Model Inversion Attack with Least Information and an In-depth Analysis of its Disparate Vulnerability

Anonymous Authors*

***Abstract*—In this paper, we study *model inversion attribute inference* (MIAI), a machine learning (ML) privacy attack that aims to infer sensitive information about the training data given access to the target ML model. We design a novel black-box MIAI attack that assumes the least adversary knowledge/capabilities to date while still performing similarly to the state-of-the-art attacks. Further, we extensively analyze the *disparate vulnerability* property of our proposed MIAI attack, i.e., elevated vulnerabilities of specific groups in the training dataset (grouped by gender, race, etc.) to model inversion attacks. First, we investigate existing ML privacy defense techniques– (1) *mutual information regularization*, and (2) *fairness constraints*, and show that none of these techniques can mitigate MIAI disparity. Second, we empirically identify possible disparity factors and discuss potential ways to mitigate disparity in MIAI attacks. Finally, we demonstrate our findings by extensively evaluating our attack in estimating binary and multi-class sensitive attributes on three different target models trained on three real datasets.**

***Index Terms*—model inversion attribute inference, privacy, disparate vulnerability**

I. Introduction

In recent years, machine learning (ML) has become very popular because of its wide range of applications in real-life, including predictive modeling [1], speech and speaker recognition [2], medical diagnosis [3], and image analysis [4], [5]. These increasing applications have introduced new ways for attack vectors to perform privacy attacks [6]. Among the different existing privacy attacks against ML models, e.g., membership inference [7], model extraction [8], and property inference [6], *model inversion attribute inference attack* (MIAI) is comparatively under-explored and more challenging due to its inherent characteristics [9]. Although some research has focused on attribute inference in different contexts, e.g., in social media [10], these studies do not specifically focus on ML privacy or model inversion vulnerability. In an MIAI attack, the adversary leverages its access to the target model to infer sensitive attributes in the training instances of the target ML model [1], [11]. For instance, a social survey (e.g., National Longitudinal Surveys) dataset to model how individuals rate their lives may also include their responses to questions on sensitive topics, such as drug usage behavior, sexual activities, etc. If an attacker knows background information of a target individual and can also infer some sensitive features of that target individual by querying the ML model that predicts life ratings, a significant privacy violation could occur. In-depth analysis of the vulnerabilities of model inversion attacks in the domain of tabular datasets [1], [3], [11] is even less studied.

State-of-the-art MIAI attacks in the literature consider strong adversarial capabilities [3], [11], e.g., access to the target model's confusion matrix, marginal priors, knowledge of all training instances' non-sensitive attribute values, possible values of sensitive attribute, ground truth labels of all training data instances, etc. These capabilities are not only stronger assumptions but also might not always prevail in the real world. For example, an adversary might not have access to the ground truth labels (e.g., unlabeled data in the unsupervised learning scenario like anomaly detection), might not know all non-sensitive attribute values of individual training instances, or even might not know the training data confusion matrix. These limitations are crucial challenges to existing MIAI attacks in the literature and restrict the applicability of these attacks to a more practical scenario. Therefore, in our work, we explore the following research question: *How can we design an effective MIAI attack with the least adversarial capabilities in a more realistic scenario?* We define *least* capabilities as being able to only (i) query the black-box target model and obtain only the prediction labels (not confidence scores), and (ii) have knowledge of possible values of sensitive and non-sensitive attributes (however, no information about individual training instances). A real-life scenario of our attack can be the Amazon Fraud Detector API [12] that predicts whether an input event is a fraud. In this scenario (also for other MLasS APIs), an adversary with only these least capabilities can query the model via API with unlabeled (real or synthetic) data and perform MIAI to infer sensitive attributes, e.g., a valid user's demographic or personal information.

To investigate our research question, we design a novel black-box MIAI attack with the *least* capabilities mentioned above. Our attack, dubbed as the synthetic data-based MIAI attack (SDMIA), makes the same number of queries to the target model compared to the state-of-the-art MIAI attacks [3], [11]. We query the target model to generate synthetic data, which we then use to train the attack model. We further consider limiting the number of queries in our attack to half of the state-of-the-art (denoted as the SDMIA*) attacks and demonstrate that our proposed synthetic data-based attack still matches the state-of-the-art MIAI attacks in terms of effectiveness. Our proposed SDMIA renders the following benefits over the existing attacks– (i) novel back-box attack with the least capabilities, (ii) query-efficient attack (SDMIA* requires 50% less query), (iii) more realistic and broad applicability, (iv) performs similarly in most cases compared to existing attacks, and (v) achieves more stable attack performances than state-of-the-art MIAI attacks varying target models.

Dataset characteristics and algorithmic bias in the training process may contribute to having unfair outcomes in ML

models [13], [14]. These unfair outcomes might also impact different training instances' privacy vulnerability differently, commonly expressed as *disparate vulnerability* property in privacy attacks [11], [15]. This notion captures serious privacy concerns since there could be scenarios where MIAI attacks' average accuracy measured over the entire training dataset is not significant, but the attacks are highly effective on a particular set of training instances or subgroups, e.g., instances grouped by race, gender, ethnicity, etc. Although the existing work on ML privacy has explored this *disparate vulnerability* issue, it has only been studied in-depth in the context of membership inference attacks [15]. Since the MIAI attack is a more consequential and challenging privacy attack, in our work, we perform an extensive analysis of *disparate vulnerability* in the MIAI context.

We extend our study to explore potential disparity contributing properties of training dataset/ target model (e.g., correlation, marginal priors, etc.) termed as *'factors'* in this paper. First, we investigate two popular ML privacy defense techniques against the *disparate vulnerability* property of the MIAI attack– *mutual information regularization* [16] and *fairness constraints* [15]. We analyze the impacts of these techniques on (i) individual subgroup vulnerability, (ii) disparity among subgroup vulnerabilities, and (iii) target model utility. Our experiments show that although these techniques do not significantly affect target model utility, they fail to perform consistently in reducing disparity and subgroup vulnerability. Second, we expand our MIAI disparity analysis and empirically identify possible factors for the *disparity* in MIAI. We consider two types of disparity factors– factors dependent on the target model (e.g., overfitting) and factors largely dependent on training data distribution (e.g., skewness, kurtosis, etc.). We further explore how these factors contribute towards *disparity* in MIAI, i.e., their normalized weights. Finally, based on our findings, we discuss potential ways to design effective defenses for mitigating disparity in MIAI.

We empirically evaluate our SDMIA and SDMIA* on three different target models (decision tree (DT), logistic regression (LR), and deep neural network (DNN)) trained on three real datasets (Adult [17], NLSY [18], and FiveThirtyEight [19]). Our analysis shows that SDMIA can achieve more stable and similar performances (even better in some metrics in some scenarios) compared to the existing MIAI attacks, despite assuming the least capabilities for the adversary and even with fewer queries. Our in-depth disparity factor analysis demonstrates that while the target model impacts the most influential disparity factor (e.g., mutual information in the deep neural network, correlation in the decision tree), the overall disparity in MIAI results from the holistic impact of multiple factors. This finding calls for designing robust multi-factor-based disparity mitigation strategies to adopt in the future for effective MIAI disparity mitigation.

The contributions of this paper are as follows:

1) We design a novel black-box MIAI attack that considers the *least* adversarial knowledge/capabilities to date while performing similarly to the state-of-the-art MIAI attacks.
2) We formally define and analyze our proposed MIAI attack's *disparate vulnerability* property and introduce two metrics for *disparity* measures.
3) We investigate existing ML privacy defense techniques– *mutual information regularization* and *fairness constraints* and show that none of these techniques can mitigate MIAI *disparity* consistently.
4) We empirically identify the factors behind the *disparity*, their contributions, i.e., their importance/weights towards *disparity*, and discuss possible directions for disparity mitigation.
5) We extensively evaluate the proposed SDMIA using three different target models trained on three real datasets.

II. Preliminaries

MIAI Attack. Let f be a deterministic function representing the target ML model. The input of f is a d-dimensional vector $x = [x_s, x_2, ..., x_d]$, where d signifies the number of input attributes. Without loss of generality, let x_s be the sensitive attribute and $\{x_2, ..., x_d\}$ be the set of the non-sensitive attributes. An adversary exploits available auxiliary information in MIAI attacks to infer the training data sample's sensitive attribute (x_s) value. Adversaries can query the target models through APIs made available by the ML-as-a-service (MLaaS) providers [20]–[22]. The auxiliary information may include full or partial knowledge of the non-sensitive attribute values $\{x_2, ..., x_d\}$ in the training dataset, confusion matrix, marginal priors, confidence scores, and predicted labels returned by the target model, etc.

A comparative analysis of the adversarial capabilities assumed in Fredrickson et al. attack on decision tree (FJRMIA) [3], LOMIA [11], and our SDMIA is illustrated in Table I. In FJRMIA, an adversary has access to the confusion matrix, all non-sensitive attribute values of target individuals' training instances, and marginal priors. It queries the target model with all possible sensitive attribute values and predicts the one that maximizes the posterior value computed by $cm[y, y'] * p_s$, where cm is the confusion matrix, y, and y' are the true label and the model's predicted label, respectively, and p_s denotes marginal prior of the sensitive attribute value. LOMIA [11] considers similar adversarial capability assumptions except access to the marginal priors.

Attribute Importance. An ML model, either classification or regression, predicts output y' based on input attributes $\{x_s, x_2, ..., x_d\}$. However, all input attributes might not be equally impactful toward prediction. Therefore, the attribute/feature importance is commonly considered to assign scores to all input attributes [23]. Importance of an attribute x_d in the DT model can be formally defined [24] as below:

$$\mathcal{I}mp_{AI}(x_d) = \mathcal{M}_A(x_d) * \mathcal{S}_{x_d} - \mathcal{M}_A(L_c) * \mathcal{S}_{L_c} - \mathcal{M}_A(R_c) * \mathcal{S}_{R_c} \quad (1)$$

where the $\mathcal{I}mp_{AI}(x_d)$ is the importance of attribute x_d (i.e., node in DT), $\mathcal{M}_A$ is the metric value (e.g., Entropy or Gini), $\mathcal{S}_{x_d}$ is the number of samples at the node, and R_c, L_c indicate right and left child, respectively. So, importance in DT is

computed by multiplying the attribute's (node) metric value with the number of samples at the node and subtracting scores (metric value * samples) of its left and right child. For DNN or LR models, attribute importance can be computed using weighted scores of the attribute coefficients returned by the trained models [25], [26].

III. Related Work

The concept of model inversion (MI) attack is fairly recent. Fredrickson et al. [1] proposed the concept of MI attack in 2014 by successfully conducting an attack on linear regression models to uncover genomics information about target individuals. MI attacks are applicable to data domains including image, tabular, text, or audio data [1], [5]. MI attacks in the image domain (i.e., image reconstruction) have been a central research focus [3], [5], [27]. However, MIAI attacks on tabular data are underexplored [3], [11]. Fredrickson et al. [3] conducted MI attacks on both image and tabular data. They evaluated the maximum a posteriori (MAP) algorithm from [1] for black-box MIAI attacks on tabular data (decision tree). We denoted this baseline by FJRMIA. Another baseline MIAI attack on tabular data is LOMIA, proposed by Mehnaz et al. [11], where the adversary trains an attack model querying the target. Both attacks consider a wide range of capabilities.

