On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques

Yan Zhou * Murat Kantarcioglu* Chris Clifton †

Abstract

Biased AI models result in unfair decisions. In response, a number of algorithmic solutions have been engineered to mitigate bias, among which the Synthetic Minority Oversampling Technique (SMOTE) has been studied, to an extent. Although the SMOTE technique and its variants have great potentials to help improve fairness, there is little theoretical justification for its success. In addition, formal error and fairness bounds are not clearly given. This paper attempts to address both issues. We prove and demonstrate that synthetic data generated by oversampling underrepresented groups can mitigate algorithmic bias in AI models, while keeping the predictive errors bounded. We further compare this technique to the existing state-of-the-art fair AI techniques on five datasets using a variety of fairness metrics. We show that this approach can effectively improve fairness even when there is a significant amount of label and selection bias, regardless of the baseline AI algorithm.

Keywords— AI fairness, sensitive feature, synthetic data, SMOTE

1 Introduction

AI algorithms are being criticized for reflecting and potentially exacerbating human biases in data. Biased AI models perpetuate inequalities in job hiring, credit lending, health care, predictive policing, and criminal sentencing [34]. In response, many bias mitigation techniques have been developed to improve fairness (e.g., [14, 41, 6, 22, 30, 13, 42, 25, 26, 7, 24, 33, 20, 23, 28]), among which synthetic minority oversampling techniques (e.g. fair-SMOTE) have been shown very effective [8, 19, 39, 29]. However, there lacks theoretical justification for the successful use of SMOTE for de-biasing, as well as clear error/fairness bounds. In this paper, we attempt to address both issues.

SMOTE randomly fabricates a new sample along the line segment between an instance x and one of its random neighbor $x^{(k)}$ [9] in the feature space. SMOTE-based de-biasing techniques directly address data bias, thus can be considered as pre-processing. Other types of pre-processing based de-biasing techniques focus on transforming a given input by editing its features and labels, assigning weights to selected training samples, or learning a latent representation excluding sensitive features to increase fairness of the trained model. The consensus is that the cause of algorithmic bias lies within the data, favoring the privileged group as decisions are made. This interpretation of bias is imperfect, but it simplifies and allows for the formalization of incorporating fairness into AI models from upstream in the machine learning pipeline.

Despite its clear efficacy, there is a weak understanding of how SMOTE helps improve fairness. Most SMOTE-based de-biasing techniques are given credits for their empirical success that relies on hunches such as a more balanced subgroup and a more authentic representation of real data distribution. The main question to investigate is whether synthetic data in the SMOTE style can in fact bridge the gap between the distributions of the (privileged and unprivileged) subgroups, and thus help in addressing label bias (e.g. human errors, normally introduced into data as a result of human decisions made in a certain historical context) and selection bias (e.g. linguistic data favoring English, an unexpected correlation between demographic attributes and decision output when data is sampled) associated with the original data. Resolving bias at the data level frees us from specific definition and mathematical notations of fairness. Furthermore, with synthetic data, we do not count on data surveillance to collect more data at the expense of risking individual privacy.

The main contributions of this paper include: 1.) Theoretical justification for the SMOTE-based de-biasing techniques; 2.) Empirical evidence that synthetic data generated in SMOTE fashion can address label bias and selection bias; 3.) Extensive empirical studies that compare the SMOTE-based de-biasing technique with the existing pre-processing, in-processing, and post-processing techniques using a variety of fairness metrics.

The rest of the paper is organized as follows: Section 2 discusses existing related work. Section 3 presents the problem and our theoretical results. Section 4 presents our SMOTE-based de-biasing algorithm. Section 5 presents the experimental results, and Section 6 concludes our work.

*Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080, {yan.zhou2, muratk}@utdallas.edu

†Department of Computer Sciences, Purdue University, West Lafayette, Indiana, 47907, clifton@cs.purdue.edu

2 Related Work

De-biasing techniques progress in three directions: pre-processing, in-processing, and post-processing. Meanwhile, many fairness metrics have been proposed [40, 17], implying the interlocking complexity of the fair AI problem.

De-biasing through pre-processing typically transforms training data by reducing the influence of demographic changes on the positive base rate. Feldman et al. [14] propose to alter the unprotected attributes to remove disparate impact. Zemel et al. [41] encode the input with a latent representation that obfuscates sensitive attributes. Optimized preprocessing [6] learns a probabilistic transformation for the input to improve group fairness while limiting individual data distortion. Reweighing [22] assigns different weights to selected samples to ensure fairer predictions by trained classifiers. In-processing de-biasing is done at the algorithmic level where learning algorithms are tweaked to ensure fairness. Some de-biasing techniques coupled with adversarial objectives consider learning fair representation under the constraint of different adversarial objectives for group fairness [30, 13]. De-biasing with adversarial learning [42] aims to maximize predictive accuracy while minimizing adversary's ability to predict sensitive attributes. Other in-processing de-biasing techniques focus on fairness constraints on structured subgroups [25, 26], training an optimized classifier with respect to a given fairness metric [7], training a classifier with data augmentations that deliberately manipulate subgroup features [18], or adding a regularization term to the objective against discrimination [24]. Post-processing techniques modify output labels to meet different fairness objectives. Some calibrates classifier outputs to ensure equalized odds [33, 20], some makes favorable predictions for unprivileged groups and unfavorable predictions for privileged groups in the vicinity of decision boundaries [23], and in the case where only black-box access is granted, a classifier satisfying multi-accuracy fairness conditions can be learned to improve fairness and subgroup accuracy [28].

