Unfair AI: It Isn't Just Biased Data

Chowdhury Mohammad Rakin Haider
Purdue University
West Lafayette, IN, USA
chaider@purdue.edu

Chris Clifton
Purdue University
West Lafayette, IN, USA
clifton@cs.purdue.edu

Yan Zhou
University of Texas
Dallas, TX, USA
yan.zhou2@utdallas.edu

2022 IEEE International Conference on Data Mining (ICDM) | 978-1-6654-5099-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDM54844.2022.00114

***Abstract*—Conventional wisdom holds that discrimination in machine learning is a result of historical discrimination: biased training data leads to biased models. We show that the reality is more nuanced; machine learning can be expected to induce types of bias not found in the training data. In particular, if different groups have different optimal models, and the optimal model for one group has higher accuracy, the optimal accuracy joint model will induce disparate impact even when the training data does not display disparate impact. We argue that due to systemic bias, this is a likely situation, and simply ensuring training data appears unbiased is insufficient to ensure fair machine learning.**

***Index Terms*—Machine Learning, Fairness, Systemic Bias**

I. Introduction

We are witnessing an unprecedented rise in automatic decision making models with long-lasting impact on human lives. Examples include hiring [1], bail/sentencing [2], lending [3], and medical treatment [4]. Machine learning (ML) models typically target high predictive accuracy. However, debates on ML model fairness introduce a new element.

Fairness is defined as equal treatment towards individuals irrespective of their sensitive attributes such as gender, race, age, etc. Much of the fairness-aware ML research (discussed more in Section II) assumes that unfair models result from historical unfairness in the input data. FairML literature commonly assumes that the training data contains a disproportionate number of individuals from the unprivileged group who received an incorrect/unfavorable outcome, which leads to group-wise disproportionate predictions.

The key contribution of this paper is to show that ML models can be expected to produce group-wise inequality *even when the training data is fair*. We start with data having no base rate disparity (equal favorable/unfavorable outcomes between groups), no majority or minority group (balanced), and all labels are presumed correct. We show that even with such fair training data, in some circumstances (e.g. disparately accurate features or *feature disparity*) an optimal accuracy ML model is *expected* to introduce disparate impact (different favorable outcome rates between groups) in predictions. In particular, when the optimal group-wise models are different we show theoretically that Bayes-optimal ML models can be expected to introduce such bias; we back this up with empirical results using multiple machine learning algorithms.

Furthermore, the circumstances where this occurs are a likely situation due to systemic bias: the features used in building models tend to be those that work well for the privileged group. For example, in college admission prediction, Test Score may be highly predictive for the privileged but perform poorly on the unprivileged. On the flip side, GPA may be more effective for the unprivileged group, but not as predictive as Test Score is for the privileged. In such a scenario, we show that even when the training data labels are correct and balanced between privileged and unprivileged groups, a Bayes-optimal classifier will be biased towards one group.

Specifically, when resources are scarce (as with selective admission colleges), ML models are expected to disproportionately favor the privileged group. In contrast, with an overabundance of resources (as with admissions to some for-profit schools), ML models disadvantage the unprivileged group by disproportionately offering resources to the wrong individuals in the unprivileged group (i.e., saddling individuals who do not succeed with student loan debt.) This holds even with equal group-wise base rates in the training data.

We discuss fairness metrics and a summary of literature in Section II. Section III theoretically establishes model-induced unfairness. We validate the theory empirically using synthetic fair datasets in Section IV. Then, we discuss implications of our findings, their relation to systemic bias, and feature disparity in real-world datasets in Section V.

II. Background

A. Fairness Metrics

Fairness notions are typically defined as group-wise equality of prediction statistics. Let dataset $\mathcal{D} = \{\mathbf{x}^{(k)}, s^{(k)}, y^{(k)}\}_{k=1}^{N}$, where $\mathbf{x}, s \in \{p, u\}$ and $y \in \{+, -\}$ are the set of n features, the sensitive attribute, and the target variable, respectively. Let, $s = p(= u)$ or in short $p(u)$ indicates privileged (unprivileged) group membership. Let the predicted outcome be $\hat{y}$. The most popular fairness metrics, *Disparate Impact*, concentrates on base rate equality. It requires equality of positive prediction probability among the groups, i.e., $P(\hat{y} = +|u) = P(\hat{y} = +|p)$, often presented as a ratio.

$$DI = \frac{P(\hat{y} = +|u)}{P(\hat{y} = +|p)}$$

Disparate impact can also be defined as a characteristic of the dataset rather than the model, $DI = \dfrac{P(y = +|u)}{P(y = +|p)}$

Other definitions include group-conditioned accuracy (e.g. equalized odds, equal opportunity), and group-conditioned calibration. For a comprehensive discussion of fairness metrics, see [5] and [6].