In the privacy research domain, membership inference attack [7], [28], [29] and MIAI attack [3], [11] have different goals. While the membership inference infers whether a sample is in the target model training dataset (i.e., binary classification), MIAI infers the sensitive attribute value of a training sample. Several works have explored the membership inference attack based on only the predicted label [28], [29]. Christopher et al. [29] proposed three label-only membership inference attacks leveraging the robustness of the model, which can achieve comparable performance as the confidence score-based attack. Zheng et al. [28] proposed two decision boundary-based label-only membership inference attacks. However, recent research shows that membership inference attack has different characteristics than MIAI attack, and MIAI does not show any consistent pattern [9]. For example, dataset complexity and training epochs positively impact membership inference attack performance, whereas these have no such impact on the MIAI attack. Similarly, unlike membership inference, there is no explicit relationship between overfitting and MIAI attacks. All these distinct characteristics and inconsistent patterns make an MIAI attack more challenging than other privacy attacks, including membership inference.

Commonly used privacy attack defense techniques include *regularization, differential privacy, fairness constraints,* and *rounding confidence score* [3], [16]. Wang et al. [16] proposed a *mutual information regularization* technique to defend against MI attacks (both image and tabular data). However, this work did not explore the impact of the *mutual information regularization* technique on *disparate vulnerability* property in MIAI. Therefore, we extend the technique in [16] to investigate its impact on MIAI *disparate vulnerability* and whether it can consistently reduce disparity in MIAI without worsening individual subgroup vulnerability magnitudes. Another work by Kulynych et al. [15] investigates *fairness constraints* and *differential privacy* as techniques to reduce disparity in membership inference attacks. *Differential privacy* is not effective against MIAI attacks without significantly compromising model utility [1], [30]. Also, our proposed attack does not require confidence scores, so *rounding confidence* is not applicable. Therefore, we only investigate *mutual information regularization* and *fairness constraints* as two viable MIAI disparity mitigation techniques.

Kulynych et al. perform the first in-depth analysis on disparate vulnerability in membership inference attacks [15]. They identified *overfitting* and subgroup size as factors impacting *disparity* in membership inference attacks. We perform an in-depth analysis of *disparity* in MIAI attacks to identify possible disparity factors with their impacts on MIAI disparity.

IV. Model Inversion Attack with the Least Adversary Capabilities

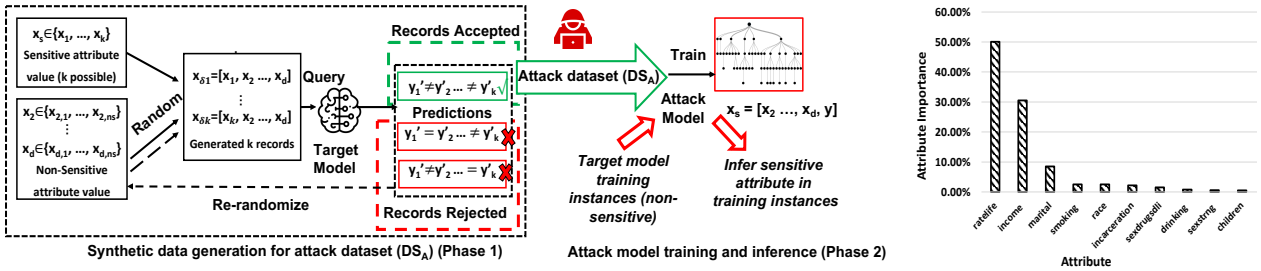
In this section, we first present the threat model and then illustrate our proposed synthetic data-based MIAI attack (SDMIA) with the least adversary capabilities.

TABLE I: Comparison among different attacks in terms of adversary capabilities.

Capability	FJRMIA [3]	LOMIA [11]	SDMIA
(1) Target model's confusion matrix	●	●	
(2) Marginal prior (sensitive attribute)	●		
(3) Marginal prior (non-sensitive attributes)	●		
(4) Possible values of attributes	●	●	●
(5) Training data instances' non-sensitive attribute values	●	●	
(6) Training data instances' actual labels	●	●	
(7) Target model's prediction labels	●	●	●

A. Threat Model

In our threat model, two parties are involved– the model owner and the adversary. Individual training data instances are an adversary's targets. The adversary aims to infer an individual's sensitive attribute in the training data. We assume the adversary does not impact the target model training process or performance; the adversary only interacts with the model owner by issuing API queries to the black-box model with the least capabilities, thereby designing the attack model to infer target individual's sensitive attribute in the training data. The attack is considered successful if the adversary can accurately infer the target individual's sensitive attribute value. In Table I, we illustrate the adversarial capabilities of existing attacks and our proposed MIAI attack. In our attack, the adversary has black-box access to the target model and can only obtain the predicted labels (y') by querying the model. Also, the adversary knows the possible values of sensitive and non-sensitive attributes. Unlike the existing attacks, our attack does not assume knowledge of marginal priors, target model confusion matrix cm, or any knowledge about the target model training dataset instances.



(a) *SDMIA Overview* (b) *Attribute Importance*

Fig. 1: (a.) Overview of our proposed SDMIA. It works in 2 phases– phase 1: Synthetic data generation for attack dataset (DS_A), and phase 2: Attack model training and inference. (b.) Importance of attributes in attack model, while estimating *drug_marijuana* in NLSY dataset.

B. Our Proposed Synthetic Data-based Model Inversion Attack (SDMIA)

In this section, we introduce our synthetic data-based MIAI attack (SDMIA). We also consider an instance of the SDMIA (SDMIA*) that uses half the number of queries (i.e., 50%) as all of the existing attacks (i.e., LOMIA [11] and FJRMIA [3]). We also experiment with 25% and 75% queries and show that SDMIA is effective in both cases (see Fig. 15 in Appendix). For each instance in the training dataset, the adversaries in LOMIA [11] and FJRMIA [3] query k times (k is the number of possible values of the sensitive attribute), meaning they require $k * m$ queries in total, where m is the number of instances in the training dataset. Unlike these attacks, we generate random queries in SDMIA but constrain the query numbers to $k*m$, which is further reduced to $\frac{k*m}{2}$ in SDMIA*. SDMIA and SDMIA* follow the same attack steps; the only difference is the number of queries to the target model. Our attack works for both binary ($k = 2$) and multi-class ($k > 2$) sensitive attributes. However, if the number of possible values of the sensitive attribute (k) is larger than the number of class labels, there will always be at least two sensitive attribute values for which the target model would predict the same output label (according to the pigeonhole principle), resulting in no records accepted in Fig.1a. Therefore, we consider $k <$ number of classes (label), which holds for different real-life datasets and is consistent with the baseline [11]. SDMIA has two phases: (1) Synthetic data generation for the attack dataset (DS_A) and (2) attack model training and inference. In Fig. 1a, we present the overview of our proposed SDMIA strategy.

1) Phase 1: Synthetic Data Generation: The core component of our proposed SDMIA is the synthetic data generation algorithm (Algorithm 1 in Appendix A). This algorithm is inspired by the synthetic data generation technique proposed in [7] for membership inference attacks. However, unlike the existing algorithm, our algorithm is designed for MIAI attacks only considering access to prediction labels (without confidence scores). Also, our algorithm generates k query instances varying k possible sensitive attribute values in each iteration. The complexity of our algorithm is $O(c_2 + c_1 * m)$, where m, c_1, and c_2 stand for loop counter (number of instances queried), array access time (constant time to access an attribute's array for generating records), and outside loop run time (constant time for initialization purposes outside the loop). Therefore, our algorithm has linear time complexity.

Our synthetic data generation algorithm consists of three steps: i) *initialization*, ii) *querying*, and iii) *selection*. In *initialization*, the algorithm randomly initializes values for $d-1$ non-sensitive attributes to form the partial query x^-. It also initializes d_{max} and d_{min}, which denote the max and min number of non-sensitive attributes that would change in every iteration, respectively. Additionally, we initialize max record rejection threshold r_{max}. In *querying*, the adversary varies the sensitive attribute value and generates k query instances from x^-, i.e., $x_{\delta_1}, \ldots, x_{\delta_k}$, where k is the number of unique possible values of the sensitive attribute. The target model f_{tar} is then queried with these generated instances, and the adversary receives the corresponding predictions (y's).

In *selection*, the adversary either includes the generated instances to the attack model training dataset (DS_A) or rejects them based on the returned predictions. If all the k predictions using k possible values of the sensitive attribute (where $k \geq 2$) are different, we accept the instances to form DS_A; otherwise, we reject them (Fig. 1a). In the above scenario, k query instances set $x_{\delta_1}, \ldots, x_{\delta_k}$ is accepted to form DS_A if all the k predictions $y'_1, \ldots, y'_k$ are different (green marked in Fig. 1a). If at least two predictions are the same, the set of instances is rejected (red marked in Fig. 1a). Regardless of acceptance/rejection, in every iteration we re-randomize non-sensitive attribute values by changing d_{max} non-sensitive attributes (chosen uniformly at random from $d-1$ non-sensitive attributes). However, in case of consecutive rejections, we update d_{max}. If the number of consecutive rejections exceeds the max rejection threshold (r_{max}), we update d_{max} by $max(d_{min}, \lceil d_{max}/2 \rceil)$. We reduce d_{max} to control the search diameter around the accepted records while re-randomizing. This enhances search speed and ensures accepted records vary in the least number of attributes enabling better capture of the correlation between sensitive attribute and label.

The goal of generating synthetic data is to capture the dependency between the input attributes and the output (label). Since we initialize the non-sensitive attributes randomly

TABLE II: Sensitive attributes (binary) distribution in the training datasets.

Dataset	Training Instances	Sensitive attribute	Positive class label	Positive class count	Positive class %
NLSY	5096	drug_marijuana	dm_yes	961	18.9%
Adult	35222	marital	married	16893	47.9%
FiveThirtyEight	331	alcohol	yes	266	80.36%

(controlling search diameter) and vary the sensitive attribute values, the synthetically generated DS_A dataset primarily emphasizes the correlation between the sensitive attribute and the label of the target model (varying the sensitive attribute yields different target model predictions). In Fig. 1b, we demonstrate attribute importance in the attack model to infer *drug_marijuana* sensitive attribute in NLSY dataset, where the target model label *ratelife* has the highest importance in the attack model (at 50.1%). Also, in the Adult dataset, among 18,329 *single* individuals, only 1208 have $\geq 50K$ label, i.e., *income*, whereas, among the remaining 16,893 *married* individuals, 7565 have $\geq 50K$ label. Therefore, to capture this strong dependency, we consider accepting the set of instances to form DS_A only when all the k predictions from the target model using k possible sensitive attribute values are different.

2) Phase 2: Attack Model Training and Inference: In this phase, we train an attack model f_{adv} using the generated synthetic dataset DS_A from phase 1 (Section IV-B1). This attack model performs the model inversion, i.e., it predicts sensitive attribute values when given the non-sensitive attributes as the input. More specifically, we train the f_{adv} attack model with non-sensitive attribute values $(x_{ns} = x_2, ...x_d)$ and y' (i.e., prediction) in DS_A dataset as the input features where the output label is the sensitive attribute x_s, i.e., $x_s = f_{adv}(y' \cup x_{ns})$, as shown in Fig. 1a. Once the f_{adv} model is trained, it can be deployed to infer the sensitive attribute of any target model training dataset instance. We evaluate the performance of SDMIA and SDMIA* in Section IV-C4.

C. SDMIA Evaluation

This subsection describes datasets, SDMIA experimental setup, performance metrics, and comparison among our proposed SDMIA and the existing MIAI attacks.

1) Dataset Description: We use publicly available datasets described below for our experiments in this paper.

Texas. This dataset is based on Texas Hospital discharge data [31]. We consider 1st quarter hospital data of five consecutive years (2006-2010) and extract features related to patient status, length of hospital stays, illness severity, and demographic information[1]. We use *pri_proc* as the label, indicating the primary procedure the patient has undergone (e.g., type1, type2, type3 classes). First, we apply stratified sampling to randomly select 20K instances from each year, based on the most frequent *pri_proc*. We then combine instances and obtain a dataset of 100K instances. After additional pre-processing (removing all duplicate instances), we end up with 89,924 instances which we split into 67,443 ($\sim$ 75%) and 22,481 ($\sim$ 25%) for training and testing, respectively.