Accuracy-fairness trade-off optimization is a popular algorithmic treatment for bias in current de-biasing techniques [10, 15]. In addition, there is an active line of work on optimizing Pareto fairness of the problem [31, 2]. Breugel et al. [36] recently proposed a GAN structure with causal knowledge to generate fair synthetic data from unfair data. Features are generated sequentially according to the given causal structure while biased edges are removed to achieve fairness. The technique is not applicable to fairness definitions established directly on downstream models. Causal fairness-aware GAN (CFGAN) [37] also attempts to create fair synthetic data from given causal knowledge. The technique is designed for a single protected attribute. FairGAN [38] generates synthetic data following the distribution of the real data while ensuring there is no correlation between synthetic data and the protected attribute. FairGAN improves fairness, however, it may lead to significant accuracy loss [36].

SMOTE is one of the most widely used approaches to oversampling the minority class, and has recently been reported to synthesize new samples with excellent utility and good privacy compared to other synthetic data generators [32]. SMOTE-based oversampling has also been shown an effective technique for bias mitigation [8, 19, 39, 29]. The Fair-SMOTE algorithm balances subgroup data distributions so that the privileged and the unprivileged groups have an equal number of positive and negative instances [8]. This method has been demonstrated to be effective at reducing bias and achieve higher classification performance than the state-of-the-art fairness algorithms. The parameterised data sampling method has been shown to produce fairness-optimal predictions with a small loss in predictive power [19]. A K-Means SMOTE-based algorithm has also been proposed to enhance model fairness and prediction accuracy [39]. Although existing SMOTE-based fairness algorithms have demonstrated promising fairness outcome, little is known about its working principle in theory. Furthermore, existing literature lacks an extensive study comparing SMOTE-based approaches to the state-of-the-art pre-processing, in-processing, and post-processing fairness algorithms using a wide range of fairness metrics. This paper aims to fill such gaps. We provide theoretical support for synthetic oversampling that helps bridge the gap between the distributions of subgroups in data. We also provide extensive and comparative empirical study on its effectiveness.

3 Problem Statement and Theoretical Results

We now formally define the problem and provide theoretical justification for SMOTE-based fairness algorithms.

3.1 Preliminaries

Target Error: Given the underlying data distribution D, the target error $\epsilon_D(h)$ is:

$$\epsilon_D(h) = \mathbb{E}_{x\sim D}[|h(x) - f(x)|],$$

The error measures the expected difference between the output of the hypothesis h and the target function f for a given instance x.

Favorable Class: Given a privileged group $\mathcal{X}_p$ and an unprivileged group $\mathcal{X}_u$, the favorable class c^+ is the category in which $\mathcal{X}_p$ has a disproportionately larger probability:

$$Pr(c^+|X \subset \mathcal{X}_p) \gg Pr(c^+|X \subset \mathcal{X}_u).$$

Two-sample Test Statistic: Schilling developed a test statistic for a measure of the discrepancy between two distributions [35]. The general multivariate two-sample problem is based on the k nearest-neighbor-type coincidences:

$$T_{k,n} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{r=1}^{k}\mathbf{I}_i(r)$$

given independent random samples in $\mathbb{R}^d$ $\{x_1, \ldots, x_{n_1}\}$ and $\{x_{n_1+1}, \ldots, x_n\}$ from unknown distributions $F(x)$ and $G(x)$, where $\mathbf{I}_i(r)$ is the indicator function that measures to 1 if the r nearest-neighbor is in the same sample as x_i. Large values of the test statistics give evidence *against* the hypothesis of $F(x) = G(x)$.

$\mathcal{H}$-Divergence: Ben-David et al. [5] proposed the *$\mathcal{H}$-divergence* to make feasible measuring divergence between two distributions D and D' over domain $\mathcal{X}$:

$$\begin{aligned} d_{\mathcal{H}}(D, D') &= 2 \sup_{h \in \mathcal{H}} [Pr_D[I(h)] - Pr_{D'}[I(h)]] \\ &\geq 2|Pr_D[h(x) = 1] - Pr_{D'}[h(x) = 1]| \end{aligned}$$

where $\mathcal{H}$ is a hypothesis class of finite VC dimension on a domain $\mathcal{X}$, $I(h)$ is the set such that $\mathbf{x} \in I(h) \Leftrightarrow h(\mathbf{x}) = 1$.

3.2 Problem Statement Given a domain $\mathcal{X} \sim D$ and a target function $f : \mathcal{X} \to [0, 1]$, let D_p and D_u be the distributions over the privileged and the unprivileged subgroups in $\mathcal{X}$, and $\hat{D}_u$ be the distribution of the oversampled unprivileged group. We formalize de-biasing as bounding the discrepancy of the predictive errors over D_p and $\hat{D}_u$:

$$\Delta\epsilon(h) = \epsilon_{D_p}(h) - \epsilon_{\hat{D}_u}(h)$$

where $\epsilon_D(h)$ is the target error for $x \sim D$, and the discrepancy of fairness over D_p and $\hat{D}_u$:

$$\Delta\epsilon(h(x) = 1) = |Pr_{D_p}[h(x) = 1] - Pr_{\hat{D}_u}[h(x) = 1]|$$

This formalization is inspired by the concept of *transfer learning*: if two populations only differ in demographic and socioeconomic background, models trained on one population (e.g. white defendants) should be readily applicable to the other (e.g. black defendants), with bounded predictive errors and positive base rate difference. Thus, the de-biasing technique is designed to oversample the (un)privileged group so that the divergence of the two distributions D_p and D'_u is bounded, and consequentially the discrepancies of predictive errors and fairness measures are bounded.