978-1-6654-5099-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDM54844.2022.00114

B. Related Work

The goal of fairness-aware machine learning is to develop non-discriminatory models with respect to sensitive attributes such as race, sex, etc. Fairness-interventions are broadly categorized as pre-processing, in-processing, and post-processing techniques. Typically, training data bias were re-captured in traditional models. Therefore, early work emphasized on pre-processing techniques such as class-label modification [7], sampling [8], [9], altering distributions to hide correlation to the sensitive attributes [10], etc. In-processing techniques involved modification of existing algorithms [11], [12], fair regularization terms [13], and optimization with fairness constraints [14], [15] as fairness-interventions. Agarwal et al. [16] modeled the fairness-intervention as a turn-taking game between fairness and accuracy optimizer. Post-processing techniques [12], [17] modify the predicted outcomes or model parameters to obtain non-discriminatory predictions.

Introduced bias is recently studied in [18] which defined *requisite features* as features d-connected to both the utility function and the sensitive attributes. Requisite features lead to introduced bias when p-admissible loss functions are optimized. Instead, we show requisite features lead to disparate group-wise classifiers. Consequently, the joint optimized classifier skews towards the more accurate one. We further report the impacts of resource constraints on induced bias.

Besides group-fairness, individual fairness (treat similar individual similarly) [14] and sub-group (middle ground between individual and group-fairness) [19] fairness were proposed. Specific fair learning applications, such as natural language processing, graph embedding, computer vision, and causal inference are also investigated. A detailed literature survey can be found in [6].

III. Formal Analysis

We show that an optimal classifier built on a fair balanced dataset (FBD) can still produce unfair outcomes. Let $P(p) = \beta P(u)$. When $\beta \neq 1$, the predictions are dominated by the majority which was addressed in [9] through oversampling minority group. In this work, we assume this issue has been rectified, and show model-induced bias with balanced datasets.

A. Fair Balanced Dataset (FBD)

We assume that our fair, balanced training dataset has identical group *outcomes* (i.e., equal base rates and sample count). While all labels are correct, the features are insufficient for a "perfect" (100% accuracy) classifier. The only distinction between the privileged and unprivileged group is that the Bayes-optimal model on the privileged group is different and more accurate than the one on unprivileged. Keeping with legal requirements in many countries, we discard the explicit use of sensitive attributes to separate the privileged and unprivileged groups when making predictions. We refer to the disparity in feature predictivity as *feature disparity*. We show that feature disparity leads to outcome bias in the joint optimal model.

Formally, an FBD $\mathcal{D}$ has $DI(\mathcal{D}) = 1$ and $y \not\perp s$. $y^+(y^-)$ indicates positive(negative) samples. Then $P(y^+) = \alpha = 1 - P(y^-)$. Assuming normally distributed features, a feature with higher distinction between samples from different classes is more predictive. In contrast, a non-predictive feature's distribution is independent of its class labels. A proxy measure of feature disparity is the separation of positive and negative sample distributions. Here, separation sep between two normal distributions $\mathcal{N}(\mu_1, \sigma)$, and $\mathcal{N}(\mu_2, \sigma)$ is,

$$sep = \frac{|\mu_1 - \mu_2|}{\sigma} \tag{1}$$

Let the most predictive feature set for the privileged and unprivileged group be $\{x_1, x_2, \ldots, x_r\}$ and $\{x_{r+1}, x_{r+2}, \ldots, x_{2r}\}$ respectively. Let, x_i^{sy} indicate the random variable corresponding to the feature x_i of the members of group s in class y and $\mu_i^{sy} = E[x_i^{sy}]$. If $i \leq r$,

$$x_i^{p+} \sim \mathcal{N}(\mu_i^{p+}, \sigma_i^2) \qquad x_i^{p-} \sim \mathcal{N}(\mu_i^{p-}, \sigma_i^2)$$

$$x_i^{u+}, x_i^{u-} \sim \mathcal{N}(\mu_i^{p_{avg}} + \delta, \sigma_i^2) \qquad \mu_i^{p_{avg}} = \frac{\mu_i^{p+} + \mu_i^{p-}}{2}$$

Here, $\delta(= 0)$ is a constant. Similarly, we define the distributions of $x_{r+1 \leq i \leq 2r}$. Finally, $x_{2r < i \leq n} \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Without loss of generality, we assume $\mu_i^{s-} < \mu_i^{s+}$.