NLSY. The National Longitudinal Survey of Youth (NLSY) dataset consists of responses from the survey conducted in 1997 [18] on 8984 individuals born between 1980-1984 living in the US. We extract features related to participant demographic information, including age, marital status, race, habits (e.g., smoking, drinking), and ratings of life (participants rate their life into classes, e.g., *excellent, very good, good, fair, poor*) [18]. The dataset has 15 attributes, and we consider ratings of life, i.e., *ratelife* (classes: *excellent, very good, good, fair, poor*) as the label attribute. After pre-processing, we obtain 6795 instances which we split into 5096 ($\sim$ 75%) and 1699 ($\sim$ 25%) for training and testing, respectively.

Adult. The adult dataset is extracted from the 1994 Census database. The purpose is to predict whether an individual has income is higher or lower than 50K in a year [17] from the individual's other features. The dataset has 48,842 instances and 14 attributes, including marital status, occupation, education, and race. After pre-processing, we obtain 45,222 instances which we split into 35,222 ($\sim$ 75%) and 10,000 ($\sim$ 25%) for training and testing, respectively.

FiveThirtyEight. FiveThirtyEight Datalab [19] surveyed 553 individuals to identify their steak preferences, given their habitual features and demographic data, e.g., smoking, drinking, age, and gender. This dataset has 15 attributes with $steak\_type$ as the target label. After pre-processing, we obtain 331 instances that we use exclusively for training.

2) Experiment Setup: In Table II, we present the total instances in the training sets and the binary sensitive attributes we consider, along with their positive class counts in the training set. In Table IX, we present multi-class sensitive attributes that we consider. There is no positive class for multi-class; therefore, we present the average of all class estimation performances in our experiments. For each dataset, we train 3 different target models in the BigML [32] platform– decision tree (DT), logistic regression (LR), and deep neural network (DNN). We use BigML's 1-click supervised feature for training the target models with default parameters. We present performances of all 3 target models trained on the Adult dataset, tested on training and test datasets in Tables III, IV, respectively. We observe that target models perform similarly on training and test datasets, indicating no significant overfitting on target models. Target model performances trained on other datasets are presented in Tables XI, XII in Appendix. We query the trained target model with API provided by the BigML to obtain the attack dataset DS_A (using Algorithm 1). For each target model, we train the attack model as a bootstrap decision-tree forest of 10 decision tree models trained on the attack dataset DS_A (i.e., bagging technique), denoted by 1-click ensemble model in BigML [32]. We use decision-tree forest, an ensemble model, to achieve better attack performance.

We compare performances of our proposed SDMIA with baselines for both binary (k=2) and multi-class (k>2) sensitive attributes in different datasets, varying the target models. In

[1]Our experiments comply with data sharing and re-identification agreements.

TABLE III: Target model performances trained on Adult dataset (tested on *training* dataset)

Model	Precision	Recall	Accuracy	F1 score
DT	83.05%	78.04%	86.15%	80.00%
DNN	77.16%	80.68%	82.83%	79.00%
LR	79.95%	76.02%	84.31%	78.00%

TABLE IV: Target model performances trained on Adult dataset (tested on *test* dataset)

Model	Precision	Recall	Accuracy	F1 score
DT	81.59%	76.16%	85.29%	78.00%
DNN	75.96%	79.29%	82.11%	77.00%
LR	79.34%	75.2%	84.11%	77.00%

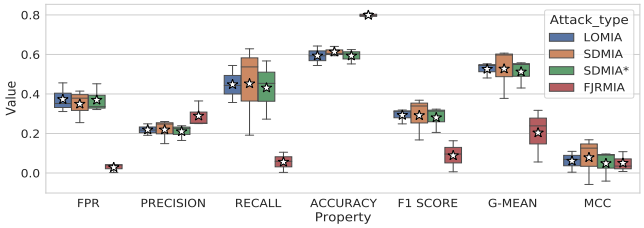


Fig. 2: Comparison of performance metrics distributions among SDMIA, SDMIA*, LOMIA, and FJRMIA; while inferring *drug_marijuana* (lower prior *dm_yes*=18.9% positive class) obtained from 3 target models trained on NLSY dataset.

binary sensitive attribute (k=2), we consider three different scenarios (depending on sensitive attribute's positive class marginal priors): (i) estimating *drug_marijuana* in NLSY dataset, where the positive class (i.e., 'dm_yes') has very low (18.9%) prior, (ii) estimating *alcohol* in FiveThirtyEight dataset, where the positive class (i.e., 'Yes') has very high (80.3%) prior, and (iii) estimating the *marital* in the Adult dataset, where priors are balanced, i.e., positive class (i.e., 'married') has 47.96% marginal prior. For each scenario, we perform all attacks on three target models (DT, LR, DNN) trained on NLSY, FiveThirtyEight, and Adult datasets. We present results in Section IV-C4.

3) Evaluation Metrics: Since accuracy is not an effective evaluation metric for imbalanced class data, we also consider the F1 score, which balances precision and recall of the positive class only. This metric also does not capture the overall performance. We consider Mathews correlation coefficient (MCC) and Geometric-mean (G-Mean) [33], which account for all confusion matrix entries and provide more balanced measures. We also consider false-positive rates (FPR) to measure the model's incorrectness in prediction. In different metrics, besides numerical comparison, we compare attacks in *attack stability* property, i.e., the resistance of the attack performance to target model variations (model agnostic).

4) SDMIA Performance Comparison:

(1) Estimating binary sensitive attributes (k=2): In Fig. 2, we present attack performance distributions (horizontal bars ($-$) and stars ($\star$) in boxes indicate *'median'* and *'mean'*) from three different target models in scenario (i), as discussed in Section IV-C2. Since FJRMIA is impacted by positive

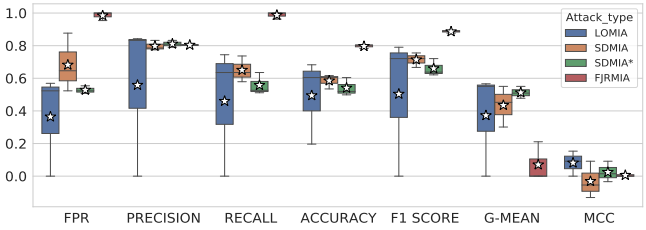


Fig. 3: Comparison of performance metrics distributions among SDMIA, SDMIA* and existing LOMIA, FJRMIA; while inferring *alcohol* (higher marginal prior *Yes*=80.3% positive class) obtained from 3 target models trained on FiveThirtyEight dataset.

class marginal priors, in scenario (i), we obtain the lowest performances, i.e., F1 score (8.85%) and G-mean (20.38%) for FJRMIA (Fig. 2). Mean FPR, F1 score, and accuracy for SDMIA* are similar to those of LOMIA, despite having the least adversary capabilities, and queries to the target model are limited to half. For example, mean FPR and accuracy scores in LOMIA are 37.3% and 59.23%, whereas SDMIA* has 36.94% and 59.29% mean FPR and accuracy, respectively. Also, SDMIA performs similarly to LOMIA. In summary, all attacks, except FJRMIA, achieve similar performances. In Table V, we present the attack performances for DT target models (Tables XVI and XVII in Appendix for DNN, LR).

In Fig. 3, we present distributions of attack performances in scenario (ii). For higher marginal prior positive class, i.e., in scenario (ii), FJRMIA achieves the highest recall at 98.75% but results in an FPR of 98.46% (Fig. 3). In contrast, all other attacks achieve comparable and more consistent performance. Moreover, SDMIA and SDMIA* have more stable distributions compared to LOMIA across different models (Fig. 3). For example, LOMIA has a higher *interquartile range (IQR)* of accuracy (24.32), whereas SDMIA and SDMIA* have 4.07 and 5.28 IQRs, respectively. In some cases, our SDMIA and SDMIA* outperform LOMIA. For example, the mean values of precision and F1 score in SDMIA* are 81.08% and 65.98%, whereas SDMIA achieves 79.76% and 71.42%, and LOMIA has the lowest values at 55.84% and 50.37%, respectively. In Table VI, we present the attack performances for DT target models (Tables XVIII and XIX in Appendix for DNN, LR).

In scenario (iii), when the classes have more balanced priors, we find LOMIA, SDMIA, and SDMIA* to have similar performances, and all three outperform FJRMIA. We present the distribution in the Appendix in Fig. 18; FJRMIA has the lowest F1 score (35.30%) compared to $\sim$59% in other attacks. In Table VII, we present the attack performances for the DT target model (Tables XX and XXI in Appendix for DNN, LR).

(2) Estimating multi-class sensitive attributes (k>2): We experiment with multi-class sensitive attributes (k=4) *age*={*18-29, 30-44, 45-60, >60*} in FiveThirtyEight and *race*={*nonblhis, black, hispanic, mixed*} in NLSY dataset. We compute average class performances for each attack for each target model. In Fig. 4a and 4b, we present the mean and

TABLE V: Attack performance while inferring *drug_marijuana* sensitive attribute, against DT model trained on NLSY dataset

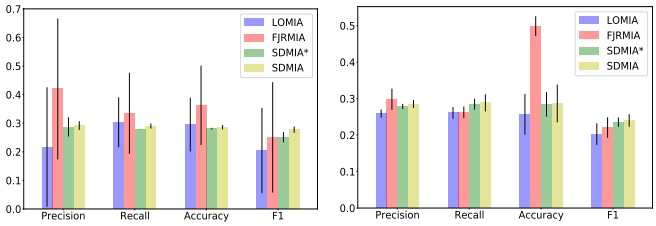
Attack Strategy	Query#	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	10192	101	3959	176	860	36.46%	10.51%	79.67%	16.32%	31.72%	10.79%
LOMIA [11]	10192	426	2849	1286	535	24.88%	44.33%	64.27%	31.87%	55.27%	10.96%
SDMIA	10192	516	2569	1566	445	24.78%	53.69%	60.54%	33.91%	59.72%	12.59%
SDMIA*	5096	436	2747	1388	525	23.90%	45.37%	62.46%	31.31%	54.90%	9.63%

TABLE VI: Attack performance while inferring *alcohol* sensitive attribute, against DT model trained on FiveThirtyEight dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	662	266	0	65	0	80.36%	100.00%	80.36%	89.11%	0.00%	0.00%
LOMIA [11]	662	198	28	37	68	84.26%	74.44%	68.28%	79.04%	56.63%	15.33%
SDMIA	662	196	8	57	70	77.47%	73.68%	61.63%	75.53%	30.11%	−13.11%
SDMIA*	331	140	32	33	126	80.92%	52.63%	51.96%	63.78%	50.90%	1.48%

TABLE VII: Attack performance while inferring *marital* sensitive attribute, against DT model trained on Adult dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	70444	3788	17818	511	13105	88.11%	22.42%	61.34%	35.75%	46.69%	29.97%
LOMIA [11]	70444	7574	17132	1197	9319	86.35%	44.84%	70.14%	59.02%	64.74%	44.25%
SDMIA	70444	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%
SDMIA*	35222	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%



(a) *age* (b) *race*

Fig. 4: Performances comparison among different attack strategies, while estimating multi-class (*k=4*) sensitive attribute *age* in *FiveThirtyEight* and *race* in *NLSY* dataset.

standard deviations (error bars) of average class performances across models. In both plots, observe that SDMIA and SDMIA* perform slightly better than LOMIA in precision and F1 scores and do not perform significantly differently in accuracy and recall than LOMIA. FJRMIA has higher accuracy than all attacks due to its biases to class marginal priors, which highly impacts multi-class accuracy. SDMIA and SDMIA* achieve slightly better F1 scores than all attacks. Significantly higher error bars in FJRMIA and LOMIA in most metrics indicate significant performance variation across models.

