Ensuring Fairness with Transparent Auditing of Quantitative Bias in AI Systems

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—With the rapid advancement of AI, there is a growing trend to integrate AI into decision-making processes. However, AI systems may exhibit biases that lead decision-makers to draw unfair conclusions. Notably, the COMPAS system used in the American justice system to evaluate recidivism was found to favor racial majority groups; specifically, it violates a fairness standard called equalized odds. Various measures have been proposed to assess AI fairness. We present a framework for auditing AI fairness, involving third-party auditors and AI system providers, and we have created a tool to facilitate systematic examination of AI systems. Including independent and trusted auditors is pivotal for objectivity and accountability; third parties are often further necessary for their specialized expertise in relevant domains. Auditors, equipped with our tool, can thoroughly review AI systems for bias and fairness violations. The tool is open-sourced and publicly available. Unlike traditional AI systems, we advocate a transparent white-box and statistics-based approach. It can be utilized by third-party auditors, AI developers, or the general public for reference when judging the fairness criterion of AI systems.**

***Index Terms*—AI, fairness, auditing**

I. Introduction

The accelerating pace of artificial intelligence (AI) technologies has revolutionized numerous fields in recent years. In healthcare, AI-driven diagnostic systems streamline disease identification, while in finance, automated trading algorithms analyze market trends to execute optimal trades swiftly. This remarkable progress has reached diverse industries, sprouting out various applications for different purposes.

One of its most profound impacts is how AI has transformed decision-making processes across sectors. By harnessing vast amounts of data and employing advanced algorithms, AI empowers organizations to make more informed and strategic decisions. From assessing job applicants to determining school admissions, AI-driven insights offer efficient and analytical advantages.

AI system, however, may be biased. Factors such as inherent biases in original data sets or flaws in algorithm designs could contribute to bias within AI systems. When left unchecked, the ramifications of such biases extend far beyond mere inaccuracies; they can lead to devastating consequences, such as amplifying systemic injustices, perpetuating group discrimination, and exacerbating societal inequalities.

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a tool for predicting recidivism—the tendency of criminals to reoffend. It is used in the criminal justice system in multiple states in the United States. In 2016, it was discovered in an investigation by the journalists at ProPublica that the COMPAS system is, in fact, unfair towards minority and disadvantaged groups.

Cases such as COMPAS underscore the critical importance of rigorously examining the fairness of AI systems by third parties. Fairness is fundamentally a subjective social construct, heavily influenced by cultural context and deeply rooted in historical inequalities. However, there have been developments in research leveraging statistical metrics to quantify fairness that provide transparent and objective insights.

By employing such mathematically rigorous methodologies, our research can offer a deeper look into the fairness of AI systems, enabling third parties to make more informed and robust judgments. For example, it would reveal that the COMPAS system violates a fairness measure called equalized odds.

In addition to equalized odds, there are a number of fairness measures for AI systems. Our framework aims to help third parties apply these metrics and navigate complex datasets with ease, pinpointing instances of bias and discrimination.

With our tool, a third party such as ProPublica could clearly and easily demonstrate that COMPAS violates equalized odds according to their datasets. Our platform empowers organizations to conduct thorough assessments of AI fairness and uncover discrepancies in predictive models.

This functionality not only facilitates sound judgment but also fosters objectivity and transparency regarding AI systems, ultimately bolster the claim of fairness and justice in AI-enabled decision-making.

Our tool is written in Python and is offered as a Python package. It supports common dataset formats such as CSV. It is open-sourced and publicly available for downloads.

II. Fairness Measures

Fairness is about making sure the disadvantaged and unprivileged groups of individuals are treated equitably. However, there have been several interpretations of it proposed in the past. To have a constructive discussion on fairness, we

Identify applicable funding agency here. If none, delete this.

must first have precise definitions of it. Pessach and Shmueli [1] formulated a number of fairness measures in a unified mathematical notation. We will base our framework on their formulation.