3.3 Theoretical Results We present two theoretical results: 1.) SMOTE-based oversampling can bridge the gap between the distributions of the privileged and the unprivileged subgroups (Theorem 3.1); and 2.) the discrepancies of predictive errors and fairness measures over D_p and $\hat{D}_u$ are bounded (Theorem 3.2). The intuition behind Theorem 3.1 is when the entire population in the favorable (unfavorable) class is treated fairly, both the privileged and the unprivileged groups should follow a well-mixed distribution. When there is bias, generating synthetic data from existing "unprivileged favorable" or "privileged unfavorable" samples can reduce the Schilling statistic that measures the gap of the distributions between the privileged and the unprivileged groups. With the guarantee of a smaller two-sample distributional divergence through oversampling by Theorem 3.1, Theorem 3.2 provides an error bound given the distributional divergence between the two groups.

THEOREM 3.1. *Given input $X \sim D$, let $X_p \sim D_p$ denote the privileged subset $X_p \subset X$ following the distribution D_p, and similarly, $X_u \sim D_u$ be the unprivileged subset following D_u. $\hat{X}_u$, as a result of oversampling X_u in the favorable class c^+, reduces the Schilling test statistic:*

$$T_{k,n}(X_p, \hat{X}_u) < T_{k,n}(X_p, X_u).$$

Proof. Given the unprivileged subgroup X_u, SMOTE samples an instance $x_u \in X_u$ that belongs to the favorable class c^+. Instances in X_u that are the nearest neighbors of the selected sample x_u are used to generate the synthetic data to improve the underrepresented populations in the favorable class. Suppose m synthetic samples are generated such way and added to X_u. Without loss of generality, among these m samples, let $M_p = \{x_1, \ldots, x_{m_p}\}$ be the ones to which the privileged samples in X_p are nearer neighbors, as the red dot shown in Figure 1, and $M_u = \{x_{m_p+1}, \ldots, x_m\}$ has nearer neighbors from the unprivileged in X_u, as the black dot shown in Figure 1.

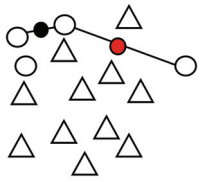


Figure 1: Synthetic samples that are nearer neighbors to the privileged and unprivileged, respectively.

Hence, $M = M_p \cup M_u$ essentially follows a mixture of D_u and D_p such that

$$P(x \in M) = \alpha_p \int_M f_p(x)dx + (1 - \alpha_p) \int_M f_u(x)dx$$

where α_p is the probability $\hat{x_u} \in M$ follows the privileged distribution, f_p and f_u are the density functions of the two distributions D_p and D_u. Let $\hat{X}_u = X_u \cup M_p \cup M_u$, then

$$T_{k,n}(X_p, \hat{X}_u) = \frac{1}{(n+m)k} \left[\sum_{i=1}^{n} \sum_{r=1}^{k} \mathbf{I}_i(r) + \sum_{i=n+1}^{n+m} \sum_{r=1}^{k} \mathbf{I}_i(r) \right],$$

in which the second term $\sum_{i=n+1}^{n+m} \sum_{r=1}^{k} \mathbf{I}_i(r)$ are associated with samples in M. The more biased the source data, the higher the probability α_p that a synthetic sample $x_i \in M$ is drawn from the privileged distribution since $P(NN_i(r) \in X_p | c^+) > P(NN_i(r) \in X_u | c^+)$, where $NN_i(r)$ is the r^{th} nearest neighbor of x_i. Therefore, $x_i \in \hat{X}_u$ is more likely from the well-mixed distributions of D_p and D_u, and the average number of neighbors of x_i coming from D_u is smaller:

$$\sum_{i=n+1}^{n+m} \sum_{r=1}^{k} \mathbf{I}_i(r) < \frac{m}{n} \sum_{i=1}^{n} \sum_{r=1}^{k} \mathbf{I}_i(r),$$

thus,

$$T_{k,n}(X_p, \hat{X}_u) < T_{k,n}(X_p, X_u).$$

□

This process of oversampling X_u from nearest neighbors is essentially searching for a reasonable set of $\hat{X}_u$ to bridge the gap between D_p and D_u so that the k nearest neighbors of $x \in X_p \cup \hat{X}_u$ is equally likely from either D_p or D_u, that is,

$$\lim_{n \to \infty} T_{k,n}(X_p, \hat{X}_u) = 0.5$$

In other words, X_p and $\hat{X}_u$ are more likely from the same well-mixed distribution. In an ideal case where X_p and $\hat{X}_u$ become identically distributed under $\hat{D}$, then:

$$\mathbb{E}(y|f(X_p) = h, x) = \mathbb{E}(y|f(\hat{X}_u) = h', x)$$

for $x \in X_p \cup \hat{X}_u$, assuming X_p and $\hat{X}_u$ are sufficiently large.