We consider a Bayesian classifier θ, the theoretical optimal classifier for normally distributed features. Estimating prediction probabilities with Bayesian boundary and multivariate normal distribution requires generalized chi-squared distribution estimation [20]. Except for limited cases such as lower-dimensional linear boundary, the lack of known closed form probability estimates has led to several computational methods [20]. As such, we limit analytical results to $\sigma_i \not\perp y$ (linear boundary) with $r = 1$, $n = 2$ in Section III-B. Consequently, the unprivileged group is less separable than the privileged if $\sigma_1 < \sigma_2$; the groups are equally separable when $\sigma_1 = \sigma_2$. We show that even in such a simplified case, the model is expected to produce outcome bias not present in the training data; computational estimates (omitted due to space constraints) show this also holds for $n > 2$.[1]

B. Less Separable Unprivileged Group

A Bayesian classifier θ computes μ_i^θ and σ_i^θ for each feature x_i. For $\mathbf{x} \in \mathcal{D}$, the decision boundary of θ is defined as,

$$P(\hat{y}^+|\mathbf{x}) > P(\hat{y}^-|\mathbf{x}) \tag{2}$$

Here, $P(\hat{y}^+|y^+, s)$ and $P(\hat{y}^+|y^-, s)$ are group-wise true positive rate (TPR_s) and false positive rate (FPR_s). Therefore, the group-wise selection rate (SR_s) is,

$$\mathbb{P}(\hat{y}^+|s) = \alpha \mathbb{P}(\hat{y}^+|y^+, s) + (1 - \alpha)\mathbb{P}(\hat{y}^+|y^-, s) \tag{3}$$

$$\text{where } \mathbb{P}(\hat{y}^+|y, s) = \int_{\substack{\mathbf{x} \text{ s.t.} \\ P(\hat{y}^+|\mathbf{x}) > P(\hat{y}^-|\mathbf{x})}} P(\mathbf{x}|y, s) d\mathbf{x} \tag{4}$$

Applying the conditional independence assumption of the Naive Bayesian classifier (NBC) to (2), we get (5).

$$\frac{(x_1 - \mu_1^{avg})\Delta}{2(\sigma_1^\theta)^2} + \frac{(x_2 - \mu_2^{avg})\Delta}{2(\sigma_2^\theta)^2} + c_\alpha \geq 0 \tag{5}$$

[1] https://github.com/rakinhaider/Inherent-AI-Bias

Here, $c_\alpha = \log\frac{\alpha}{1-\alpha}$. The range of feature x_i that secures positive prediction is,

$$b_i(x_j) \leq x_i < \infty \qquad i,j \in \{1,2\} \text{ s.t. } i \neq j \tag{6}$$

$$\text{Here, } b_i(x_j) = \mu_i^{avg} - \frac{2(\sigma_i^\theta)^2}{\Delta}\left(c_\alpha + \frac{(x_j - \mu_j^{avg})\Delta}{2(\sigma_j^\theta)^2}\right)$$

Equation (4) is expanded as follows,

$$\mathbb{P}(\hat{y}^+|y,s) = \frac{1}{2} - \frac{1}{2}E_{x_1\sim P(x_1^{sy})}\left[erf\left(\frac{b_2(x_1) - \mu_2^{sy}}{\sqrt{2}\sigma_2}\right)\right] \tag{7}$$

Here, erf is the error function [21]. According to [21],

$$E_{x\sim\mathcal{N}(\mu,\sigma^2)}\left[erf\left(mx+n\right)\right] = erf\left(\frac{m\mu+n}{\sqrt{1+2m^2\sigma^2}}\right) \tag{8}$$

Theoretically, NBC is expected to yield

$$\mu_i^{\theta+} = \frac{1}{2}(\mu_i^{p+} + \mu_i^{u+}); \quad \mu_i^{\theta-} = \frac{1}{2}(\mu_i^{p-} + \mu_i^{u-}); \quad \sigma_i^\theta \geq \sigma_i$$

According to our assumptions,

$$\mu_2^{p_{avg}} = \mu_2^{p+} \quad \mu_1^{p_{avg}} - \mu_1^{p+} = -\frac{\Delta}{2}$$
$$\mu_1^{u_{avg}} = \mu_1^{p+} \quad \mu_2^{u_{avg}} - \mu_2^{p+} = -\frac{\Delta}{2}$$

