transactions on Latest Trends in Artificial Intelligence*, 4(4), 2023.

[26] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[27] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies*, 1:460–480, 2022.

[28] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. Property inference from poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1120–1137. IEEE, 2022.

[29] Shagufta Mehnaz, Sayanton V Dibbo, Roberta De Viti, Ehsanul Kabir, Björn B Brandenburg, Stefan Mangard, Ninghui Li, Elisa Bertino, Michael Backes, Emiliano De Cristofaro, et al. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4579–4596, 2022.

[30] Texas Department of State Health Services. Texas Inpatient Public Use Data File (PUDF) | Texas DSHS — dshs.texas.gov. https://www.dshs.texas.gov/texas-health-care-information-collection/health-data-researcher-information/texas-inpatient-public-use, 2006. [Accessed 29-05-2024].

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

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

[33] C Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 1904.

[34] Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *Proceedings on Privacy Enhancing Technologies*, 2022.

[35] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792, 2022.

[36] Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11666–11673, 2021.

[37] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

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

[39] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 253–261, 2020.

[40] Da Zhong, Ruotong Yu, Kun Wu, Xiuling Wang, Jun Xu, and Wendy Hui Wang. Disparate vulnerability in link inference attacks against graph neural networks. *Proceedings on Privacy Enhancing Technologies*, 2023.

A Overview of Notations

Symbol	Description
$\mathbb{D}$	The original training dataset
$\mathcal{M}$	Target model trained on $\mathbb{D}$
$Pr(\mathcal{M}(.))$	Confidence score output by $\mathcal{M}$
$\mathcal{A}$	Attack Algorithm
$s(.)$	Sensitive attribute value of a record
$\mathbf{n}(.)$	Vector of non-sensitive attribute values of a record
$\mathcal{S}$	Set of possible values of the sensitive attribute
$\mathcal{N}(\mathbb{D})$	Non-sensitive portion of dataset $\mathbb{D}$
$\mathcal{C}$	Confidence matrix
ρ	Correlation

Table 7: Notations used and their descriptions

B Methodology

B.1 Correctness Proof of Sampling Technique

Let, s and y denote the sensitive attribute and output respectively and n denote the total number of records. Then, according to definition,

$$c = \frac{n\sum sy - (\sum s)(\sum y)}{\sqrt{(n\sum s^2 - n(\sum s)^2)\,(n\sum y^2 - n(\sum y)^2)}} \tag{3}$$

Now, $\sum y$ essentially means the sum of output values of n records. Since, the output can either take 0 or 1 and takes 1 a total of $n_+^+ + n_-^+$ times, $\sum y = n_+^+ + n_-^+$ and $\sum y^2 = n_+^+ \times 1^2 + n_-^+ \times 1^2 = n_+^+ + n_-^+$. Similarly, $\sum s^2 = \sum s = n_+^+ + n_+^-$. Substituting these values in equation 3 and performing a few algebraic operations we get,

$$c = \frac{n_+^+ \times n_-^- - n_+^- \times n_-^+}{\sqrt{(n_+^+ + n_+^-)(n_+^+ + n_-^+)(n_-^+ + n_-^-)(n_+^- + n_-^-)}} \tag{4}$$

According to our requirement, the number of positive samples $(n_+^+ + n_-^+)$ is equal to the number of negative samples $(n_+^- + n_-^-)$ and the ratio between positive and negative samples $(\frac{n_+^+ + n_+^-}{n_-^+ + n_-^-})$ is m. If we combine $n_+^+ + n_+^- + n_-^+ + n_-^- = n$ with the above we get,

$$n_+^+ + n_-^+ = n_+^- + n_-^- = \frac{n}{2}$$
$$n_+^+ + n_+^- = \frac{n}{m+1}, \quad n_-^+ + n_-^- = \frac{mn}{m+1}$$

Substituting all these in equation 4 we get,

$$n_+^+ \times n_-^- - n_+^- \times n_-^+ = c\sqrt{m} \times \frac{n}{m+1} \times \frac{n}{2} \tag{5}$$

Substituting n_+^+ with $\frac{n}{2} - n_-^+$ and n_+^- with $\frac{n}{2} - n_-^-$ we get,

$$n_-^- - n_-^+ = c\sqrt{m} \times \frac{n}{m+1} \tag{6}$$

Now we can solve to get the values of n_-^- and n_-^+ first, and using those to get the values of n_+^+ and n_+^- as shown in section 6.1. Note that, these values are integers which is why the fractions are rounded up using floor and ceiling. Therefore, the sampled records may not have the exact correlation as c but the difference would be negligible for our purpose.