A. Preliminaries

For the following definition, we will use Y to denote the ground truth of an outcome; $\hat{Y}$ to denote the predicated result of an outcome; $Y = 1$ and $\hat{Y} = 1$ to denote them being accepted or positive. For example, let Y be recidivism. Then $Y = 1$ means the case of an individual actually recidivating and $\hat{Y} = 0$ means the prediction of an individual recidivating is negative. In addition, we use V and $\hat{V}$ when the truth and the prediction aren't binary. For example, we denote the COMPAS score by $\hat{V}$, which ranges from 1 to 10.

Protected attributes are the characteristics of individuals that are, for example, legally or ethically, considered sensitive and warrant protection against discrimination and bias. We denote by S some protected attribute. We write $S = 1$ to represent the privileged group and $S \neq 1$ to represent the unprivileged group. For example, let S be Caucasian. Then $S = 1$ represents the case of an individual's race being Caucasian and $S \neq 1$ vice versa.

We denote by ϵ some threshold that is used to limit the fairness measures.

A positive outcome or prediction may be either beneficial or detrimental to the individual in different cases. Various measures were devised when considering different scenarios. For presentation, we consider $Y = 1$ to be the detrimental case, like in the COMPAS prediction scenarios.

B. Disparate Impact

In 1971 [2], the US Supreme Court ruled that it is illegal for hiring decisions to have "disparate impact" by race, thus coining the term. It is taken as unintentional discrimination, as opposed to intentional discrimination, which is called "disparate treatment".

Legal cases involving disparate impact often refer to the "80% Rule", advocated by the US Equal Employment Opportunity Commission [3], where it requires the selection rate of a minority group to be no less than 80% of that of a majority group. Formally [4], it is:

$$\frac{P[\hat{Y} = 1 | S \neq 1]}{P[\hat{Y} = 1 | S = 1]} \geq 1 - \epsilon$$

In the case of 80% rule, $\epsilon = 20\%$.

C. Demographic Parity

Demographic parity [5], also known as statistical parity, is similar to disparate impact but, instead of ratio, the difference is taken. It is named so to suggest that each demographic group(such as race, gender, or age) should have equal representation or opportunity. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1] - P[\hat{Y} = 1 | S \neq 1]| \leq \epsilon$$

D. Conditional Statistical Parity

Conditional statistical parity [6] is similar to demographic parity, but, in addition to protected attributes, it further takes into account some "legitimate" attributes that are *legitimately* related to the case. For example, a legitimate attribute when considering future recidivism could be the number of prior crimes committed. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, L = l] - P[\hat{Y} = 1 | S \neq 1, L = l]| \leq \epsilon$$

where L denotes the legitimate attributes.

E. Overall Accuracy Equality

Overall accuracy equality [7] is similar to demographic parity, but instead of the case of $\hat{Y} = 1$, it considers the case of $Y = \hat{Y}$; that is, the case where the prediction is accurate. Formally, it is:

$$|P[Y = \hat{Y} | S = 1] - P[Y = \hat{Y} | S \neq 1]| \leq \epsilon$$

F. Mean Difference

Mean difference [8] considers the expected value of the prediction. Formally, it is:

$$|E[\hat{Y} | S = 1] - E[\hat{Y} | S \neq 1]| \leq \epsilon$$

G. Equalized Odds

Equalized odds [9] is similar to demographic parity, but it further takes into account the ground truth. It considers the cases of true positive and false positive. It solves the downsides of a fully accurate classifier that might be deemed unfair by measures that do not consider ground truth. For example, consider a group A that is predicated to recidivate and they do in fact recidivate and a group B that is predicated to recidivate but end up never recidivating. A fully accurate classifier will always predict A to recidivate while B to never recidivate, and this will violate demographic parity. By taking ground truth into account, equalized odds avoids these pitfalls. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, Y = 0] - P[\hat{Y} = 1 | S \neq 1, Y = 0]| \leq \epsilon$$
$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$

H. Equal Opportunity

Equal opportunity [1] [9] is a relaxation of equalized odds by only considering the true positive case. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$

I. Predictive Equality

Predictive equality [6] is also a relaxation of equalized odds by only considering the false positive case. Formally, it is:

$$|P[\hat{Y} = 1 | S = 1, Y = 0] - P[\hat{Y} = 1 | S \neq 1, Y = 0]| \leq \epsilon$$

[1] It was named so by considering the cases where positive outcomes benefit individuals, such as school admissions, hence the word opportunity.