To simplify the expressions, we define the following notations,

$$c_{denom} = 2\sqrt{2}\Delta\sqrt{(\sigma_1^\theta)^4\sigma_2^2 + (\sigma_2^\theta)^4\sigma_1^2} \quad c_2 = \frac{4(\sigma_1^\theta\sigma_2^\theta)^2}{c_{denom}}$$
$$c_1 = \frac{\Delta^2(\sigma_2^\theta)^2}{c_{denom}} \qquad c_3 = \frac{\Delta^2(\sigma_1^\theta)^2}{c_{denom}} \qquad c_{avg} = \frac{c_1+c_3}{2}$$

Let, $l = \mathbb{1}_{y=y^-}$. Using (7) and (8), we get,

$$\mathbb{P}(\hat{y}^+|y,s) = \frac{1}{2} + \frac{(-1)^l}{2}erf\left[c_1\mathbb{1}_{s=p} + c_3\mathbb{1}_{s=u} + (-1)^l c_2\right] \tag{9}$$

Having $\sigma_1 < \sigma_2$ and $\mu_i^{s+} - \mu_i^{s-} = \Delta; \forall i$, according to the definition of pooled variance, $\sigma_1^\theta < \sigma_2^\theta$. Since the error function is a strictly increasing function, from (9), we conclude

$$P(\hat{y}^+|y^+,p) > P(\hat{y}^+|y^+,u) \tag{10}$$
$$\text{and, } P(\hat{y}^+|y^-,p) < P(\hat{y}^+|y^-,u) \tag{11}$$

Equations (10) and (11) concludes that feature disparity results in imbalanced TPR and FPR. We now show the introduced disparity in selection rates. More precisely, $SR_p > SR_u$ when $\alpha < 0.5$. Extending (3) with (9), it is equivalent to show that:

$$\alpha\left[erf(c_1+c_2c_\alpha) - erf(c_3+c_2c_\alpha)\right] > (1-\alpha)\left[erf(c_1-c_2c_\alpha) - erf(c_3-c_2c_\alpha)\right] \tag{12}$$

Since $\sigma_1 \neq \sigma_2$, it implies that $c_1 \neq c_3$. Therefore, from Fig. 1, we conclude that:

$$erf(c_1 \pm c_2c_\alpha) - erf(c_3 \pm c_2c_\alpha) \propto \frac{d}{dx}erf(x)\Big|_{c_{avg}\pm c_2c_\alpha} \tag{13}$$

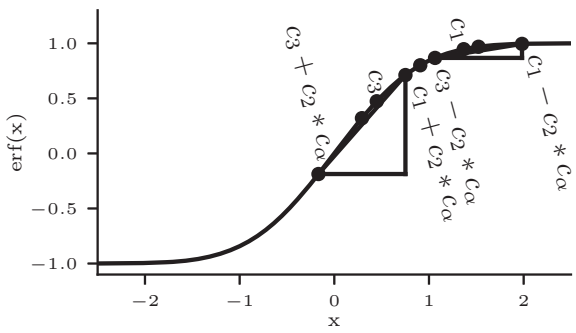

Fig. 1. Erf Function with $\Delta = 10, \sigma_1 = 2$ and $\sigma_2 = 5$

From (12),(13) and the definition of the *erf* function,

$$\frac{\exp\left(-(c_{avg}-c_2c_\alpha)^2\right)}{\exp\left(-(c_{avg}+c_2c_\alpha)^2\right)} - \frac{\alpha}{1-\alpha} < 0$$
$$\therefore \mathrm{e}^{4c_{avg}c_2c_\alpha} - \mathrm{e}^{c_\alpha} < 0$$
$$\therefore (4c_{avg}c_2 - 1)c_\alpha < 0 \tag{14}$$

With $c_\alpha < 0$, (14) holds only when $4c_{avg}c_2 > 1$. Moreover, $\sigma_i < \sigma_i^\theta$, implies that $4c_{avg}c_2 > 1$. Therefore, (12) holds and we conclude that $SR_p > SR_u$ when $\alpha < 0.5$ Similarly, we can show that $SR_p < SR_u$ when $\alpha > 0.5$. (When $\alpha = 0.5$, the selection rates are expected to be equal.) To summarize, an NBC trained on FBD $\mathcal{D}$ with disparate group-wise optimal model accuracy (due to feature disparity) can result in disparate selection rates among the groups.