In Fig. 15 (Appendix), we present average SDMIA performances, varying query numbers (i.e., 25%, 50%, and 75% fewer queries; and the same queries as existing attacks). It shows SDMIA is still effective even with queries reduced beyond 50%. Observe that SDMIA performances drop slightly as query numbers are reduced (except some outliers). A similar positive relationship between attack performance and query numbers is demonstrated by Chandrasekaran et al. [34] for model extraction attacks.

Discussion. Target model architecture impacts the attack stability, i.e., causes performance variations (e.g., error bars in Fig. 4 or higher IQRs in Fig. 3). However, among all attacks, our proposed SDMIA is more resilient to this phenomenon due to its ability to capture sensitive attribute-label dependency effectively compared to the existing attacks. FJRMIA and LOMIA query with the training data instances (known non-sensitive attributes). Comparatively higher dissimilarity among these instances (many non-sensitive attributes differ) causes higher prediction variations across models, negatively impacting attack stability. In other words, non-sensitive attributes also have significant attribute importance in the attack. Whereas, in SDMIA, random record generation with limiting search space enables comparatively higher similar record generation (only d_{max} attributes differ). This contributes to lower prediction variations across target models, i.e., similar distributions in attack datasets (label has significantly higher importance in the attack). Therefore, SDMIA captures sensitive attribute-label dependency strongly, achieving better stability. *In summary*, existing attacks are less stable in both binary and multi-class sensitive attribute estimation, while SDMIA is more stable. We conclude that SDMIA effectively estimates both binary and multi-class sensitive attributes, given that it requires the least capabilities and 50% fewer queries in SDMIA*.

V. Disparate Vulnerability in MIAI

Disparate Vulnerability captures uneven impacts of the attack on subgroups, i.e., attack performances vary across disjoint training instances of subgroups, and particular subgroup instances might be more vulnerable than other subgroups. For example, while inferring marital happiness or other personal sensitive information, *divorced* individuals might have higher vulnerability than *married* individuals. First, we formally define *disparity* in subgroup vulnerability and then analyze two probable solutions for mitigating MIAI disparity– *mutual information regularization* and *fairness constraints*.

A. Formal Definition of Disparity

In an MIAI attack, the performance of the adversary ($\mathcal{A}$) quantifies the term *'vulnerability'*. We define two types of *'vulnerability'* in MIAI– i) *average vulnerability*, when $\mathcal{A}$ has no specific target subgroup knowledge, we refer it as $\mathcal{V}_a(\mathcal{A})$, and ii) *subgroup vulnerability*, when $\mathcal{A}$ has subgroup knowledge and targets a specific subgroup, we refer it as $\mathcal{V}_s(\mathcal{A})$.

Average Vulnerability $\mathcal{V}_a(\mathcal{A})$: We mathematically define $\mathcal{V}_a(\mathcal{A})$ as follows:

$$\mathcal{V}_a(\mathcal{A}) = Pr[MIAI(f_{tar}, V(x_s, x_{ns}), d_{ta}) = x_s] \quad (2)$$

where $V(x_s, x_{ns})$ is the set of possible values of sensitive and non-sensitive attributes, and d_{ta} is the targeted instance in the training dataset (the randomness is from d_{ta} selection). The vulnerability is determined by the correct estimation of sensitive attribute values (i.e., $MIAI(f_{tar}, V(x_s, x_{ns}), d_{ta}) = x_s$), where x_s is the actual value of the sensitive attribute, and $MIAI(f_{tar}, V(x_s, x_{ns}), d_{ta})$ represents the estimated sensitive attribute value.

Subgroup Vulnerability $\mathcal{V}_s(\mathcal{A})$: Let the attribute of interest for subgroup vulnerability be x_b (e.g., gender), i.e., we are interested in understanding whether MIAI attacks pose significantly different vulnerability for different subgroups (e.g., male, female) based on this x_b attribute, where a subgroup $S = s_g$ is defined as the disjoint set of samples that belong to a particular subgroup. We mathematically define $\mathcal{V}_s(\mathcal{A})$ as:

$$\mathcal{V}_s(\mathcal{A}) = Pr[(MIAI(f_{tar}, V(x_s, s_g, x_{ns}), d_{ta}|S = s_g) = x_s] \quad (3)$$

where all notations are similar to $\mathcal{V}_a(\mathcal{A})$ with the additional capability of $\mathcal{A}$ to target a specific subgroup $S = s_g$ of subgroup attribute x_b (the randomness is from d_{ta} selection for $S = s_g$). Hence, $\mathcal{V}_s(\mathcal{A})$ in Eqn. 3 represents the vulnerability of subgroup s_g only.

Disparity: For different subgroups $S = s_1, ..., s_n$, the subgroup vulnerability $\mathcal{V}_s(\mathcal{A})$ can vary, i.e., $\mathcal{V}_{s1}(\mathcal{A}) \neq \mathcal{V}_{s2}(\mathcal{A}) \ldots \neq \mathcal{V}_{sn}(\mathcal{A})$. We define *disparity* in $\mathcal{V}_s(\mathcal{A})$ as:

$$\mathcal{V}_{ds}(\mathcal{A}) = gap(\mathcal{V}_{s1}(\mathcal{A}), ..., \mathcal{V}_{sn}(\mathcal{A})) \quad (4)$$

where the gap stands for the generic distance among subgroup vulnerabilities $\mathcal{V}_s(\mathcal{A})$ and can be computed using a suitable metric. We propose two metrics to compute this gap.

B. Proposed Metrics for Disparity $\mathcal{V}_{ds}(\mathcal{A})$ Measure

We introduce two metrics as defined below to measure disparity, i.e., gap among subgroup vulnerabilities.

Absolute Maximum Disparity (AMD): We denote this metric by the absolute difference between the maximum and minimum subgroup vulnerabilities ($\mathcal{V}_s(\mathcal{A})$) in an attribute, expressed as:

$$\mathcal{V}_{ds}^{amd}(\mathcal{A}) = |V_{max} - V_{min}| \quad (5)$$

where V_{max} is the maximum subgroup vulnerability and V_{min} is the minimum subgroup vulnerability.

Maximal Mean Disparity Deviation (MMDD): We use this metric to capture the central tendency, i.e., the maximal deviation of subgroup vulnerabilities from the mean value. We denote MMDD by the absolute difference between maximum subgroup vulnerability $\mathcal{V}_s(\mathcal{A})$ and average subgroup vulnerability $\mathcal{V}_a(\mathcal{A})$ defined in Section V-A. We express MMDD as:

$$\mathcal{V}_{ds}^{mmdd}(\mathcal{A}) = |V_{max} - V_a| \quad (6)$$

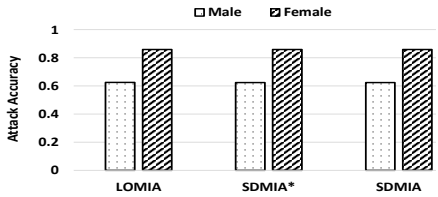


Fig. 5: Illustrating *disparity* in MIAI attacks– vulnerabilities in *sex* subgroups in Adult dataset, while estimating *marital*.

C. Disparity in SDMIA

We perform experiments with subgroup vulnerabilities ($\mathcal{V}_s(\mathcal{A})$), and show that disparity exists in both existing and proposed MIAI attacks. In Fig. 5, we present the subgroup vulnerabilities for DT target model trained on the Adult dataset where the adversary estimates *marital* sensitive attribute. As first demonstrated by Mehnaz et al. [11] in their LOMIA strategy, the attack accuracy against *male* and *female* subgroups are 62.5% and 85.8%, respectively. In our SDMIA and SDMIA*, we also observe similar trends, i.e., a significant disparity with ∼62.4% and ∼85.9% attack accuracy against the two subgroups, respectively. This indicates that an adversary, even with the least capabilities, can perform SDMIA and achieve better performance on higher vulnerable subgroups than other subgroups due to disparity.

Therefore, we perform an in-depth study of MIAI disparity to investigate the impact of existing privacy defense techniques on disparity and potential factors contributing to disparity.

D. Impact of Existing Techniques in Disparity Mitigation

We empirically analyze the effectiveness of the existing privacy defense techniques to mitigate disparity, i.e., $\mathcal{V}_{ds}(\mathcal{A})$ among subgroups without increasing individual subgroup vulnerabilities, i.e., $\mathcal{V}_s(\mathcal{A})$ and without compromising target model utility significantly. We investigate the impact of mutual information regularization and fairness constraints on MIAI disparity mitigation.

1) Mutual Information Regularization: Wang et al. [16] proposed a *mutual information (MI) regularization* based approach to defend against model inversion attacks. The key idea is to reduce dependency between target model inputs and labels by applying a regularization penalty. We extend this in the context of subgroup vulnerability $\mathcal{V}_s(\mathcal{A})$. We aim to leverage attribute level dependency (i.e., between subgroup attribute x_b and label attribute y') in the form of *mutual information* (MI) and reduce that to investigate its impact on subgroup vulnerabilities and disparity in MIAI. In our work, we consider a similar approach as [16] to reduce attribute level MI by incorporating an MI-based regularizer in the training loss function. We apply the regularizer as an additional penalty term in the loss function. We consider the following overall loss function to train the target model:

$$\mathcal{L}_{mi} = \min_{f \in \mathcal{H}} \mathcal{L}_e(y, f(x)) + \lambda \mathcal{I}(x_b, y') \quad (7)$$

where $\mathcal{L}_{mi}$ is the mutual-information-based regularization loss which is the sum of cross entropy loss ($\mathcal{L}_e(y, f(x))$) and a penalty term ($\lambda\mathcal{I}(x_b, y')$), where λ is the regularization weight coefficient (we consider $\lambda = 0.01$), and $\mathcal{I}(x_b, y')$ is the mutual information (Appendix B2).

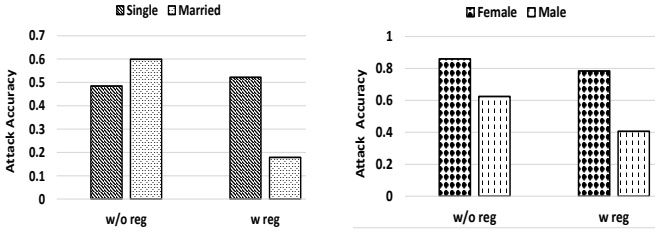
2) Fairness Constraints: Since disparate vulnerability is attributed to the distributional generalization gap in the target model's property functions (e.g., loss function that computes distributional difference for models on input data and outputs a numeric vector) among different subgroups [15], we aim to investigate the role of this distributional gap in the case of MIAI disparity. *Fairness* is commonly used to reduce this distributional generalization gap among different subgroups. For example, let ϕ be the target model's (f_{tar}) property function, and x be the input vector. For two subgroups s_1 and s_2, the distributional gap in the target model ($\mathcal{D}_{gap}$) can be formally defined as the δ deviations (generic distance) between subgroups in terms of property function ϕ:

$$\mathcal{D}_{gap} = \delta(Pr[\phi(f_{tar}, x \in s_1)|s_1], Pr[\phi(f_{tar}, x \in s_2)|s_2]) \tag{8}$$

We consider two fairness constraints– i) demographic parity (DP), and ii) equalized odds (EO) [35]. We train the model with the commonly used Exponentiated Gradient reduction algorithm [36] for classification with fairness. For the experiment, we use the accuracy score metric as a measure of the gap. DP reduces the distributional gap among subgroups, ensuring subgroups have a similar percentage of instances with each label. In contrast, EO reduces the distributional gap by minimizing true positive and false positive rates' differences among subgroups [37].