J. Conditional Use Accuracy Equality

Conditional use accuracy equality [7] is similar to equalized odds, but instead of conditioning on the ground truth, it conditions on the prediction and calculates the probability of the ground truth. It can be seen as checking the prediction accuracy across groups, thus the name. It further requires the measure in the case of positive predictive values to be less than that of the negative predictive values. Formally, it is:

$$\begin{array}{cr} |P[Y=1|S=1,\hat{Y}=1]-P[Y=1|S\neq 1,\hat{Y}=1]| & \leq \epsilon \\ \wedge & \\ |P[Y=0|S=1,\hat{Y}=0]-P[Y=0|S\neq 1,\hat{Y}=0]| & \leq \epsilon \end{array}$$

K. Predictive Parity

Predictive parity [10] is a relaxation of conditional use accuracy equality by only considering the positive predictive value case. Formally, it is:

$$|P[Y=1|S=1,\hat{Y}=1]-P[Y=1|S\neq 1,\hat{Y}=1]| \leq \epsilon$$

L. Equal Calibration

Equal calibration [10] is similar to equal opportunity, but instead of having a binary $\hat{Y}$, it is conditioned on the range of the predicted value $\hat{V}$. For example, this could be conditioned on the highest COMPAS score $\hat{V} = 10$. Calibration is a concept of having a fair score function [11]. Formally, it is:

$$|P[Y=1|S=1,\hat{V}=v]-P[Y=1|S\neq 1,\hat{V}=v]| \leq \epsilon$$

M. Positive Balance

Positive balance [13] is similar to equal opportunity, but instead of taking the difference of probability of binary prediction $\hat{Y}$, it takes the difference of the expected value of the score $\hat{V}$, which may be non-binary, such as the score of COMPAS. Formally, it is:

$$|E[\hat{V}|Y=1,S=1]-E[\hat{V}|Y=1,S\neq 1]| \leq \epsilon$$

N. Negative Balance

Negative balance [13] is like positive balance except it conditions on the case of $Y=0$. Formally, it is:

$$|E[\hat{V}|Y=0,S=1]-E[\hat{V}|Y=0,S\neq 1]| \leq \epsilon$$

These fairness measures are compiled in Table I.

III. Auditing Framework

Per the review by Pessach and Shmueli [1], we designed an auditing framework for calculating the various fairness measures. We offer two versions of fairness checkers: one for when the prediction results are readily available in CSV input and one for when a model is provided. In most auditing cases the model version is preferred because CSV results can be easily fabricated.

A. Notations

Types α, β are stand-ins for any type. Let $\alpha \to \beta$ denote a function from type α to type β. α_i denotes some particular type; $\prod_i \alpha_i$ denotes the Cartesian product of multiple α of possibly different types. We write $x : \alpha$ to mean x is of the type α. We write `str` for the string type, `bool` for the boolean type, and `int` for the integer type.

A database $\mathcal{D} = \{r_1, r_2, ...\}$ is a collection of rows. A row r_i : `key` $\to$ `value` is a lookup table or dictionary, where the concrete type of `key` and `value` is `str`. For example, r_n("sex") = "Female" means r_n's sex is female. Henceforth, we will use `row` as a type synonym of `key` $\to$ `value`.

A model $\mathcal{M}$: `row` $\to \alpha$ is a black-box predictor that takes a row and returns its prediction result of some type α; for example, if the model returns a string then $\alpha =$ `str`.

A privileged predicate R : `row` $\to$ `bool` takes a row and determines if it belongs to the privileged group. For example, $R(r_i) := r_i$("race") $==$ "Caucasian" means the privileged group is those with race being Caucasian.

A positive predicate of rows $\hat{P}$: `row` $\to$ `bool` takes a row and determines if its prediction is positive. For example, $\hat{P}(r_i) :=$ `int`$(r_i$("score")$) > 7$ means a row's prediction is positive if its score is greater than 7. A positive predicate of model results $\hat{P} : \alpha \to$ `bool` takes a model's result and determines if it is positive. In the simplest case, the model returns a `bool`, and the positive predicate can just be the identity function.