C. Less Separable Unprivileged Group with Resource Constraints

Resource availability are often not aligned with the demand. Scarcity of resources compels selection of a sub-sample of deserving candidates. In contrast, for higher utilization, surplus resources can distributed with relaxed qualification considerations. Typical approaches to enforce resource constraints may involve decision boundary modification or probabilistic ranking with pre-established cut-off for selection. While the former method approximates the constraint, the latter guarantees exact resource allocation. Let, the decision boundary is constrained with a constant c_{res} that controls the rate of positive predictions. Using (15), we can obtain (16).

$$P(\hat{y}^+|\mathbf{x}) > c_{res}P(\hat{y}^-|\mathbf{x}) \tag{15}$$
$$\mathbb{P}(\hat{y}^+|y,s) = \frac{1}{2} + (-1)^l\frac{1}{2}erf(z) \tag{16}$$

Here, $z = c_1\mathbb{1}_{s=p} + c_3\mathbb{1}_{s=u} + (-1)^l c_2 - (-1)^l\frac{\log c_{res}}{c_{denom}}$

Since $c_{res} \not\perp\!\!\!\perp s$, equation (16) is similar to (9) and the relationship in (10) and (11) still hold. Later, in Section IV-C2, we show that probabilistic ranking with pre-established cut-off further exacerbates the outcome fairness.

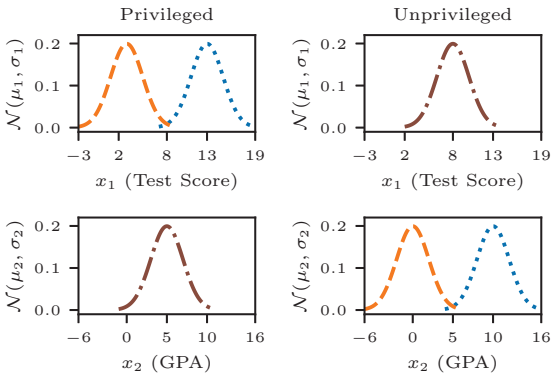

Fig. 2. Example of feature distributions when the unprivileged group is equally separable to the privileged group, but optimal models are different. The blue (dotted) and orange (dashed) curves represent positive and negative sample distributions. The brown (dashdot) curve indicates two distributions (positive and negative) are overlapping.

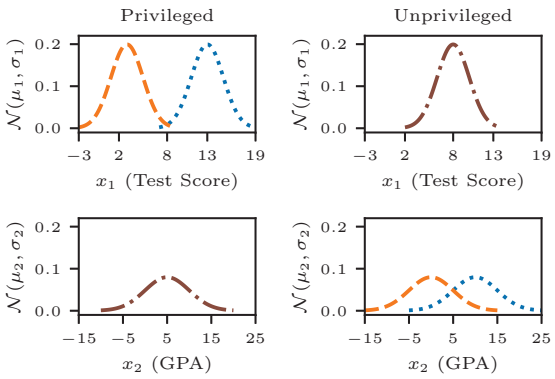

Fig. 3. Example of feature distributions when the unprivileged group is less separable than the privileged group. The plot colors and patterns convey similar meaning as described in Fig. 2.

TABLE I
MODEL PERFORMANCES ON EQUALLY SEPARABLE UNPRIVILEGED GROUP

α	AC_p	AC_u	SR_p	SR_u	FPR_p	FPR_u
0.25	99.6	99.3	20.7	21.0	00.4	00.5
0.50	99.4	99.5	50.1	49.8	04.2	03.7
0.75	99.5	99.3	78.6	78.8	15.8	16.7

IV. EXPERIMENTAL RESULTS

A. Fair Balanced Datasets

Real world data typically contain historical unfairness. Instead, we conduct experiments on synthetic fair balanced datasets (SFBD). Following Section III, the r group-specific predictive attributes are sampled conditioned on class labels whereas the $n - 2r$ non-predictive attributes are unconditionally sampled from a random normal distribution. For example, in college admission prediction, privileged samples are generated by sampling x_{TS}^{p+} and x_{TS}^{p-} from two distinct distributions. In contrast, x_{GPA}^{p+} and x_{GPA}^{p+} are sampled from the same distribution. In this work, each *SFBD* contains 10000 samples. We limit the discussion to SFBD with 2 attributes ($r = 1, n = 2$) due to space limitations. Experiments with more features produced a similar outcome. We generate one SFBD for each $\alpha \in \{0.25, 0.5, 0.75\}$ with $\mu_1^{p+} = 13, \mu_2^{u+} = 10$ and $\Delta = 10$.

B. Equally Separable Unprivileged Group

We first consider equally separable privileged and unprivileged group, i.e., $\sigma_1 = \sigma_2 = 2$. In the college admission scenario, this corresponds to $\sigma_{TS} = \sigma_{GPA}$. Fig. 2 shows the distribution of the attributes where the groups are equally separable. Clearly x_1 and x_2 are equally predictive for the privileged and the unprivileged group respectively. Table I denotes the group-wise optimal model accuracy as AC_s. We observe similar AC_s, SR_s and FPR_s for each group indicating unbiased joint-optimal model.