3) Disparity Mitigation Experiment Setup: In the mutual information regularization experiment, we apply regularization between subgroup attribute x_b (e.g., marital) and label attribute y (e.g., income) considering the DNN target models trained on Adult and NLSY datasets. We train the target model with regularization and iteratively reduce the loss function defined in Section V-D1. In our implementation, we have modified the default loss function provided by the Scikit-learn ML library. We have computed (between x_b and y) another penalty term (as illustrated in Eqn. 7), and added it with the default cross-entropy loss function in Scikit-learn. We perform SDMIA to compute subgroup vulnerabilities after target model training with regularization. We present the results in Section V-D4.

For the fairness experiment, we experiment with one subgroup as *control* (e.g., *white* in *race* attribute in the Adult dataset) with a fixed number of instances, and change the number of other subgroup's (*target*) instances. For each subgroup size, we randomly select *target* (e.g., *black*) subgroup instances (and repeat the experiment three times). We train corresponding DT target models (three models) based on these subgroup sizes, perform SDMIA on each model, and compute the mean of subgroup vulnerabilities across these three models. For binary subgroups (e.g., *sex*), we calculate $\mathcal{V}_{ds}(\mathcal{A})$ from subgroup vulnerabilities using our proposed disparity metrics. For multi-valued subgroups (e.g., *race*), we use 1-way ANOVA with a pairwise t-test to identify significant subgroups before calculating $\mathcal{V}_{ds}(\mathcal{A})$. Since we perform multiple comparisons, we also apply Benjamini-Hochberg *p-value* correction [38]. In Section V-D5, we present the results.



(a) x_b: *marital*, x_s: *sex* (b) x_b: *sex*, x_s: *marital*

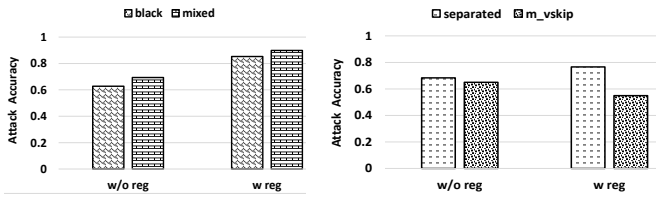
Fig. 6: Comparison between *marital* and *sex* subgroup attack accuracies with and without MI regularization, while estimating *sex* and *marital* sensitive attributes, respectively, in Adult dataset.

4) Disparity Mitigation via Mutual Information (MI) Regularization: In this section, we compare subgroup vulnerabilities, disparity, and target model utilities between two cases– target models trained (i) with and (ii) without MI regularization. We consider the setup discussed in Section V-D3.

Subgroup Vulnerabilities $\mathcal{V}_s(\mathcal{A})$. For experiments with binary subgroups, we consider *sex* and *marital* attributes in the Adult dataset. In Fig. 6a, we present the subgroup vulnerabilities of individuals with a different marital status where the adversary estimates the *sex* attribute. After MI regularization, *single* subgroup's vulnerability is increased to 53.23% from 48.46%, whereas for *married* subgroup, it is reduced to 46.56% from 59.94%. In contrast, as shown in Fig. 6b, *male* and *female* subgroups have 62.43%, and 85.90% vulnerabilities without regularization whereas with regularization both are reduced to 40.68% and 78.43%, respectively.

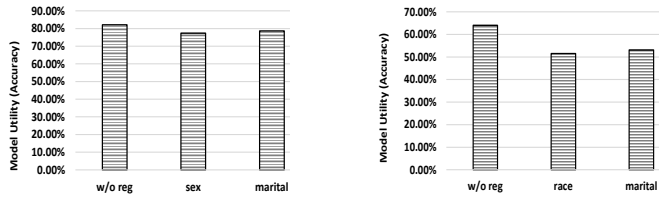
For multi-valued subgroups, we consider *race* and *marital* attributes in the NLSY dataset. From ANOVA test for *race* attribute, we obtain *mixed* and *black* subgroups as the most significant pair (Table XIV). Similarly, for *marital* attribute, *separated* and *m_vskip* represent the most significant pair (Table XV). In Fig. 7a and 7b, we present the comparisons among significant *race* and *marital* subgroups, respectively. With MI regularization, both subgroups' vulnerabilities in *race* increase (Fig. 7a). In contrast, *marital* subgroups show a different pattern. While the *separated* subgroup's vulnerability increases to 76% from 68%, it is reduced to 55% from 65% for *m_vskip*. We conclude that MI regularization does not consistently reduce subgroup vulnerabilities in MIAI attacks.

Disparity $\mathcal{V}_{ds}(\mathcal{A})$. We measure disparity using our proposed AMD and MMDD metrics, as presented in Section V-B. Table VIII shows the computed disparity values. In the Adult dataset, for *marital*, both AMD and MMDD values increase with MI regularization. For example, AMD is increased to 34.33 from 11.48, and MMDD is increased to 16.46 from 5.97 after MI regularization, which we also observe in Fig. 6a. A similar pattern can be observed in *sex* subgroups in the Adult



(a) x_b: *race*, x_s: *drug_marijuana* (b) x_b: *marital*, x_s: *drug_marijuana*

Fig. 7: Comparison between significant *race* and *marital* subgroup attack accuracies with and without MI regularization, while estimating *drug_marijuana* sensitive attribute in NLSY dataset.



(a) Adult target model utility (b) NLSY target model utility

Fig. 8: Comparison among target model utility with and without MI regularization in Adult and NLSY datasets.

dataset. For multi-valued subgroups in the NLSY dataset, i.e., both *marital* and *race* subgroups, AMD and MMDD values show a mixed trend. Although AMD in *marital* increases from 3.33 to 21.66, MMDD decreases from 4.73 to 1.12. In contrast, AMD in *race* decreases from 6.62 to 4.59 while MMDD increases from 5.79 to 7.89. Experiments from both binary and multi-valued subgroup attributes show that disparity in MIAI is not consistently reduced with MI regularization. As illustrated in [16], a highly accurate model is more vulnerable to MIAI attacks. While regularization reduces overall MIAI attacks by marginally reducing target model accuracy, it does not ensure each subgroup's accuracy is reduced consistently, thereby causing subgroup vulnerabilities to change in different margins and impacting disparity.

TABLE VIII: Performance comparisons between models with and without MI regularization, using proposed *disparity* metrics (AMD, MMDD)

Dataset	Subgroup Attribute	AMD w/o reg	AMD w reg	MMDD w/o reg	MMDD w reg
Adult	$marital$	11.48	34.33	5.97	16.46
NLSY	$race$	6.62	4.59	5.79	7.89
Adult	sex	23.47	37.74	15.82	25.44
NLSY	$marital$	3.33	21.66	4.73	1.12

Target Model Utility. In Fig. 8a, we present the model utility with and without MI regularization on *sex*, and *marital* attributes in the Adult dataset. With regularization, the accuracy score is slightly reduced to 77.43% and 78.62% from 82.11% in the cases of *sex* and *marital* subgroup attributes, respectively. In Fig. 8b, we present model utility with and without regularization on *race*, and *marital* attributes in the NLSY dataset. Again, with regularization, the target model accuracy score is slightly reduced. This indicates MI regularization does not significantly worsen target model performance.

5) Disparity Mitigation via Fairness Constraints: We experiment with *sex* and *race* subgroup attributes, where the adversary estimates the *marital* attribute in the Adult dataset. We follow the setup in Section V-D3.

Subgroup Vulnerabilities $\mathcal{V}_s(\mathcal{A})$. First, we consider *female* as the *target* subgroup with sizes 100, 500, 1k, 5k, and 10k and concatenate these with fixed 10k *male* instances representing the *control* subgroup. In Fig. 9, we present mean vulnerabilities from three models for each size of the target subgroup. We

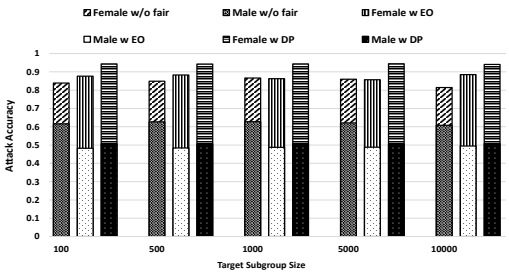


Fig. 9: Comparison between *sex* subgroups' vulnerabilities (*female* target) without fairness, with DP, and EO constraints, while estimating *marital* (Adult dataset). Lower and higher bars are two subgroup vulnerabilities, so the difference represents disparity.

observe that *male* subgroup vulnerability is reduced with both fairness constraints. In contrast, *female* subgroup vulnerability increases slightly with EO while significantly with DP. This trend remains consistent across different target subgroups (female) sizes. For example, in the case of size 500, *female* subgroup has vulnerability $\sim$94% and $\sim$87% with DP and EO, respectively, compared to $\sim$85% without any fairness constraints. Note that there is only a slight change in disparity with sample size in a model without fairness since disparity depends on the holistic impact of other factors. We perform a similar set of experiments on *sex* subgroup, with *male* as *target* subgroup and observe similar trends (Fig. 16 in Appendix).

For the multi-valued *race* subgroup, we consider *black* as

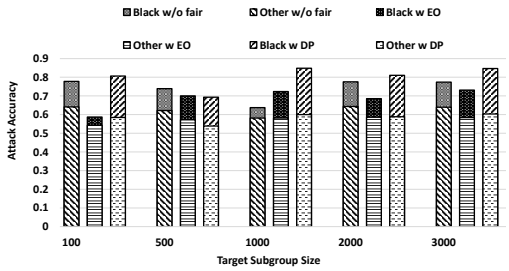


Fig. 10: Comparison between significant *race* subgroups' vulnerabilities without fairness, with DP, and EO constraints, while estimating *marital* (Adult dataset). Lower and higher bars are two subgroup vulnerabilities, so the difference represents disparity.

the *target* and *white* as the control subgroup, keeping *other* subgroup instances fixed. From the ANOVA test, we obtain *black* and *other* as the most significant pair (Table XIII). In Fig. 10, we present their vulnerabilities. Although in most cases *other* subgroup's vulnerability is reduced with fairness, it is not consistent. In contrast, *black* subgroup's vulnerability increases significantly with DP. For example, in size 1000, with DP, the *black* subgroup vulnerability is increased to 85% with fairness compared to 63% without fairness. Also, with EO, the vulnerability increases in some subgroup sizes, e.g., at 1000. Hence, we conclude that fairness does not consistently reduce subgroup vulnerabilities in MIAI attacks.

Disparity $\mathcal{V}_{ds}(\mathcal{A})$. In Fig. 9, we observe that the disparity $\mathcal{V}_{ds}(\mathcal{A})$ increases with either of the fairness constraints compared to without fairness scenario. Also, from our experiments, another observation is that disparity with DP is higher than disparity with EO (Fig. 13 and Fig. 14 in Appendix show the disparity in separate plots). Fig. 10 shows a similar trend with DP, i.e., the disparity is increased compared to the model without fairness in most cases, except for some outliers. Also, in Fig. 16, in most cases, disparity also increases with either fairness constraint. In summary, fairness constraints do not consistently reduce disparity in MIAI. Note that fairness equalizes subgroups training samples by reducing the distributional gap. However, subgroup vulnerabilities in MIAI attacks are estimated on the attack model, where the training set contains other features than that one equalized by fairness constraints (i.e., label attribute). Therefore, unlike membership attack [15], fairness in MIAI does not guarantee consistent disparity mitigation.