THEOREM 3.2. *Given a hypothesis h trained on data X, the differences of predictive errors and fairness measures on data $X_p \subset X$ from D_p and $X_u \subset X$ from D_u by h are bounded.*

Proof. Let $d(D_p, D_u)$ be the discrepancy of the two distributions D_p and D_u, then [3]:

$$\begin{aligned} d(D_p, D_u) = &\int ||x^p - x^u|| dD_p \otimes D_u(x^p, x^u) \\ &- \frac{1}{2}\int ||x_1^p - x_2^p|| dD_p \otimes D_p(x_1^p, x_2^p) \\ &- \frac{1}{2}\int ||x_1^u - x_2^u|| dD_u \otimes D_u(x_1^u, x_2^u) \geq 0 \end{aligned}$$

where $||\cdot||$ denotes the Euclidean norm of $\mathbb{R}^d$ and the equality holds if and only if $D_p = D_u$. x^p and x^u are two samples from D_p and D_u respectively, x_1^p and x_2^p are two samples from D_p, and x_1^u and x_2^u are two samples from D_u. Synthetic oversampling asymptotically reduces $d(D_p, D_u)$ according to Theorem 3.1. Given a hypothesis h trained on data $X \sim D$, the difference of predictive errors on data from D_p and D_u by h is bounded with respect to the divergence between D_p and D_u [5]:

$$\epsilon_{D_p}(h) - \epsilon_{D_u}(h) \leq d(D_p, D_u) + \lambda$$

where $\lambda = \arg\min_{h\in\mathcal{H}}[\epsilon_{D_p}(h) + \epsilon_{D_u}(h)]$, which is the combined error of the ideal joint hypothesis.

The divergence $d(D_p, D_u)$ cannot be accurately estimated from limited samples [16]. However, the discrepancy between favorable predictions by h on data from D_p and D_u is bounded by the $\mathcal{H}$-*divergence* between D_p and D_u:

$$|Pr_{D_p}[h(x) = 1] - Pr_{D_u}[h(x) = 1]| \leq \frac{1}{2} d_{\mathcal{H}}(D_p, D_u).$$

Kifer et al. [27] provide a theoretical bound for the true $\mathcal{H}$-*divergence* given any $\delta \in (0, 1)$ with probability at least $1 - \delta$ [5]:

$$d_{\mathcal{H}}(D, D') \leq \hat{d}_{\mathcal{H}}(U, U') + 4\sqrt{\frac{d\log(2m) + \log(\frac{2}{\delta})}{m}}$$

where U and U' are samples of size m from D and D'. As the sample size increases, the empirical $\mathcal{H}$-*divergence* asymptotically approaches the true $\mathcal{H}$-*divergence*.

With the concept of $\mathcal{H}$-*divergence*, we can bound the difference of predictive errors and the discrepancy of favorable predictions on data U_p and U_u of size m from D_p and D_u as follows:

(3.1)
$$\Delta\epsilon(h(x) = 1) \leq \frac{1}{2}\hat{d}_{\mathcal{H}}(U_p, U_u) + 2\sqrt{\frac{d\log(2m) + \log(\frac{2}{\delta})}{m}}$$

(3.2)
$$\Delta\epsilon(h) \leq \frac{1}{2}\hat{d}_{\mathcal{H}}(U_p, U_u) + 2\sqrt{\frac{d\log(2m) + \log(\frac{2}{\delta})}{m}} + \lambda$$

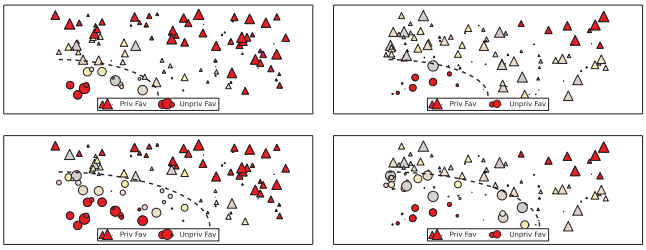
where $\Delta\epsilon(h) = \epsilon_{D_p}(h) - \epsilon_{D_u}(h)$, $\Delta\epsilon(h(x) = 1) = |Pr_{D_p}[h(x) = 1] - Pr_{D_u}[h(x) = 1]|$, and λ is the combined error of the ideal joint hypothesis as defined earlier. ∎

As (3.1) and (3.2) suggest, we can limit both the discrepancy of favorable predictions and the difference in predictive errors by making the distributions of the two groups D_p and D_u diverge less, especially when one group is underrepresented in terms of favorable prediction.

4 De-biasing with Synthetic Data

Given a demographic attribute $A = \{a_1, a_2\}$ in a dataset X, the majority group defined on A is $X_{A=a^*} \subset X$ if $Pr(A = a^*|X) > Pr(A \neq a^*|X)$ where $a^* \in \{a_1, a_2\}$. The privileged group defined on A is $X_{A=\hat{a}} \subset X$ if, historically, $Pr(y = 1|X_{A=\hat{a}}) > Pr(y = 1|X_{A\neq\hat{a}})$ where $\hat{a} \in \{a_1, a_2\}$, $y \in \{0, 1\}$ and $y = 1$ is favored. For example, in the COMPAS dataset, the black race is the majority and the white race is privileged. We investigate the following scenarios:

1.) When the majority is privileged, we oversample the unprivileged group with favorable labels (Figures 2a). We can also oversample the privileged group with unfavorable labels, but it tends to distort the prior distribution easily, hence is not recommend.
2.) When the minority is privileged, we oversample the privileged group with unfavorable labels (Figure 2b).



(a) Oversample the unprivileged favorable samples in the majority group.

(b) Oversample the privileged unfavorable samples in the minority group.

Figure 2: Oversampling strategies when the majority/minority of the population is privileged.

The idea of this de-biasing approach is to generate synthetic data to reduce the difference in the positive base rate between the privileged and the unprivileged populations. When more *underrepresented data is generated, the divergence between D_p and D_u is reduced*, and consequently we bridge the gap between the two groups with fairer favorable predictions. Method 1 and Method 2 generate synthetic data in the minority group for greater efficiency (smaller amount needed) and less corruption on the original data distribution. This sampling strategy may not be ideal. A more rigorous (expensive) way is to run a permutation test to identify where D_p and D_u differ most.