TABLE II
MODEL PERFORMANCES ON LESS SEPARABLE UNPRIVILEGED GROUP

Method	α	AC_p	AC_u	SR_p	SR_u	FPR_p	FPR_u
NBC	0.25	99.6	87.4	21.1	15.4	00.3	07.2
	0.50	99.4	84.4	49.9	49.4	03.0	25.9
	0.75	99.5	87.1	78.3	85.6	13.6	61.3
SVM	0.25	99.6	87.4	21.7	19.4	01.5	07.2
	0.50	99.4	84.2	50.0	49.9	05.9	20.4
	0.75	99.5	87.1	77.3	81.3	14.2	45.6
DT_5	0.25	99.5	87.4	21.7	21.7	02.0	09.0
	0.50	99.4	83.8	52.5	57.9	08.6	30.5
	0.75	99.4	86.5	77.6	82.7	16.4	49.0
PR	0.25	99.5	87.5	25.0	21.4	00.3	06.1
	0.50	99.4	84.4	50.1	49.1	00.6	14.7
	0.75	99.5	87.1	75.6	75.8	02.8	30.2
RBC	0.25	99.5	87.5	22.1	21.8	01.1	10.1
	0.50	99.4	84.4	50.1	49.0	04.7	22.1
	0.75	99.5	87.1	76.8	78.4	11.9	43.2

C. Less Separable Unprivileged Group

1) Without Resource Constraints: The less separable unprivileged group has higher standard deviation in predictive attributes. The less separable unprivileged group has $\sigma_{TS} = \sigma_1 = 2$ and $\sigma_{GPA} = \sigma_2 = 5$. Fig. 3 is the distribution of a less separable unprivileged group, which is similar to Fig. 2 except that $\sigma_1 < \sigma_2$. Clearly, x_1 is a better predictor for the privileged group than x_2 is for the unprivileged. NBC achieves lower unprivileged model accuracy on the less separable unprivilegd group (Table II). We observe significant SR_s and FPR_s disparity for $\alpha \in \{0.25, 0.75\}$. SR_p is higher when $\alpha = 0.25$, but opposite for $\alpha = 0.75$. This confirms the theoretical result of Section III-B. The higher SR_u at $\alpha = 0.75$ is largely due to the high FPR_u of 61.28%. In the running example of college admission predictions, highly selective campuses offer 37.36% more admission to privileged students. With low selectivity, the unprivileged individuals are admitted at a 9.44% higher rate, but a significant portion of acceptances are received by under-

TABLE III
MODEL PERFORMANCES ON LESS SEPARABLE UNPRIVILEGED GROUP WITH SCARCE OR SURPLUS RESOURCES

Resources	α	AC_p	AC_u	SR_p	SR_u	FPR_p	FPR_u
Scarce	0.25	92.4	83.6	14.9	05.1	00.0	01.3
	0.50	69.6	64.1	17.3	02.7	00.0	00.3
	0.75	44.9	41.4	18.1	01.9	00.0	00.0
Surplus	0.25	45.0	40.5	81.3	98.7	75.0	98.3
	0.50	69.6	63.6	82.4	97.6	64.7	95.5
	0.75	92.7	83.2	84.5	95.5	37.8	85.6

TABLE IV
MODEL PERFORMANCES FOR COMPAS SFBD DE-BIASED BY INFLATING PRIVILEGED OR UNPRIVILEGED UNFAVORED CLASS

Inflated Class	α	AC_p	AC_u	SR_p	SR_u	FPR_p	FPR_u
Priv.	0.25	73.3	74.0	21.1	10.9	15.5	06.5
	0.50	61.8	61.6	62.4	41.6	52.1	30.3
	0.75	75.9	72.2	89.0	81.9	76.5	68.7
Unpriv.	0.25	74.6	72.5	16.2	05.4	12.0	06.0
	0.50	61.2	61.2	75.2	51.1	67.6	34.0
	0.75	74.7	73.6	91.7	78.3	83.3	68.7

qualified students, with the risk of higher drop-out rate and additional student loan burden among the unprivileged group.