C Details of Experiment Setup

C.1 Datasets

(1) Census19. Originating from the 2019 US Census Bureau Database [7], the Census19 dataset includes over 1.6 million

records and 12 variables capturing a wide range of personal and demographic details of US residents. The goal of this dataset is to classify individuals by their annual income, setting the threshold at over $90,000, an adjustment from the Adult dataset's $50,000 threshold to account for inflation over the years. Marital status is chosen as the sensitive attribute in this dataset. The attribute can take multiple values but we convert the attribute into binary by labeling all values except Married as Single. To streamline analysis, we categorize all instances of marital status into two groups, married or single, in the initial preprocessing steps.

(2) Texas-100X. The Texas-100X dataset expands upon the Texas-100 hospital dataset [30] originally introduced by Shokri et al [32] and contain 925,128 records from 441 hospitals. We use the PRINC_SURG_PROC_CODE column from the dataset as the output attribute for this dataset. However, PRINC_SURG_PROC_CODE is a categorical column that can take up 100 different values. The other columns do not contain sufficient information related to the surgery procedure to allow training of a target model for a 100-class classification problem with good classification accuracy on a held-out test set. Therefore, we project the 100 values into 2 distinct categories: top-10 most frequent surgery procedures and the rest of the procedures. After this mapping, the classification problem becomes a binary one. For this dataset, SEXCODE is selected as the sensitive attribute which can take up values corresponding to 'Male' or 'Female'.

(3) Adult. This dataset [5] is used to predict whether an individual earns over 50,000 a year. The dataset contains 48,842 instances and has 14 attributes. Following the preprocessing technique in [29] We merge the marital status attribute into two distinct clusters: Married, which includes 'Married-civ-spouse,' 'Married-spouse-absent,' and 'Married-AF-spouse'; and Single, which includes 'Divorced,' 'Never-married,' 'Separated,' and 'Widowed.' We then consider this attribute (Married/Single) as the sensitive attribute that the adversary aims to learn. After removing records with missing values, the final dataset consists of 45,222 records. We split the Adult dataset and use 35,222 records to train the target models, and the remaining 10,000 records to evaluate attacks on data from the same distribution but not in the training set.

C.2 Model Architecture and Training Hyperparameters

The default neural network used consists of three hidden layers having 32, 16, and 8 neurons respectively with ReLU activation function. For 4-layer MLP, another layer with 64 neuron was added adjacent to input layer and for 2-layer MLP, the layer with 32 neurons were dropped. The output layer is a softmax layer consisting of one neuron for each output class. This is a standard neural network architecture used in prior inference works [23]. The Adam optimization algorithm is incorporated with the initial learning rate set to 0.001. The training is run for 500 iterations.

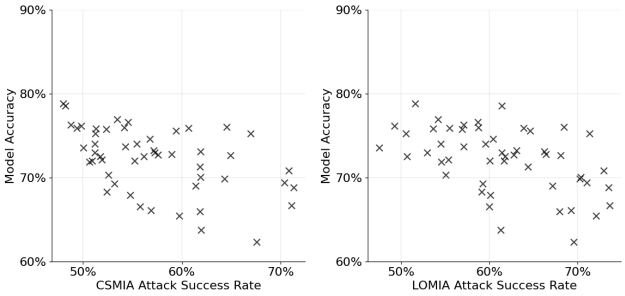


Figure 7: Model utility vs. attack performance (CSMIA - left, LOMIA - right) of 51 states from Census19 dataset. Accuracy is used as a metric for both axes.

D Additional Experiment Results

D.1 Model Utility vs. Vulnerability at Group Level

Figure 7 plots the correlation between the sensitive attribute and the output of the 51 states on the X-axis and the angular difference on the Y-axis. This is from the same scenario as section 6.3. The results reveal that there is no correlation between group-level model utility and group vulnerability. This behavior explains the failure of fairness constraint-based defenses in reducing disparity as groups with similar model utility can have different levels of vulnerability.