5 Experimental Results

In our experiment, synthetic (non-existing) data is generated from one group using a SMOTE [9] adaptive variant [21] (favoring low density region) until the other group is not disproportionately (dis)advantaged in the training set. The technique is naturally applicable to multi-class problems. Each experiment was run 10 times on Intel Xeon 2.30 GHz CPU with 256 GB memory, and we report the averaged results. Source code is available on Github[1].

Data Set We test our de-biasing technique on five data sets: *Adult*, *Compas*, *German Credit*, *Medical Expanse*, and *Bank* data [12]. These datasets represent the two general cases where the majority is privileged and the minority is privileged. Although the *Adult* dataset has been criticized for inadequate pre-processing [11], it presents a good example where model predictions are strongly correlated with the sensitive attributes in the data (the accuracy drops from 74% to 66% without the sensitive attributes). Each dataset was split into independent training (70%) and test (30%) sets.

Baseline Learning Algorithms We use *Logistic Regression* (LR), *Random Forest* (RF), *Support Vector Machine* (SVM), and *Neural Network* (NN) as the baseline learning algorithms. All hyper-parameters for baseline algorithms were tuned using grid search with cross validation.

De-biasing Algorithms We compare the oversampling technique (**syn**) to the vanilla baseline (**orig**) and a variety of mitigation algorithms, featuring sample reweighing, feature editing, and regularization. These algorithms include *Disparate Impact Remover* (**dir**, pre-processing) [14], *Reweighing* (**rew**, pre-processing) [22], *Prejudice Remover* (**pr**, in-processing) [24], *Exponentiated Gradient Reduction* (**egr**, in-processing) [1], *Calibrated EqOdds* (**cpp**, post-processing) [33], and *Reject Option* (**ro**, post-processing) [23]. We do not choose the *adversarial debiasing* technique [42] since it either fails to de-bias or suffers significant accuracy drop. De-biasing algorithms used for comparison are implemented in the IBM AI Fairness 360 library, licensed under the Apache License 2.0 [4].

Fairness Metrics We measure fairness using a number of individual and group fairness metrics, including *average odds difference*, *disparate impact*, *statistical parity difference*, *equal opportunity difference*, and *Theil index* [4].

5.1 Experiment without Injection of Label and Selection Bias

In this experiment, we do not inject either label or selection bias to the data.

Figure 3 illustrates the results (with error bars) on the Adult dataset in which the privileged (also majority) group is *Male* and the favorable class is *'>50K'*. The positive base rate difference is around 23%. The baseline algorithms shown are LR, RF, and NN (from top to bottom: *balanced accuracy*, *average odds difference*, *disparate impact*, *statistical parity difference*, *equal opportunity difference*, *Theil index*). The SVM baseline has poor accuracy with very little bias on the Adult data, hence not shown. Our mitigator (**syn**, 2^{nd} col.) has the least bias overall while preserving the accuracy, followed by Reweighing (**rew**, 4^{th} col., pre-processing) and Reject Option (**ro**, last col., post-processing). Prejudice Remover (**pr**, 6^{th}*col.*, in-processing) is specific to LR, hence not shown in the RF and NN plots.

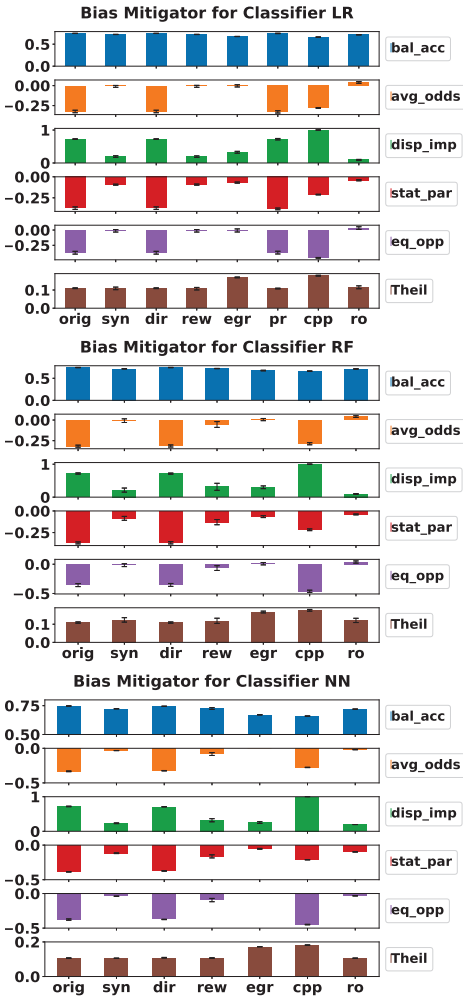
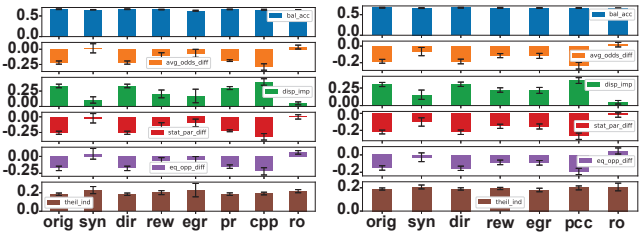


Figure 3: Fairness measures on Adult data.