A score predicate of rows $\hat{S}$: `row` $\to \alpha$ takes a row and gives its predicted score. Similar to positive predicate, a score predicate of model results $\hat{S} : \alpha \to \beta$ takes a model's result and gives its score.

A ground truth predicate T : `row` $\to$ `bool` takes a row and gives the ground truth of the result.

A legitimate predicate $L : \prod_i \alpha_i \to$ `row` $\to$ `bool` takes n parameters and returns a row predicate. For example, $L(x)(r_i) :=$ `int`$(r_i$("priors_count")$) > x$ first take a parameter x and then decides if priors count is larger than it.

A calibration predicate $C : \prod_i \alpha_i \to$ `row` $\to$ `bool`, takes n parameters and returns a row predicate. For example, $C(u,l)(r_i) := l <$ `int`$(r_i$("score")$) < u$ first takes two parameters u, l as upper bound and lower bound, and then decides if r_i's prediction score is between them.

B. Definitions

We abstracted the idea of privileged groups and positive prediction as predicates to maximize flexibility. With the proper predicates, any model output can be used; even prose-like responses of generative models can be included given adequate predicates.

Given a dataset $\mathcal{D}$ for auditing use, and given CSV prediction results or the prediction model $\mathcal{M}$ itself, we can calculate the fairness measures by modeling the statistical random variables mentioned in Section II with our predicates. We demonstrate this process in the following for a few selected measures; all the other measures can be modeled similarly.

Fairness Measure	**Definition**
Disparate Impact	$\frac{P[\hat{Y}=1\|S\neq 1]}{P[\hat{Y}=1\|S=1]} \geq 1-\epsilon$
Demographic Parity	$\|P[\hat{Y}=1\|S=1]-P[\hat{Y}=1\|S\neq 1]\| \leq \epsilon$
Conditional Statistical Parity	$\|P[\hat{Y}=1\|S=1,L=l]-P[\hat{Y}=1\|S\neq 1,L=l]\| \leq \epsilon$
Overall Accuracy Equality	$\|P[Y=\hat{Y}\|S=1]-P[Y=\hat{Y}\|S\neq 1]\| \leq \epsilon$
Mean Difference	$\|E[\hat{Y}\|S=1]-E[\hat{Y}\|S\neq 1]\| \leq \epsilon$
Equalized Odds	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
Equal Opportunity	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
Predictive Equality	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
Conditional Use Accuracy Equality	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
	$\|P[Y=0\|S=1,\hat{Y}=0]-P[Y=0\|S\neq 1,\hat{Y}=0]\| \leq \epsilon$
Predictive Parity	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
Equal Calibration	$\|P[Y=1\|S=1,\hat{V}=v]-P[Y=1\|S\neq 1,\hat{V}=v]\| \leq \epsilon$
Positive Balance	$\|E[\hat{V}\|Y=1,S=1]-E[\hat{V}\|Y=1,S\neq 1]\| \leq \epsilon$
Negative Balance	$\|E[\hat{V}\|Y=0,S=1]-E[\hat{V}\|Y=0,S\neq 1]\| \leq \epsilon$

TABLE I: Fairness measures.

For equal opportunity, recall its formal definition:

$$|P[\hat{Y}=1|S=1,Y=1]-P[\hat{Y}=1|S\neq 1,Y=1]| \leq \epsilon$$

To model the Y, $\hat{Y}$, and S, we define the corresponding T, $\hat{P}$, and R. We have $Y=1$ if and only if T is true; $\hat{Y}=1$ if and only if $\hat{P}$ is true; $S=1$ if and only if R is true; and vice versa.

$$\text{equal\_opportunity}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$

For positive balance, its formal definition is:

$$|E[\hat{V}|Y=0,S=1]-E[\hat{V}|Y=0,S\neq 1]| \leq \epsilon$$

We model Y and S as the previous example. As for $\hat{V}$ we have it equal to the output of $\hat{S}$, and now we can calculate the measure.