Target Model Utility. We compare target model utility with and without fairness constraints, considering *female* as the *target* subgroup. We present the comparisons among model utilities (in accuracy metric) in Fig. 11. It shows that the overall model utility is mostly consistent with varying subgroup sizes. Also, utility is slightly reduced in models trained with EO compared to a model without fairness, which is further reduced with DP. We empirically show that fairness constraints do not significantly compromise target model utility.

VI. Disparity Mitigation

Our empirical evaluations show that MI regularization and fairness techniques are not consistently effective in reducing *disparity* in MIAI (Sections V-D4 and V-D5). While MI regularization aims to reduce mutual information, fairness reduces the distributional gap, i.e., overfitting. This indicates there exist other disparity factors that contribute toward MIAI disparity. Therefore, we extend our study to investigate other potential MIAI disparity factors and their contributions, (i.e., importance/weights) towards subgroup vulnerability and disparity in MIAI; and ways to mitigate disparity.

A. Disparity Factors

To identify viable disparity factors, we consider two types of factors– (i) Training dataset-based factors (7): mutual information, correlation, marginal prior, skewness, kurtosis, covariance, and standard deviation, and (ii) Target model-based factors (1): overfitting. We first obtain disjoint training samples for a particular subgroup and then compute these factors using different statistical formulations. For example, we use Shannon's entropy [39] to compute mutual information between subgroup and label. Statistical factors like skewness and kurtosis are computed on the training data label distribution (disjoint subgroup samples) since the label (target model) has the highest importance in MIAI attack (Fig. 1b). Section B (Appendix) describes the details of computing these factors.

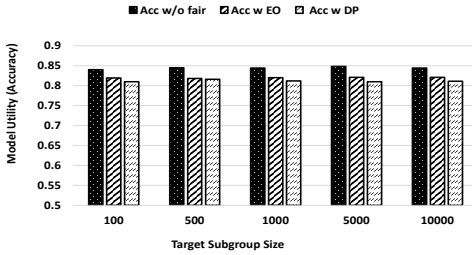


Fig. 11: Target model utility (accuracy) comparison without fairness, with DP, and with EO constraints in Adult dataset.

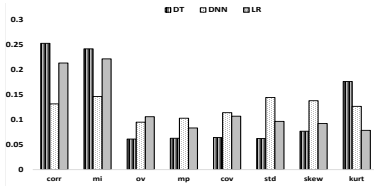


Fig. 12: Disparity factors' weights in Texas dataset, varying target models. * corr=correlation, mi=mutual information, ov=overfitting, mp=marginal prior, cov=covariance, std=standard deviation, skew=skewness, kurt=kurtosis.

B. Disparity Factors' Weight Analysis

We further experiment to identify the weights/importance of these disparity factors towards disparate vulnerability in MIAI. For this analysis, we compute disparity factors for all subgroups in a dataset (discussed in Section VI-A) and vulnerabilities of all subgroups (applying SDMIA) considering a target model (e.g., DT or DNN, or LR). We then fit the data in a linear regression model, where the disparity factors are inputs, and corresponding subgroup vulnerabilities are outputs. We obtain disparity factors' coefficients from the linear regression model. Disparity factors' weights are calculated from the coefficients of the factors in terms of factors' impact on the dependent variable (subgroup vulnerability $\mathcal{V}_s(\mathcal{A})$), i.e., odds ratio [40]. For example, let val_{coef} be a factor's coefficient, then we capture its impact on the label as follows:

$$\begin{aligned} logR_1 &= logR_0 + val_{coef} \\ \implies R_1 &= R_0 * e^{val_{coef}} \end{aligned} \tag{9}$$

where R_1 is the odds of label attribute $\mathcal{V}_s(\mathcal{A})$ with a unit change of the input factor value, and R_0 is the odds without a unit change of input factor value (initial). For each factor, we

calculate this odds ratio to identify the impact of that factor on subgroup vulnerability with a unit change of its value. Finally, we compute the importance of each factor (weight) in terms of the normalized odds ratio.

C. Impact of Target Model on Disparity Factor Weight

We compute the disparity factors' weights varying the target models. In Fig. 12, we present the results on the Texas dataset. We observe that the target model architecture significantly impacts disparity factor weight/importance. Our analysis shows that the correlation has the highest weight for the DT target model. This can be explained by the fact that the correlation between inputs and label determines optimal split criteria in DT, thus positively impacting subgroups' accuracy in the target model and hence the disparity in MIAI attack. Mutual information has a slightly higher weight than the correlation for both LR and DNN target models. This is because mutual information inherently captures layered model internal properties/parameters (invariant to parameterization) and can interpret the entire learning process, thereby better-capturing subgroups' disparate behavior as well. We observe similar patterns in the Adult dataset, presented in Fig. 17.

D. Disparity Mitigation in MIAI Attacks

According to our experiments, no single factor controls disparity in MIAI. More specifically, correlation and mutual information seem more important, although depending on data distribution, other factors might have significant weights as well (Fig. 12, Fig. 17) and holistically impact disparity. Therefore, to mitigate disparity in MIAI, it is important to reduce multiple factors, i.e., adopting a multi-factor-based disparity mitigation technique. One possible direction is to train the target models with a weighted multi-factor regularization strategy. It involves first identifying possible disparity factors' weights in the training data distribution. Then apply regularization on each factor, and adjust the overall loss function with a weighted regularization (factors' weights) penalty for all factors (similar to Eqn. 7). A more generic approach might be just considering the top few disparity factors and training target models with regularization on them (e.g., mutual information and correlation regularization). This might reduce the target model performance gap among subgroups, contributing to attack model disparity mitigation. Another possible mitigation can be guided adversarial training, where the victim can craft adversarial examples to change training data distribution. Tuning the training data distribution can regulate disparity factors' weight and hence impact disparity in the MIAI attack. However, this scenario might have a trade-off between target model performance and disparity.

VII. Discussions and Limitations

k vs. MIAI Performance. In general, for both $k = 2$ and $k > 2$, our proposed SDMIA with the least adversarial capabilities (even with 50% or 25% fewer queries) performs comparably or even better in some metrics in some scenarios (e.g., Tables V-VII, Tables XVI-XXI) compared to the existing MIAI attacks. As discussed in Section IV-B, we make a realistic assumption that $k <$ number of classes (label) and empirically evaluated SDMIA on 3 different real-life datasets with different k values (tested up to 4). Our evaluations show that all attack performances degrade with increasing k, as expected. However, existing attacks do not capture input output dependency as effectively as SDMIA and, as a result, suffer more (e.g., $\sim$ 56% and $\sim$ 50% accuracy drop in FJRMIA and SDMIA from $k = 2$ to $k = 4$).

Performance Metrics vs. MIAI Attacks. Class imbalance impacts MIAI success rate if measured with accuracy (discussed in Section V-B). FPR is highly influenced by marginal prior. Hence, FJRMIA fluctuates more in FPR. Class imbalance does not affect G-mean and MCC (used in SDMIA and LOMIA). Attack stability (i.e., IQR/error bars) depends on training data distribution (e.g., similarity), and FJRMIA/LOMIA suffer most from this measure, as discussed in Section IV-C4.

Adaptive Algorithm. Algorithm 1 re-randomizes d_{max} attributes randomly in new record generation. Leveraging confidence scores returned by target models during the previous query to craft new records might make the algorithm more adaptive. However, this also widens assumptions on adversarial capabilities (i.e., confidence score) and might not be realistic. Therefore, we want to investigate designing a more adaptive algorithm in future work.

Weak Correlation. If the correlation between the sensitive attribute and label is weak or no correlation exists, corresponding sensitive attribute is less or not at all vulnerable to an MIAI attack and our SDMIA generates fewer or no attack records.

Model Extraction and SDMIA. In model extraction attacks, the adversary queries the target model to reconstruct model parameters [34], [41]. In contrast, the adversary in SDMIA queries the target model to capture the influence of the sensitive attribute on the model. Although these two attacks have very different goals, both rely on the adversary's capability of querying the model and leveraging the returned predictions. In addition, as illustrated in Section IV-C4, in both attacks, there is a positive relationship between the number of queries and the attack performances.

VIII. Conclusions

This paper proposes a new black-box MIAI attack leveraging the least adversarial capabilities. We experimentally evaluate our proposed SDMIA with three target models trained on three datasets and show that they perform comparably to existing attacks for binary and multi-valued sensitive attribute estimation. We extend our study to investigate disparity in MIAI and show that two existing privacy defense techniques, *mutual information regularization* and *fairness constraints*, are not consistently effective in disparity mitigation. We also analyze potential disparity factors and their impacts on the disparity in MIAI. Finally, we discuss potential disparity mitigation techniques based on our experimental findings. Our work sheds light on the future of developing robust multi-factor-based defense to mitigate *disparity* in MIAI attacks.

REFERENCES

[1] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 2014, pp. 17–32.

[2] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," in *2021 IEEE symposium on security and privacy (SP)*. IEEE, 2021, pp. 730–747.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 1291–1308.

[6] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

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

[8] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[9] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "{ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[10] N. Z. Gong and B. Liu, "Attribute inference attacks in online social networks," *ACM Transactions on Privacy and Security (TOPS)*, vol. 21, no. 1, pp. 1–30, 2018.

[11] S. Mehnaz, S. V. Dibbo, E. Kabir, N. Li, and E. Bertino, "Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4579–4596.

[12] Amazon, "Amazon Fraud Detector API," https://docs.aws.amazon.com/frauddetector/latest/api/Welcome.html, 2003.

[13] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[14] Z. Lipton, J. McAuley, and A. Chouldechova, "Does mitigating ml's impact disparity require treatment disparity?" *Advances in neural information processing systems*, vol. 31, 2018.

[15] B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, and C. Troncoso, "Disparate vulnerability to membership inference attacks," *Proceedings on Privacy Enhancing Technologies*, vol. 1, pp. 460–480, 2022.

[16] T. Wang, Y. Zhang, and R. Jia, "Improving robustness to model inversion attacks via mutual information regularization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 666–11 673.

[17] U. M. L. Repository, "Adult dataset," http://archive.ics.uci.edu/ml/datasets/Adult, 1996.

[18] U. B. O. L. STATISTICS, "Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1997 cohort, 1997-2017 (rounds 1-18). Produced and distributed by the Center for Human Resource Research (CHRR), 2019," https://www.bls.gov/nls/getting-started/accessing-data.htm, 1997.

[19] W. Hickey, "Fivethirtyeight.com datalab: How americans like their steak," http://fivethirtyeight.com/datalab/how-americans-like-their-steak/, 2014.

[20] A. Marketplace, "Amazon Sentiment Analysis API," https://aws.amazon.com/marketplace/pp/prodview-szskhi4z2pohw, 2000.

[21] G. Cloud, "Google Cloud Vision API," https://cloud.google.com/vision/docs/drag-and-drop, 2016.

[22] Microsoft, "MS Azure Speech-to-text API," https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/rest-speech-to-text, 2020.

[23] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[24] S. I. Serengil, "Feature Importance in Decision Trees ," https://sefiks.com/2020/04/06/feature-importance-in-decision-trees/, 2020.

[25] Susmit, "Neural Feature Importance," https://towardsdatascience.com/neural-feature-importance-1c1868a4bf53, 2022.

[26] S. I. Serengil, "Feature Importance in Logistic Regression," https://towardsdatascience.com/neural-feature-importance-1c1868a4bf53, 2021.

[27] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. 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*, 2020, pp. 253–261.

[28] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 880–895.

[29] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International conference on machine learning*. PMLR, 2021, pp. 1964–1974.

[30] Z. Yang, E.-C. Chang, and Z. Liang, "Adversarial neural network inversion via auxiliary knowledge alignment," *arXiv preprint arXiv:1902.08552*, 2019.