Table II shows comparison of five classifiers including two state-of-the-art fairness-aware algorithms. Traditional models such as SVM and Decision Tree (maximum depth is fixed at five to reduce over-fitting) (DT_5) demonstrates similar behavior to Naive Bayes. Notably, DT_5 has the lowest false positive differences among SVM and NBC. Prejudice Remover (PR) [13] and Reduction based classifier (RBC) [16] appear inherently less biased than traditional models. Although specifically designed to reduce disparate impact, they still suffer with respect to false positive rates. A more realistic experiment, with $r > 1$ and $n > 2$ features dampened group-wise accuracy disparity and selection rate disparity (not shown due to space constraints). However, false positive rate disparity still existed and increased with α. The biased and unbiased outcomes shown in Table II and Table I respectively, demonstrate that poor feature selection is sufficient to cause outcome bias in an otherwise fair and balanced dataset.

2) With Resource Constraints: We adapt the probabilistic predictions of Naive Bayes classifier as rankings and set cut-off to limit positive (or negative) outcomes. Let, p_{res} is the resource to candidate ratio. We experiment with scarce ($p_{res} = 10\%$) and surplus ($p_{res} = 90\%$) resources. We vary the true need for the resource between α =25%, 50%, and 75%.

Table III shows that with insufficient resources, privileged individuals are selected 3 times more than unprivileged ones. This corresponds to selecting more privileged group members at highly selective prestigious colleges where the unprivileged intake rate degrades as the competition increases. In contrast, surplus resources lead to high FPR_u which yields higher acceptance rate in unprivileged group. We draw parallels between such model behavior and predatory colleges, that offers admission to unqualified minorities eventually creating higher drop out rate and additional student loan burden.

D. Fairly Sampled Balanced COMPAS Dataset

We experiment with the real-world COMPAS dataset [22]. The dataset consists of privileged (also the minority) *Caucasian*s and unprivileged *African American* group where the base positive (no-recidivism) rate difference between them is 14%. We use a de-biasing algorithm [9] that generates synthetic samples, using the SMOTE algorithm [23], to balance the group-wise positive rates.

We resampled the de-biased COMPAS dataset to obtain datasets with $\alpha \in \{0.25, 0.5, 0.75\}$. We perform a 70:30 train to test split. Since the preprocessed COMPAS dataset [24] contains only binary variables, instead of Gaussian we assume Bernoulli distributed attributes. Although the de-biased dataset has slight group-wise optimal model accuracy disparity, we observe significant disparity in group-wise selection rates. It can be ascribed to both inherent bias and data bias. Consistent with Section IV-C1, the selection rate increases for the unprivileged group as α increases.

V. SYSTEMIC BIAS: WHY WE EXPECT THESE OUTCOMES

It could be argued that in an otherwise hypothetical fair world we would not see accuracy differences between within-group optimal models. We suggest that this is instead a common and likely form of systemic bias. ML systems are generally designed by the privileged group, and the features considered are the ones that seem natural to that group. The (privileged) developers are unaware of or do not consider features that are effective for unprivileged groups, leading to inherently more accurate systems for the privileged group.

We cite two well-known examples of feature disparity from medical research. Chest pain or discomfort, most-taught symptoms of heart attack, turned out to be only dominant in men. Women are more likely to experience other symptoms, particularly shortness of breath, nausea, and back or jaw pain [25]. Similarly, cancer research long focused on lung cancer, which at the time disproportionately impacted young males (average age of diagnosis was 66 in 1975-1999 [26]). Public outcry over this gender disparity led to increased investment in breast cancer, which became overfunded relative to other cancers in terms of years of life lost [27]. While not directly a machine learning issue, we see that investments followed the stakes of the privileged group. These exemplify the situations were features used or studied are obvious to (and work well for) the privileged, eventually harming the unprivileged.

We analyzed the feature set of COMPAS dataset. The overall privileged to unprivileged group accuracy difference of NBC on COMPAS dataset is 0.5%. Since COMPAS contains historical bias, disparity in number of group samples and base-rates, a small accuracy difference could be a mixed outcome of feature disparity and other biases. Therefore, we analyze predictive power disparity of each feature f, PP_f, defined as the prediction accuracy of the models using a single feature f.

TABLE V
PREDICTIVE POWERS OF EACH FEATURE IN COMPAS DATASET.

Feature f	Avg. PP_f^p	Avg. PP_f^u	PPD_f
juv_fel_count	61.25	50.02	11.23
juv_misd_count	62.15	51.68	10.47
juv_other_count	61.82	53.26	8.56
age_cat	60.91	54.96	5.95
c_charge_degree	60.91	55.18	5.73
priors_count	64.91	60.41	4.50

We obtain predictive power difference of feature f, $PPD_f = PP_f^p - PP_f^u$ from group-wise single-feature predictive powers of f PP_f^s. We perform 10-fold cross-validation to obtain PP_f^s. Table V shows the maximum absolute PPD_f in COMPAS is 11.2%. It suggests that disproportionately predictive features are commonplace in real-world machine learning datasets. Two credit-scoring datasets showed similar trends, but are not shown due to space constraints.