$$\text{positive\_balance}(\epsilon, \mathcal{M}, R, \hat{S}, T)$$

For equal calibration, recall its formal definition:

$$|P[Y=1|S=1,\hat{V}=v]-P[Y=1|S\neq 1,\hat{V}=v]| \leq \epsilon$$

Here we model Y and S as above. As for $\hat{V}=v$, we abstracted $\hat{V}$ with C and v with $args$. We have $\hat{V}=v$ if and only if $C(args,...)$ is true.

$$\text{equal\_calibration}(\epsilon, \mathcal{M}, R, C, T, (args,...))$$

For conditional statistical parity, its formal definition is:

$$|P[\hat{Y}=1|S=1,L=l]-P[\hat{Y}=1|S\neq 1,L=l]| \leq \epsilon$$

$\hat{Y}$ and S are modeled as above. As for $L=l$, we abstracted it similarly to the case above by having $L=l$ if and only if $L(args,...)$ is true.

$$\text{conditional\_statistical\_parity}(\epsilon, \mathcal{M}, R, \hat{P}, L, (args,...))$$

This way, we can calculate each condition programmatically by testing if the predicate is true on all rows and obtain the resulting measure. All the measures can be modeled in a similar vein. So, in total, we have:

$$\text{disparate\_impact}(\epsilon, \mathcal{M}, R, \hat{P})$$
$$\text{demographic\_parity}(\epsilon, \mathcal{M}, R, \hat{P})$$
$$\text{conditional\_statistical\_parity}(\epsilon, \mathcal{M}, R, \hat{P}, L, (args,...))$$
$$\text{overall\_accuracy\_eqaulity}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$
$$\text{mean\_difference}(\epsilon, \mathcal{M}, R, \hat{P})$$
$$\text{equalized\_odds}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$
$$\text{equal\_opportunity}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$
$$\text{predictive\_equality}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$
$$\text{conditional\_use\_accuracy\_equality}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$
$$\text{predictive\_parity}(\epsilon, \mathcal{M}, R, \hat{P}, T)$$
$$\text{equal\_calibration}(\epsilon, \mathcal{M}, R, C, T, (args,...))$$
$$\text{positive\_balance}(\epsilon, \mathcal{M}, R, \hat{S}, T)$$
$$\text{nagative\_balance}(\epsilon, \mathcal{M}, R, \hat{S}, T)$$

where $\mathcal{M}$ is optional if the input contains CSV prediction results.

By calculating all the available fairness measures, we provide the auditors a comprehensive perspective on the fairness performance of a model or dataset.

We implemented the framework in Python, and it is available as a public domain open-sourced Python package at https://pypi.org/project/fairness-checker/.

IV. Application

A. Setup

In this section, we will apply the proposed framework to the ProPublica COMPAS dataset [14]–[16].

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), developed by the private company Northpointe (now Equivant), is a risk assessment software used in the American criminal justice system to evaluate the likelihood of a defendant reoffending. Defendants taking the

COMPAS test are given a questionnaire about topics ranging from family history to personal ideology. The questionnaire is then fed to the software system along with a number of parameters like the defendants' age, and then the system will assign a risk score to them from 1-10, with 10 being the highest risk.

ProPublica is an American non-profit journalism organization focused on public interests. In 2016, they conducted an investigative report into the COMPAS system. They obtained the 2013-2014 COMPAS score data of over 10,000 defendants in Florida. They also obtained criminal records of these defendants through 2016 and compared if they actually recidivate or not. They only counted misdemeanors and felonies as recidivism but not less serious crimes such as infractions. In the study, they've found that black defendants are disproportionately scored higher than they actually are, and white defendants are disproportionately scored lower than they actually are.

We shall start applying our framework. There is generally no formal guide on how to set ϵ, so we will take the 80% rule's case and set it as $\epsilon = 0.2$. For the ϵ of positive and negative balance, we shall use $\epsilon = 5$ as it is a reasonable choice.