In the *Compas* dataset, the minority is privileged. The number of examples in the favorable class (*no recidivism*) is approximately 10% higher than the unfavorable class, and the gap is approximately 22% between the privileged *Caucasian* group and the unprivileged *African-American* group. Since the minority is privileged, we invoke Method 2—generating synthetic privileged unfavorable samples in the *Caucasian* group. Figure 4 illustrates the fairness measures on the Compas dataset. Our mitigator (2^{nd} col.) is the overall winner under various combinations of baseline algorithms and fairness metrics. Reject Option (7^{th} col.) is competitive, however, its accuracy dropped significantly when the baseline was SVM (approximately 15% drop) or NN (approximately 5% drop).

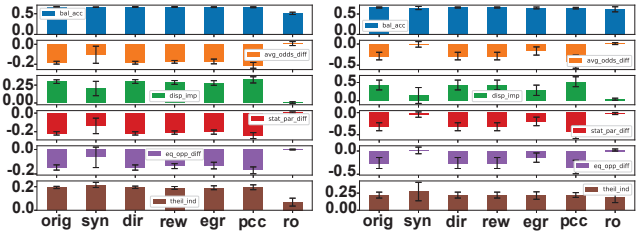
In the German Credit dataset, the protected attribute

[1] https://github.com/ut-dallas-dspl-lab/AI-Fairness



(a) Fairness with LR.

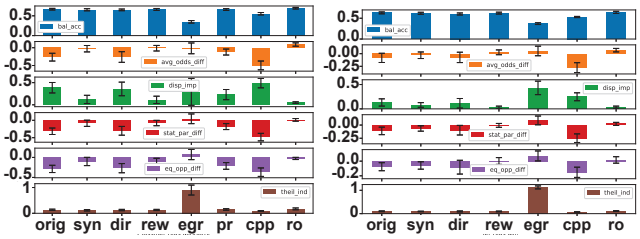
(b) Fairness with RF.



(c) Fairness with SVM.

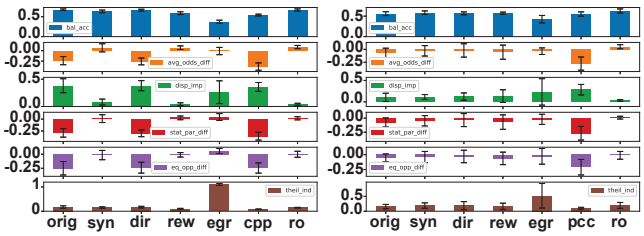
(d) Fairness with NN.

Figure 4: Fairness measures on Compas data. Rows: 1) bal_acc, 2) avg_odds, 3) disp_imp, 4) stat_par, 5) eq_opp, 6) Theil.



(a) Fairness with LR.

(b) Fairness with RF.



(c) Fairness with SVM.

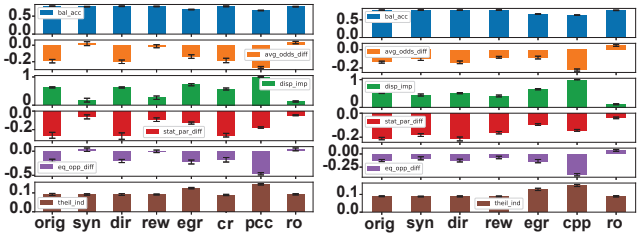
(d) Fairness with NN.

Figure 5: Fairness measures on German Credit data.

is "Age", with "old" and "young" being the privileged (also majority) and unprivileged groups, respectively. The privileged group has approximately 12% higher base rate than the unprivileged group. As can be observed in Figure 5, EGR (5^{th} col.) achieved significant bias reduction at the price of heavy accuracy loss. The best performers on German Credit include the synthetic-data mitigator (2^{nd} col.), Reweighing (4^{th} col.), and Reject Option mitigators (last col.). The pre-processing technique Disparate Impact Remover (**dir**, 3^{rd} col.) showed sensitivity to the type of AI algorithm, performing reasonably well when the baseline algorithm is RF or NN.

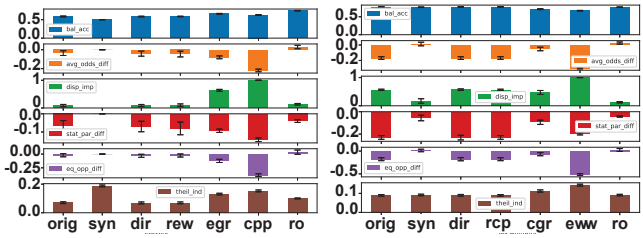
The Medical Expense Price dataset consists of approximately 97 million samples. A large amount of synthetic data would have to be generated, which appeared to be problematic for some baseline algorithms. The favorable class is *'≥ 10 Visits'* and the privileged (also majority) group is the *White* race. The base rate difference between the two groups is approximately 13%. When the baseline algorithm is LR or NN, our mitigator (2^{nd} col.) is comparable to the post-processing Reject Option technique (last col.). When the baseline is SVM, the generated synthetic data caused serious side effects that led to a significant drop in accuracy and serious individual bias (Theil index). The resulted classifier was nearly blindly assigning unfavorable labels to both the privileged group and the unprivileged group. The Reject Option was most effective, successfully improving both accuracy and fairness in nearly all cases. Detailed results are shown in Figure 6.

The privileged population in the *Bank* data accounts for 97.2% of the entire population, overwhelmingly dominant in this dataset. The favorable class is *subscribe deposit* and the privileged group is *'age ≥ 25'*. Favorable instances only



(a) Fairness with LR.