VI. CONCLUSION

In this work, we demonstrate *model*-induced bias, as opposed to data-induced bias. We show that a Bayes-optimal classifier can be expected to induce biases in the outcome that are otherwise absent in the data. Experimental results validate that if group-wise optimal model accuracy for demographic groups are different, the joint optimal Bayesian model trained on a fair dataset demonstrates disparate impact. The disparity in group-wise accuracy can arise from disproportionately predictive features. We argue feature disparity is a form of systemic bias, and machine learning exacerbates this bias. It is tempting to address this by using separate models for different groups, but this may violate ethical and legal standards (e.g., U.S. civil rights laws, E.U. GDPR Article 9). A second approach is to optimize for fairness rather than accuracy [16], as in [28] and many more recent works. We suggest that a better approach is to eliminate the underlying disparity, using methods such as participatory design to produce better predictive features for all.

REFERENCES

[1] C. C. Miller, "Can an Algorithm Hire Better Than a Human?" 2015. [Online]. Available: https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] K. Petrasic, B. Saul, J. Greig, M. Bornfreund, and K. Lamberth, "Algorithms and Bias: What Lenders Need to Know," 2017. [Online]. Available: https://www.whitecase.com/publications/insight/algorithms-and-bias-what-lenders-need-know

[4] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[5] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, (FAT*)*. Atlanta, GA, USA: ACM, 2019, pp. 329–338.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Survey*, vol. 54, no. 6, pp. 115:1–115:35, 2021.

[7] F. Kamiran and C. Toon, "Classifying without discriminating," in *Proceedings of the 2nd International Conference on Computer, Control and Communication*. Karachi, Pakistan: IEEE, 2009, pp. 1–6.

[8] K. Faisal and C. Toon, "Classification with no discrimination by preferential sampling," in *Proceedings of the 19th Machine Learning Conf. Belgium and The Netherlands*, Leuven, Belgium, 2010, pp. 1–6.

[9] Y. Zhou, M. Kantarcioglu, and C. Clifton, "Improving fairness of AI systems with lossless de-biasing," *CoRR*, vol. abs/2105.04534, 2021.

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney, Australia: ACM, 2015, pp. 259–268.

[11] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *Proceeding of the 10th IEEE International Conference on Data Mining (ICDM)*. Sydney, Australia: IEEE Computer Society, 2010, pp. 869–874.

[12] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[13] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*. Bristol, UK: Springer, 2012, pp. 35–50.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," in *Proceedings of the Innovations in Theoretical Computer Science*. Cambridge, MA, USA: ACM, 2012, pp. 214–226.

[15] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web, WWW*. Perth, Australia: ACM, 2017, pp. 1171–1180.

[16] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018*, 2018.

[17] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. J. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *CoRR*, vol. abs/1706.02409, 2017.

[18] C. Ashurst, R. Carey, S. Chiappa, and T. Everitt, "Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness," in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual Event, 2022*. AAAI Press, 2022, pp. 9494–9503.

[19] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 2018*. PMLR, 2018.

[20] A. Das and W. S. Geisler, "A method to integrate and classify normal distributions," *Journal of Vision*, vol. 21, no. 10, pp. 1–1, 2021.

[21] S. I. Wang and C. D. Manning, "Fast dropout training," in *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, GA, USA: JMLR.org, 2013, pp. 118–126.

[22] D. Dua and C. Graff, "UCI machine learning repository," http://archive.ics.uci.edu/ml, 2017.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[24] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017, pp. 3992–4001.

[25] "What are the warning signs of heart attack?" 2015. [Online]. Available: https://www.heart.org/en/health-topics/heart-attack/warning-signs-of-a-heart-attack

[26] L. Eldgridge, "What is the average age for a lung cancer diagnosis?" https://www.verywellhealth.com/what-is-the-average-age-for-lung-cancer-2249260, Jul. 30 2020.

[27] A. J. Carter and C. N. Nguyen, "A comparison of cancer burden and research spending reveals discrepancies in the distribution of research funding," *BMC Public Health*, vol. 12, no. 526, Jul. 17 2012.

[28] K. Mancuhan and C. Clifton, "Combating discrimination using bayesian networks," *Artificial Intelligence & Law*, vol. 22, no. 2, pp. 211–238, 2014.