We will set the privileged predicate be to "non-African-American" so that the unprivileged group will be African-American to test if the African-American race is discriminated.

$$R(r_i) := r_i(\text{“race”}) \neq \text{“African-American”}$$

As ProPublica referenced Northpointe's COMPAS Practitioners Guide and cited that "medium"(5-7) and "high"(8-10) categories of scores are considered to indicate a risk of recidivism, we set the positive predicate using the readily available category.

$$\hat{P}(r_i) := r_i(\text{“score\_text”}) \in \{\text{“Medium”}, \text{“High”}\}$$

Alternatively, since we know the corresponding scores of the categories, we can define it using the scores themselves, too.

$$\hat{P}(r_i) := 5 \leq \texttt{int}(r_i(\text{“decile\_score”})) \leq 10$$

On the other hand, when we're taking prediction input from our makeshift model, we simply return its result because it already gives a binary `bool` result: $\hat{P}(r_i) := \mathcal{M}(r_i)$

For the ground truth predicate, the recidivism data is already present in the ProPublica dataset, so all we have to do is a simple lookup.

$$T(r_i) := r_i(\text{“two\_year\_recid”}) == \text{“1”}$$

For the score predicate, it is again a simple lookup.

$$\hat{S}(r_i) := \texttt{int}(r_i(\text{“decile\_score”}))$$

For the legitimate predicate, we may want to look at defendants with priors. Here we shall set $args = (0)$.

$$L(0)(r_i) := \texttt{int}(r_i(\text{“priors\_count”})) > 0$$

For the calibration predicate, we may want to look at risk scores within a specific range. Here we shall set $args = (7, 5)$.

$$C(7,5)(r_i) := 5 \leq \texttt{int}(r_i(\text{“decile\_score”})) \leq 7$$

With these predicates defined, we can call the fairness measure functions and check if the measures hold or not. The results are compiled in Table II. Since some fairness measures are equivalent, we write them in the same entry.

Fairness Measure	**Criterion**	**Pass**
Disparate Impact	1.81 > 0.8	YES
Demographic Parity	0.26 < 0.2	NO
Conditional Statistical Parity	0.25 < 0.2	NO
Overall Accuracy Equality	0.02 < 0.2	YES
Mean Difference	0.26 < 0.2	NO
Equalized Odds (true positive) Equal Opportunity	0.23 < 0.2	NO
Equalized Odds (false positive) Predictive Equality	0.22 < 0.2	NO
Conditional Use Accuracy Equality (true positive) Predictive Parity	0.06 < 0.2	YES
Conditional Use Accuracy Equality (true negative)	0.06 < 0.2	YES
Equal Calibration	0.03 < 0.2	YES
Positive Balance	1.6	?
Negative Balance	1.4	?

TABLE II: Fairness measures of privileged group being Non-African-American.

B. *Fairness Analysis of African-American group*

On first blush, it is curious that it satisfies disparate impact but not demographic parity. However, if we return to the definition, we'd see that disparate impact is meant to be used when being marked positive is an advantaged thing, while here in the COMPAS example, being marked positive is a disadvantaged thing. Thus, henceforth we will exclude disparate impact from our analysis.

We can then immediately tell from the failing demographic parity, mean difference, and conditional statistical parity that COMPAS prediction results were unfair against African-American, even if we only consider the ones with prior crimes.

From the low overall accuracy equality and both conditional use accuracy equality criteria, we can tell that the accuracy is similar across African-Americans and non-African-Americans.

From the failing equalized odds we can conclude that African-Americans are indeed treated unequally by the COMPAS system even after the ground truth is taken into account. This is the same conclusion reached by the ProPublica report.

From the low equal calibration we can tell that if we only consider the medium risk score, African-Americans aren't treated fairly.

Finally, if we look at positive and negative balance, we can see that they're of similar numbers. Their average is 1.5 on a scale of 10, which means approximately a 15% difference. It remains to be said if this is fair or not. An auditor could consult a domain expert for advice on how to set a ϵ and how it is justified.

C. *Fairness Analysis of Different Races*

By setting the privilege predicate to different races we can have a more comprehensive look over the dataset. We've checked the case of privilege predicate being Non-African-American, Non-Asian, Non-Caucasian, Non-Hispanic, and Non-Native American. The results are shown in Figure 1.