[31] T. D. of State Health Services, "Texas Hospital Inpatient Discharge Public Use Data File, [first quarter 2006-10]. Texas Department of State Health Services, Austin, Texas. [May 18, 2011]," https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm, 2004.

[32] BigML, "Machine Learning made beautifully simple for everyone," https://bigml.com/, 2011.

[33] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation metrics for evaluating classification performance on imbalanced data," in *2019 international conference on computer, control, informatics and its applications (ic3ina)*. IEEE, 2019, pp. 14–18.

[34] V. Chandrasekaran, K. Chaudhuri, I. Giacomelli, S. Jha, and S. Yan, "Exploring connections between active learning and model extraction," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1309–1326.

[35] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[36] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[37] F. contributors, "Disparity mitigation with fairness," https://fairlearn.org/v0.7.0/user_guide/mitigation.html#fairness-constraints-for-binary-classification, 2018.

[38] D. Thissen, L. Steinberg, and D. Kuang, "Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons," *Journal of educational and behavioral statistics*, vol. 27, no. 1, pp. 77–83, 2002.

[39] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[40] M. Szumilas, "Explaining odds ratios," *Journal of the Canadian academy of child and adolescent psychiatry*, vol. 19, no. 3, p. 227, 2010.

[41] N. Carlini, M. Jagielski, and I. Mironov, "Cryptanalytic extraction of neural network models," in *Annual International Cryptology Conference*. Springer, 2020, pp. 189–218.

[42] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

APPENDIX

A. Synthetic Data Generation Algorithm for SDMIA

Algorithm presents the steps of the synthetic data generation.

B. Disparity Factors Computation

We compute all 8 disparity factors, either training data or target model related, contributing to the *MIAI disparity*.

Algorithm 1 Synthetic Data Generation Algorithm

Input: Sensitive attribute: $x_s \in \{x_1, ..., x_k\}$**, Non-sensitive attribute:** $x_2 \in \{x_{2,1}, ..., x_{2,ns}\}, ...x_d \in \{x_{d,1}, ..., x_{d,ns}\}$
Output: $Synthetic\ Dataset\ DS_A$

procedure $Data\ Synthesize(Sensitive : x_s)$
  $x^- \leftarrow RANDV(x_2...x_d)$ ▷ Initialize $d-1$ non-sensitive
  $j \leftarrow 0$ ▷ Counter
  $d_{max} \leftarrow c_{max}$ ▷ Max non-sensitive to change
  $d_{min} \leftarrow 1$ ▷ Min non-sensitive to change
  $r_{max} \leftarrow 2$ ▷ Max number of rejection
  for $iteration = 1\ to\ iter_{max}$ **do**
    $x_{\delta 1} \leftarrow x_1 \cup x^-$ ▷ Generate k records
    :
    $x_{\delta k} \leftarrow x_k \cup x^-$
    $y'_1 \leftarrow f_{tar}(x_{\delta 1})$ ▷ Query target with k records
    :
    $y'_k \leftarrow f_{tar}(x_{\delta k})$
    if $(y'_1 \neq null)$ & $...(y'_k \neq null)$ **then**
      if $y'_1 \neq y'_2 \ldots \neq y'_k$ **then**
        return $x_{\delta 1}, x_{\delta 2}, \ldots x_{\delta k}$ ▷ Sample Accepted
      else
        $j \leftarrow j + 1$ ▷ Sample Rejected, increment counter
        if $j \geq r_{max}$ **then** ▷ Update d_{max}
          $d_{max} \leftarrow max(d_{min}, \lceil d_{max}/2 \rceil)$
          $j \leftarrow 0$
        end if
      end if
    end if
    $x^- \leftarrow RANDV(x^-, d_{max})$ ▷ Re-randomize x^-
  end for
  return $null$
end procedure

TABLE IX: Sensitive attributes (multi-class) in Training datasets.

Dataset	Training Instances	Sensitive attribute	Classes
NLSY	5096	race	4
Adult	35222	education	3
Adult	35222	race	5
Texas	67443	severity_illness	5
FiveThirtyEight	331	age	4

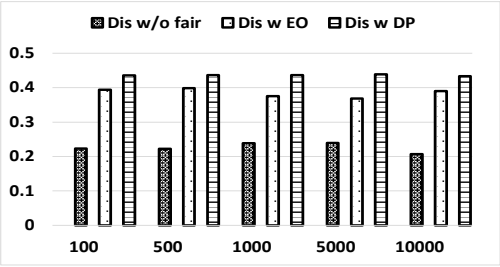


Fig. 13: Disparity in *sex* subgroups in Adult dataset (*female* target), while estimating *marital* sensitive attribute.

1) Correlation: We calculate correlations of each subgroup with label attributes in the training data. We first test the following hypotheses by performing the Chi-square (χ^2) test [42]. H_0 (null hypothesis)= S_0 and O_0 are independent, and H_a (alternate hypothesis)= S_0 and O_0 are correlated, where S_0 is the subgroup attribute and O_0 is the label, all variables are categorical. We consider significance level $\alpha = 0.05$. We present the categories in label attribute (*m*) and subgroup attributes (*n*) in a *m* x *n* contingency matrix, where $m \geq 2$, $n \geq 2$ and degrees of freedom $d_f = (m-1) * (n-1)$. We calculate the p-value applying the χ^2 test. If p-value≥ 0.05, we fail to reject the null hypothesis H_0. Otherwise, we reject H_0 and accept the alternate hypothesis H_a, indicating a significant correlation. If the test accepts H_a, we compute the correlation in terms of Cramer's values, as follows:

TABLE X: Abbreviation Table.

Abbreviation	Complete meaning
x_s	sensitive attribute
x_{ns}	non-sensitive attribute
d	total number of attributes
$\mathcal{I}mp_{AI}$	attribute Importance
$\mathcal{M}_A$	attribute Metric
$\mathcal{S}_{x_d}$	samples reaching to attribute node x_d
L_c	left Child
R_c	right Child
EO	equalized odds fairness constraint
DP	demographic parity fairness constraint
AMD	absolute maximum disparity
MMDD	maximal mean disparity deviation
DT	decision tree
LR	logistic regression
DNN	deep neural network
f_{tar}	target model
f_{adv}	attack model
DS_A	attack dataset
$\mathcal{V}_a(\mathcal{A})$	average vulnerability
$\mathcal{V}_s(\mathcal{A})$	subgroup vulnerability
$\mathcal{V}_{ds}(\mathcal{A})$	disparity
x_b	subgroup attribute
S_g	subgroup
$\mathcal{V}'_s(\mathcal{A})$	predicted subgroup vulnerability
d_{ta}	target instances
IQR	interquartile range
δ	generic distance
$\mathcal{H}$	Shannon's entropy
$\mathcal{I}$	mutual information
$\mathcal{L}_{mi}$	mutual inf regularization loss
val_{coef}	coefficient of a disparity factor
ϕ	property function
SDMIA	synthetic data-based MIAI attack

TABLE XI: Target model performances trained on NLSY dataset (tested on *training* dataset)

Model	Precision	Recall	Accuracy	F1 score
DT	43.41%	35.2%	44.72%	36.00%
DNN	32.01%	30.37%	33.48%	26.00%
LR	27.17%	26.03%	36.87%	24.00%

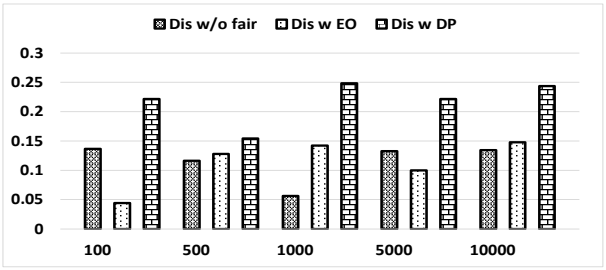


Fig. 14: Disparity in *race* subgroups in Adult dataset (*black* target), while estimating *marital* sensitive attribute.

TABLE XII: Target model performances trained on FiveThirtyEight dataset (tested on *training* dataset)

Model	Precision	Recall	Accuracy	F1 score
DT	95.18%	96.86%	96.38%	96.00%
DNN	15.31%	18.77%	8.46%	1.00%
LR	28.51%	25.82%	46.53%	22.00%

$$C_v = \sqrt{\frac{\chi^2}{[n\_total * (min(m,n) - 1)]}} \tag{10}$$

Where, *n_total* is the total instances, and C_v is Cramer's value. We define *correlation* of a subgroup by computing average Cramer's value with respect to other subgroups.

2) Mutual Information: Mutual Information (MI) captures the mutual dependency between two variables (i.e., subgroup and label attributes in our experiment). For each subgroup, we compute the mean of pairwise MI with respect to another subgroup, each time considering a different subgroup, and finally calculate the mean. We compute MI by calculating the difference between the entropy of label $\hat{Y}$, and conditional entropy of label $\hat{Y}$, given the subgroups attribute X, presented with Shannon's entropy [39] as below:

$$\mathcal{I}(X, \hat{Y}) = \mathcal{H}(\hat{Y}) - H(\hat{Y}|X) \tag{11}$$

Where, the conditional entropy $\mathcal{H}(\hat{Y}|X)$ can be formally defined as below from Shannon's entropy:

$$\mathcal{H}(\hat{Y}|X) = -\sum P(X, \hat{Y}) \log \frac{P(X, \hat{Y})}{P(X)} \tag{12}$$

where Shannon's entropy [39] is formally defined as:

$$\mathcal{H}(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{13}$$

where there are n possible outcomes $x_1, ..., x_n$ of the random variable X with the probability of them occurring being $P(x_1), ..., P(x_n)$.

3) Marginal Prior: Marginal prior is the subgroup probability of an attribute in the training data, derived as the ratio of the number of instances of that particular subgroup and total instances, denoted as follows:

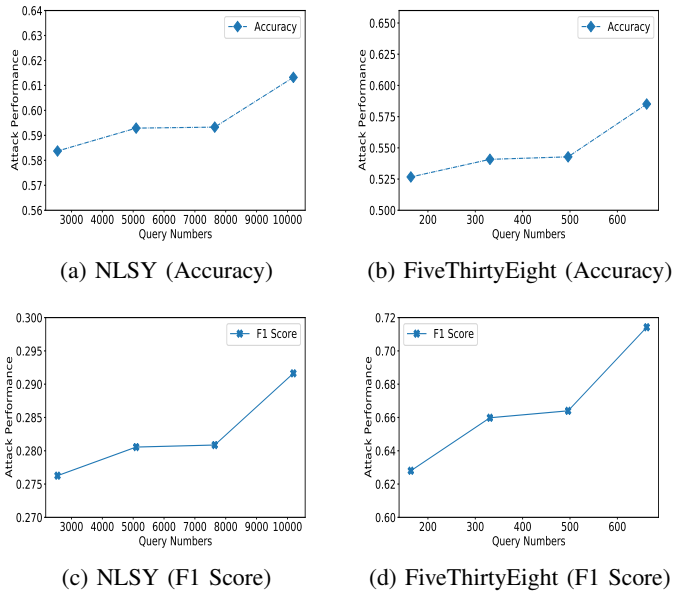
$$\mathcal{MP}_{class_1} = \frac{N_{class_1}}{N} \tag{14}$$

Where, $\mathcal{MP}_{class_1}$ is the marginal prior of a class $class_1$, N_{class_1} number of subgroup instances and N is total instances.

4) Distributional Overfitting: We consider the *overfitting* as a feature to identify the gap between model performance on training and test data (accuracy), defined as below:

$$\mathcal{OV}_{fit} = Acc_{tr} - Acc_{te} \tag{15}$$

where $\mathcal{OV}_{fit}$ is the overfitting of a subgroup, Acc_{tr} and Acc_{te} are prediction accuracy on training and test data.