D.2 Correlation vs. Angular Difference

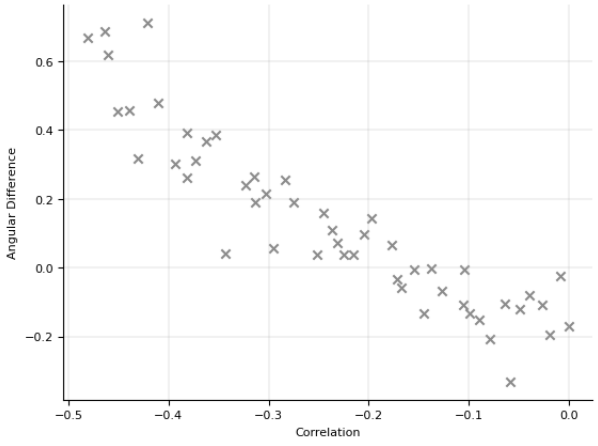


Figure 8: Correlation vs. angular difference of 51 states from Census19 Dataset

Figure 8 plots the correlation between the sensitive attribute

and the output of the 51 states on the X-axis and the angular difference on the Y-axis. This is from the same scenario as section 6.3. The results reveal that groups with a high correlation between the sensitive attribute and output tend to have a high angular difference, while those with a low correlation exhibit a low angular difference. While correlation ranges from 0 to -0.5, angular difference ranges from -0.33 to 0.71. Nonetheless, their relationship is visibly linear confirming the appropriateness of our design choice of fitting a linear regression model for the correlation estimation attack.

D.3 Angular Difference Visualization Across States

In Figure 9, we plot the confidence matrix found during the computation of angular difference for the states with indexes 0, 25, and 50 respectively. We chose these three states particularly as they denote the state with the lowest (0), median (-0.25), and the highest correlation (0.5) among the range of correlation we considered. In the Census19 dataset, the sensitive attribute is marital status and has two possible values: single and married. We chose the former as positive and the latter as negative. The output attribute, which corresponds to the output label given by the target model, also has two values - high income and low income. High income was chosen as positive and low income was chosen as negative for this attribute. The positive output records and the negative output records are plotted with different markers and two different regression lines are fit for the two sets of records. In the highest correlation case, the regression line corresponding to the low income records has a higher slope than the line corresponding to the high income records. This essentially means that the target model predicted with higher confidence when queried with the negative value for these high income records than when queried with the positive value. The root of this behavior comes from the high magnitude of correlation we set for this particular group as we identified in section 4. A high negative correlation means that there are more married high income records than their single counterpart. In summary, the plots reveal that the variation in angular difference across groups arises inherently from the variation in the correlation between the sensitive attribute and output.

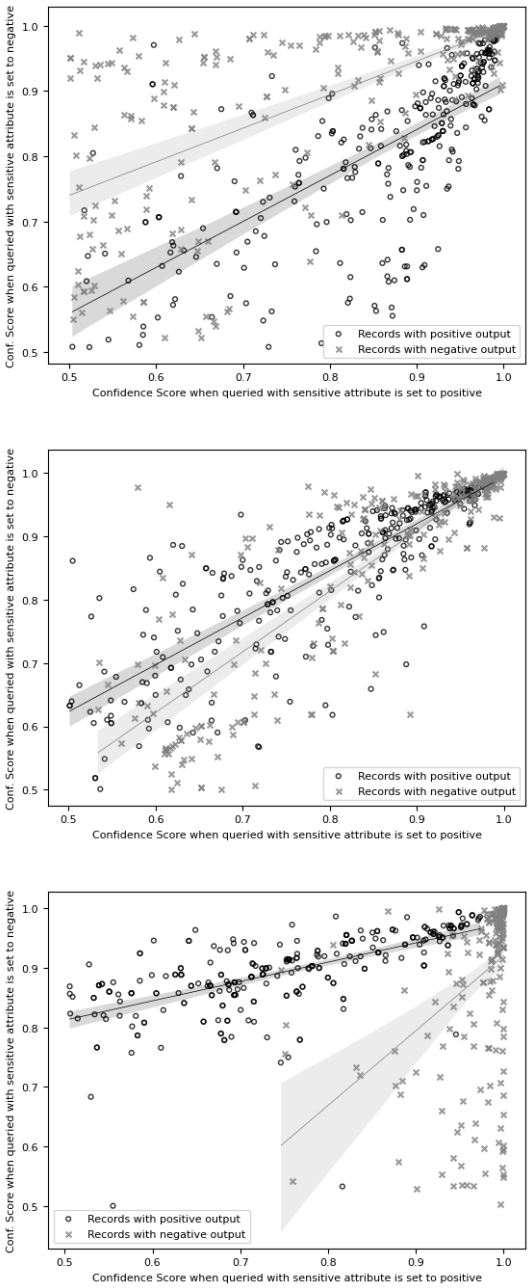


Figure 9: Confidence matrix plotted for state with index 0 (left), 25 (middle), and 50 (right). The selected three states have a correlation of 0, -0.25, and -0.5 respectively between the sensitive attribute and output.