(b) Fairness with RF.

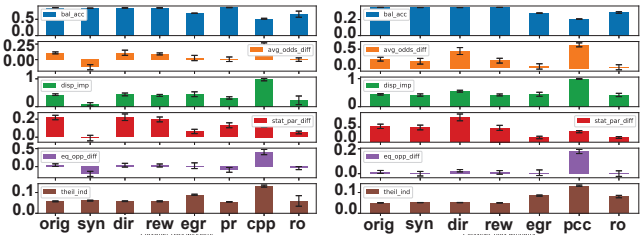


(c) Fairness with SVM.

(d) Fairness with NN.

Figure 6: Fairness measures on Medical Expense Price.

account for approximately 12% of the total. When the baseline is SVM or NN, all de-biasing techniques failed in the sense that they made little to no improvement without significantly losing accuracy, hence not shown in Figure 7. When the baseline is LR, our de-biasing technique (2^{nd} col.) was able to improve *disparate impact* and *statistical parity difference* while Reweighing (4^{th} col.) was able to improve *equal opportunity difference*, without significant accuracy loss. Little was achieved when RF was the baseline, unless accuracy was allowed to be traded for fairness. Rejection Option (last col.) suffered significant accuracy loss in both cases. Details are shown in Figure 7.



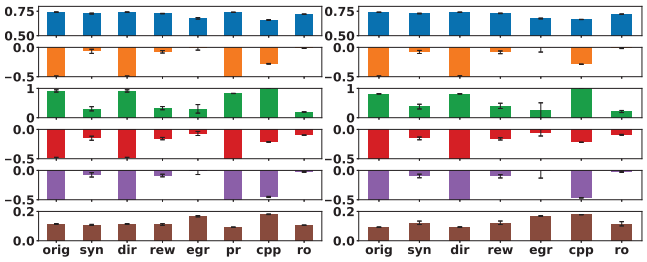
(a) LR mitigation results. (b) RF mitigation results.

Figure 7: Mitigation results on Bank data.

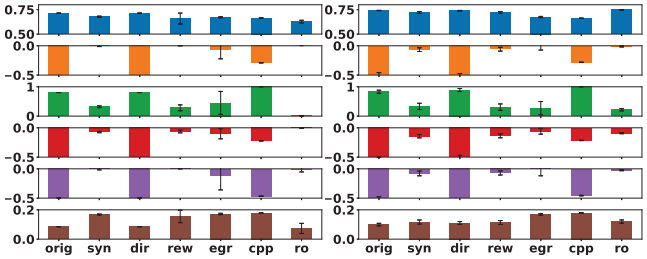
5.2 Influence of Label Bias and Selection Bias

In practice often we are not provided with the ground truth of fairness in the dataset. In this experiment, we assume the given data is unbiased, then we manually introduce a certain percentage of label/selection bias to the training data.

To introduce label bias, we randomly selected p% of unprivileged favorable instances and flipped their labels, with $p = 10, 30, 50$ respectively. In general, as the percentage of label bias increases, fairness on the test data deteriorates without the help of bias mitigators. Figure 8 shows the output of the baseline LR/RF/SVM/NN on the Adult data with our mitigator (syn) and six other mitigators, two from each of the Pre-/In-/Post-processing categories (dir,rew/egr,pr/cpp,ro).



(a) LR, 50% label bias. (b) RF, 50% label bias.



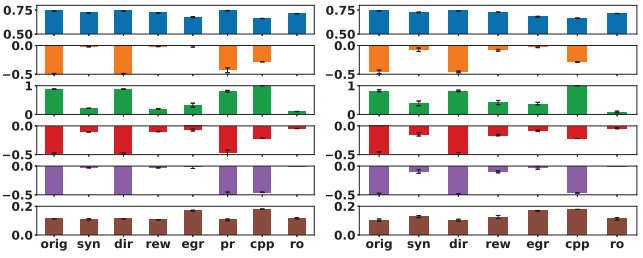
(c) SVM, 50% label bias. (d) NN, 50% label bias.

Figure 8: Fairness measures on Adult with 50% label bias. Rows: 1) bal_acc, 2) avg_odds, 3) disp_imp, 4) stat_par, 5) eq_opp, 6) Theil.

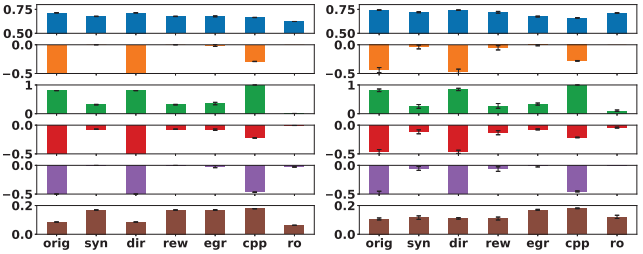
Our synthetic-data mitigator and Reject Option have the best overall performance compared to the rest of the mitigators. Reweighing works very well in general; however, when the baseline is SVM, it aggressively traded accuracy for fairness. Exponentiated Gradient Reduction (in-processing) tends to have a bad record of individual fairness (Theil index) except when the baseline is NN.

Our mitigator works best with the SVM baseline, while Reject Option works best when the baseline is NN, improving fairness significantly. However, Reject Option became very aggressive with SVM and 10–30% label bias, trading as much as 13% of accuracy for significant bias reduction. Our synthetic-data mitigator never experienced more than 2% accuracy loss in all cases. In addition, our mitigator works very well with all baselines even when there is a large percentage of label bias. Table 1 presents the fairness metrics of the original LR, Synthetic Oversampling, and Reject Option. Other combinations of different baseline algorithms and the amount of label bias draw similar conclusions.