(a) NLSY (Accuracy) (b) FiveThirtyEight (Accuracy)

(c) NLSY (F1 Score) (d) FiveThirtyEight (F1 Score)

Fig. 15: Average SDMIA attack performances, varying query numbers to the target model, while estimating *drug_marijuana* in the NLSY dataset (a, c), and *alcohol* in the FiveThirtyEight dataset (b, d).

5) Skewness: Skewness is a distributional property. It measures the deviation of the distribution from the symmetric normal distribution. We compute skewness as follows:

$$\mathcal{S}_y = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^3/N}{\sigma^3} \tag{16}$$

where $\mathcal{S}_y$ is the skewness of label y and N is the total number of instances, and $\bar{y}$ is the mean.

6) Kurtosis: Like skewness, kurtosis is a property of the distribution. Kurtosis measures the distributional peak heights with respect to the center. We compute kurtosis as follows:

$$\mathcal{K}_y = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^4/N}{\sigma^4} \tag{17}$$

where $\mathcal{K}_y$ denotes the kurtosis of label y. The higher the $\mathcal{K}_y$ means the distribution is highly tailed on the center.

7) Covariance: While *correlation* measures the strengths of association between two variables, *covariance* measures the direction of this association (positive if it moves in the same direction, otherwise, negative). We compute the *covariance* in our experiment between the subgroup and label attribute. We calculate the covariance using the following formula:

$$\mathcal{C}_{ov}(A, B) = \frac{\sum_{i=1}^{N}(A_i - \bar{A})(B_i - \bar{B})}{N} \tag{18}$$

Where $\mathcal{C}_{ov}(A, B)$ is the covariance, N is the total number of instances, A_i and B_i are instance values for the attributes, and their corresponding means are $\bar{A}$ and $\bar{B}$.

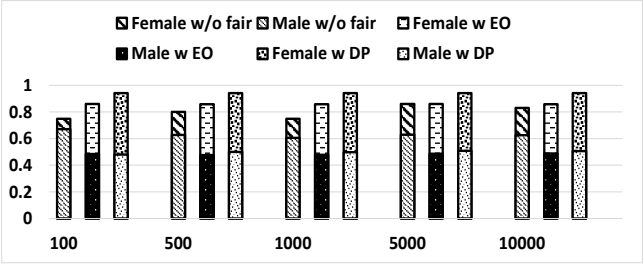


Fig. 16: Comparison between *sex* subgroups (*male* target) vulnerabilities without fairness, with DP, and EO constraints, while estimating *marital* (Adult dataset). lower and higher bars are two subgroup vulnerabilities, so the upper half of each bar is disparity.

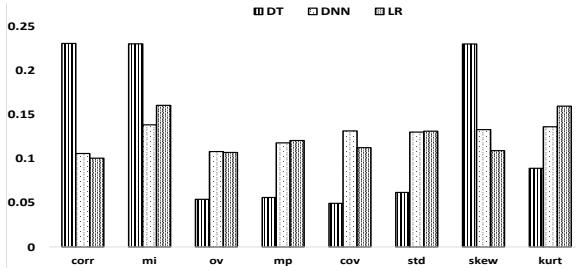


Fig. 17: Disparity factors' weights in Adult dataset, varying target models (DT or DNN or LR). *Notations: corr=correlation, mi=mutual information, ov=overfitting, mp=marginal prior, cov=covariance, std=standard deviation, skew=skewness, kurt=kurtosis.

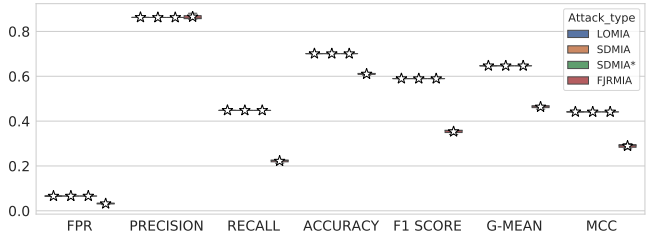


Fig. 18: Comparison of performance metrics distribution among the existing FJRMIA, LOMIA, and our proposed SDMIA obtained from 3 target models trained on Adult dataset estimating *marital* sensitive attribute (balanced marginal prior *married* positive class).

8) Standard Deviation: Standard deviation (σ) is the amount of variation of data samples from the expected value. We compute this on the label distribution of the subgroups. We calculate σ in terms of the mean μ, as follows:

$$\sigma_y = \frac{\sum_{i=1}^{N}(y_i - \mu)}{N} \tag{19}$$

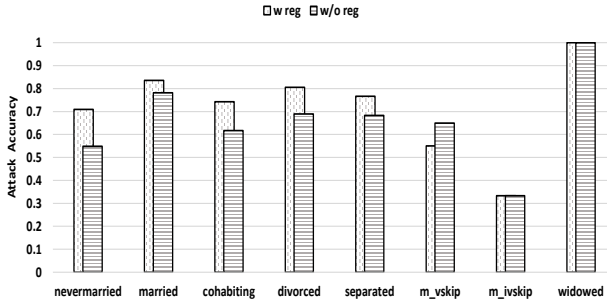


Fig. 19: Comparison among vulnerabilities of *marital* subgroups with and without MI regularization, while estimating *drug_marijuana* sensitive attribute in NLSY dataset.

TABLE XIII: *p-value* on ANOVA test & pairwise t-test with correction for *race* subgroups in Adult dataset.

p-value	White	Asian-Pac-Islander	Amer-Indian-Eskimo	Other	Black
White	1.000000	1.000000	0.535602	0.044286	0.000837
Asian-Pac-Islander	1.000000	1.000000	0.187795	0.004859	0.000837
Amer-Indian-Eskimo	0.535602	0.187795	1.000000	1.000000	0.000246
Other	0.044286	0.004859	1.000000	1.000000	0.000037
Black	0.000837	0.000837	0.000246	0.000037	1.000000

TABLE XIV: *p-value* on ANOVA test & pairwise t-test with correction for *race* subgroups in NLSY dataset.

p-value	nonblhis	black	hispanic	mixed
nonblhis	1.000000	1.000000	0.035558	0.002349
black	1.000000	1.000000	0.020586	0.002349
hispanic	0.035558	0.020586	1.000000	0.259587
mixed	0.002349	0.002349	0.259587	1.000000

TABLE XV: *p-value* after ANOVA test and pairwise t-test with Benjamini-Hochberg correction on *marital* subgroups in NLSY dataset.

p-value	nevermarried	married	cohabiting	divorced	separated	m_vskip	m_ivskip	widowed
nevermarried	1.000000	1.000000	0.96816	1.000000	0.067627	0.461070	1.000000	1.00000
married	1.000000	1.000000	1.00000	1.00000	0.461070	11.000000	1.000000	1.000000
cohabiting	0.968160	1.00000	1.000000	1.000000	1.00000	0.104640	1.000000	0.96816
divorced	1.000000	1.000000	1.000000	1.000000	0.897037	0.461070	1.000000	1.000000
separated	0.067627	0.46107	1.00000	0.897037	1.000000	0.002724	0.150231	0.46107
m_vskip	0.461070	1.00000	0.10464	0.46107	0.002724	1.000000	0.461070	1.000000
m_ivskip	1.000000	1.000000	1.0000006	1.000000	0.150231	0.461070	1.000000	1.000000
widowed	1.000000	1.000000	0.96816	1.000000	0.461070	1.000000	1.000000	1.000000

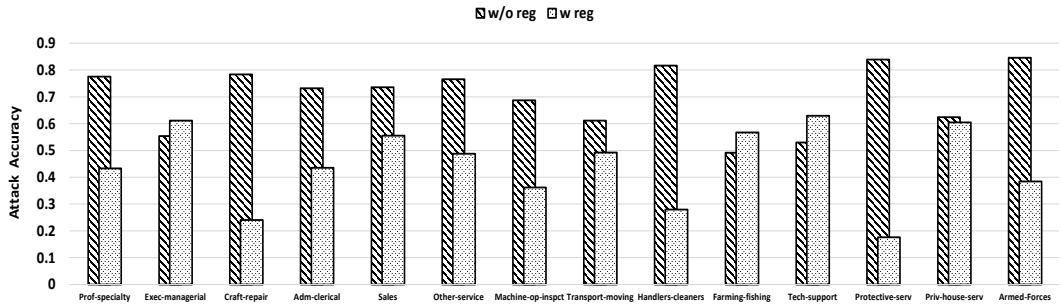


Fig. 20: Comparison between vulnerabilities of *occupation* subgroups with and without MI regularization, while estimating *marital* sensitive attribute in Adult dataset.

TABLE XVI: Attack performance while inferring drug_marijuana sensitive attribute, against LR target model trained on NLSY dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	10192	57	3966	169	904	25.22%	5.93%	78.94%	9.60%	23.85%	3.50%
LOMIA [11]	10192	523	2249	1886	438	21.71%	54.42%	54.40%	31.04%	54.41%	6.90%
SDMIA	10192	604	2421	1714	357	26.06%	62.85%	59.36%	36.84%	60.66%	16.81%
SDMIA*	5096	545	2268	1867	416	22.60%	56.71%	55.20%	32.32%	55.77%	9.06%

TABLE XVII: Attack performance while inferring drug_marijuana sensitive attribute, against DNN target model trained on NLSY dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	10192	3	4126	9	958	25.00%	0.31%	81.02%	0.62%	5.58%	0.76%
LOMIA [11]	10192	343	2680	1455	618	19.08%	35.69%	59.32%	24.86%	48.10%	0.41%
SDMIA	10192	184	3081	1054	777	14.86%	19.15%	64.07%	16.73%	37.77%	−5.79%
SDMIA*	5096	262	2807	1328	699	16.48%	27.26%	60.22%	20.54%	43.02%	−4.10%

TABLE XVIII: Attack performance while inferring *alcohol* sensitive attribute, against LR target model trained on FiveThirtyEight dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	662	266	0	65	0	80.36%	100.00%	80.36%	89.11%	0.00%	0.00%
LOMIA [11]	662	169	31	34	97	83.25%	63.53%	60.42%	72.07%	55.05%	9.16%
SDMIA	662	169	31	34	97	83.25%	63.53%	60.42%	72.07%	55.05%	9.16%
SDMIA*	331	169	31	34	97	83.25%	63.53%	60.42%	72.07%	55.05%	9.16%

TABLE XIX: Attack performance while inferring *alcohol* sensitive attribute, against DNN target model trained on FiveThirtyEight dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	662	256	3	62	10	80.50%	96.24%	78.25%	87.67%	21.08%	1.75%
LOMIA [11]	662	0	65	0	266	0%	0%	19.64%	0%	0%	0%
SDMIA	662	154	23	42	112	78.57%	57.89%	53.47%	66.67%	45.26%	−5.43%
SDMIA*	331	136	29	36	130	79.07%	51.13%	49.85%	62.10%	47.76%	−3.38%

TABLE XX: Attack performance while inferring *marital* sensitive attribute, against LR target model trained on Adult dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	70444	3861	17697	632	13032	85.93%	22.86%	61.21%	36.11%	46.98%	29.06%
LOMIA [11]	70444	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%
SDMIA	70444	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%
SDMIA*	35222	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%

TABLE XXI: Attack performance while inferring *marital* sensitive attribute, against DNN target model trained on Adult dataset

Attack Strategy	Query #	TP	TN	FP	FN	Precision	Recall	Accuracy	F1 score	G-mean	MCC
FJRMIA [3]	70444	3592	17717	612	13301	85.44%	21.26%	60.50%	34.05%	45.34%	27.62%
LOMIA [11]	70444	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%
SDMIA	70444	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%
SDMIA*	35222	7565	17121	1208	9328	86.23%	44.78%	70.09%	58.95%	64.68%	44.12%