Selection bias was introduced by deleting (10, 30, 50)% favorable unprivileged instances in the training data. We observed similar patterns on the Adult data (Figure 9) with 50% of unprivileged favorable training data. Experimental results for 30% and 10% selection bias are similar.



(a) LR, 50% select. bias. (b) RF, 50% select. bias.



(c) SVM, 50% select. bias. (d) NN, 50% select. bias.

Figure 9: Fairness of 50% selection bias in Adult data.

Oversampling-based de-biasing remains one of the best on other datasets too.

6 Conclusions

In this paper, we demonstrate, both theoretically and empirically, that SMOTE-based oversampling can deliver a promising outcome for mitigating bias in AI models. We focus on expanding the training data in the area where positive base rate difference is originated, without significantly violating the class priors. Compared to other mitigation techniques, synthetic oversampling produces better overall fairness without a significant loss of accuracy.

Table 1: Fairness metrics of the original LR, Synthetic Oversampling, and Reject Option with p label bias.

	Bal Acc	Avg Odds Diff	Disp Imp	Stat Par Diff	Eq_opp Diff	Theil Ind
	p = 50%					
LR_orig	0.737 ± 0.005	-0.535 ± 0.040	0.919 ± 0.032	-0.526 ± 0.042	-0.672 ± 0.057	0.115 ± 0.008
LR_syn	0.724 ± 0.003	-0.026 ± 0.019	0.240 ± 0.047	-0.108 ± 0.019	-0.029 ± 0.022	0.108 ± 0.009
LR_ro	0.714 ± 0.004	0.039 ± 0.011	0.094 ± 0.017	-0.041 ± 0.010	0.032 ± 0.019	0.118 ± 0.010
	p = 30%					
LR_orig	0.739 ± 0.004	-0.470 ± 0.009	0.830 ± 0.036	-0.485 ± 0.041	-0.533 ± 0.058	0.105 ± 0.011
LR_syn	0.720 ± 0.005	-0.016 ± 0.030	0.194 ± 0.060	-0.093 ± 0.029	-0.022 ± 0.035	0.109 ± 0.001
LR_ro	0.712 ± 0.005	0.035 ± 0.013	0.090 ± 0.010	-0.039 ± 0.007	0.023 ± 0.022	0.118 ± 0.010
	p = 10%					
LR_orig	0.747 ± 0.003	-0.337 ± 0.014	0.735 ± 0.011	-0.387 ± 0.010	-0.392 ± 0.023	0.107 ± 0.002
LR_syn	0.729 ± 0.002	-0.009 ± 0.009	0.200 ± 0.019	-0.091 ± 0.006	-0.012 ± 0.020	0.107 ± 0.008
LR_ro	0.715 ± 0.002	0.038 ± 0.022	0.088 ± 0.028	-0.040 ± 0.018	0.031 ± 0.030	0.115 ± 0.008

Acknowledgement

This material is based upon work supported by the NSF Program on Fairness in AI in Collaboration with Amazon under under Award FAI: Identifying, measuring, and mitigating fairness issues in AI, No. 1939728.. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69, 2018.

[2] Ananth Balashankar, Alyssa Lees, Chris Welty, and Lakshminarayanan Subramanian. What is fair? exploring pareto-efficiency for fairness constrained classifiers. *CoRR*, abs/1910.14120, 2019.

[3] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, volume 30, pages 3992–4001, 2017.

[7] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

[8] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, page 429–440, 2021.

[9] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[11] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

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

[13] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference in Learning Representations*, pages 1–14, 2016.

[14] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, 2015.

[15] Benjamin Fish, Jeremy Kun, and Adam Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152, 06

2016.

[16] Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[17] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2020.

[18] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

[19] Carlos Vladimiro Gonzlez Zelaya, Paolo Missier, and Dennis Prangle. Parametrised Data Sampling for Fairness Optimisation. In *Proceedings of Explainable AI for Fairness, Accountability & Transparency Workshop (KDD XAI)*. ACM, 2019.

[20] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323, 2016.

[21] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks*, pages 1322–1328, 2008.

[22] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, October 2012.

[23] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, pages 924–929, 2012.

[24] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the 2012th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ECMLPKDD'12, pages 35–50, 2012.

[25] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2564–2572, 10–15 Jul 2018.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT '19, pages 100–109, 2019.

[27] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, pages 180–191, 2004.

[28] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254. ACM, 2019.

[29] Ana Lavalle, Alejandro Maté, Juan Trujillo, Jorge García Carrasco, et al. A methodology based on rebalancing techniques to measure and improve fairness in artificial intelligence algorithms. 2022.

[30] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3384–3393, 10–15 Jul 2018.

[31] Natalia Martínez, Martín Bertrán, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning, 13-18, July 2020*.

[32] National Institute of Standards and Technology. Hlgmos synthetic data challenge submissions.

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. NIPS'17, pages 5684–5693, 2017.

[34] ProPublica. Machine bias, 2016. Available at `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[35] Mark F. Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986.

[36] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. DECAF: Generating fair synthetic data using causally-aware generative networks. In *Advances in Neural Information Processing Systems*, 2021.

[37] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1452–1458, 7 2019.

[38] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, 2018.

[39] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 1715–1724, 2020.

[40] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, number 3, pages 325–333, 2013.

[41] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, pages 325–333, 2013.

[42] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 335–340, 2018.