From the results, we can first notice that the overall accuracy equality is low for both the Non-Asian and Non-Native American cases. This can be explained by checking the original dataset which shows that there are only 32 and 18 rows, respectively, in a dataset of 7214 rows. Hence, the accuracy is naturally lower because of the small data size.

Looking at the remaining three groups, Non-African-American, Non-Caucasian, and Non-Hispanic, their accuracies are relatively much better.

As for fairness measures, we can immediately see from that all measures are $\epsilon < 0.2$ for Non-Caucasian and Non-Hispanic that African-American are indeed treated unfairly in a broad sense.

More specifically, African-American are treated more unfairly according to demographic parity, conditional statistical parity, mean difference, and equalized odds. This is again in accordance with the conclusion of the ProPublica report.

D. Fairness Analysis of Different Groups

On the other hand, we also analyzed the case of unprivileged group being across the three age groups and the case of privileged group of sex being Male and charged degree being misdemeanor in Figure 2.

We can see that, interestingly, the group of age "25 - 45" group receives generally fair treatment. Its data size is also the largest at 4109 rows, whereas "Less than 25" age group has 1529 rows and "Greater than 45" age group has 1576 rows, so the possibility of skewed data is unlikely.

If we look closer we can see that in the age group "Less than 25", its predictive equality is noticeably worse, meaning in young people false positive rate is higher than in older people; conversely, when we look at the age group "Greater than 45", both equal opportunity and predictive equality are worse, meaning for old people their treatment is even more unfair than young people. Only the middle age group is treated fairly.

In another vein, the case of "Male" group and "Misdemeanor" group are both treated rather fairly. Furthermore, their overall accuracy equality are both excellent at the < 0.1 range. The predictive equality of "Male" group even has a measure as low as 0.000004.

E. Analysis with Model-as-input

Although this dataset scenario falls in the CSV-as-input case in our framework, we also trained a simple makeshift model using the dataset itself for demonstration.

The model was trained by splitting the original dataset into 3 partitions: 60% of the data was used for training; 20% was used for validation; and the rest 20% was used for calculating fairness measures. Validation showed that the accuracy was about 60%.

We only chose equalized odds of the case where privileged group is Non-African-American for illustrative purposes. The equalized odds of our model is 0.35 and 0.46, while that of the COMPAS dataset is 0.23 and 0.22. So, our simple naive model is more unfair than COMPAS's model and could use some more fine-tuning.

V. Conclusion

Addressing the challenges brought upon by the rapid advancements in AI technologies and its introduced bias necessitates a rigorous assessment of AI system fairness. While fairness is subjective, it can be defined through various statistical measures. Our research contributes to this effort by proposing a comprehensive framework for auditing AI systems using multiple white-box fairness metrics, such as demographic parity, equalized odds, and overall accuracy equality.

By applying this framework to the COMPAS dataset, we confirmed that African-American defendants were unfairly scored compared to other racial groups. These results align with ProPublica's findings, demonstrating the utility and accuracy of our fairness auditing tool. Developed in Python and publicly available, this tool enables third parties to conduct detailed assessments of AI systems, promoting transparency and accountability.

References

[1] D. Pessach and E. Shmueli, "A review on fairness in machine learning," ACM Computing Surveys (CSUR), vol. 55, no. 3, pp. 1–44, 2022.

[2] Supreme Court of the United States, "Griggs v. duke power co." 401 U.S. 424, March 8, 1971.

[3] The U.S. Equal Employment Opportunity Commission (EEOC), "Uniform guidelines on employee selection procedures," March 2, 1979.

[4] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[5] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," Data mining and knowledge discovery, vol. 21, pp. 277–292, 2010.

[6] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, 2017, pp. 797–806.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," Sociological Methods & Research, vol. 50, no. 1, pp. 3–44, 2021.

[8] I. Žliobaitė, "Measuring discrimination in algorithmic decision making," Data Mining and Knowledge Discovery, vol. 31, no. 4, pp. 1060–1089, 2017.

[9] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," Advances in neural information processing systems, vol. 29, 2016.

[10] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," Big data, vol. 5, no. 2, pp. 153–163, 2017.

[11] A. Fraenkel, Fairness and Algorithmic Decision Making, 2020, lecture Notes for UCSD course DSC 167.

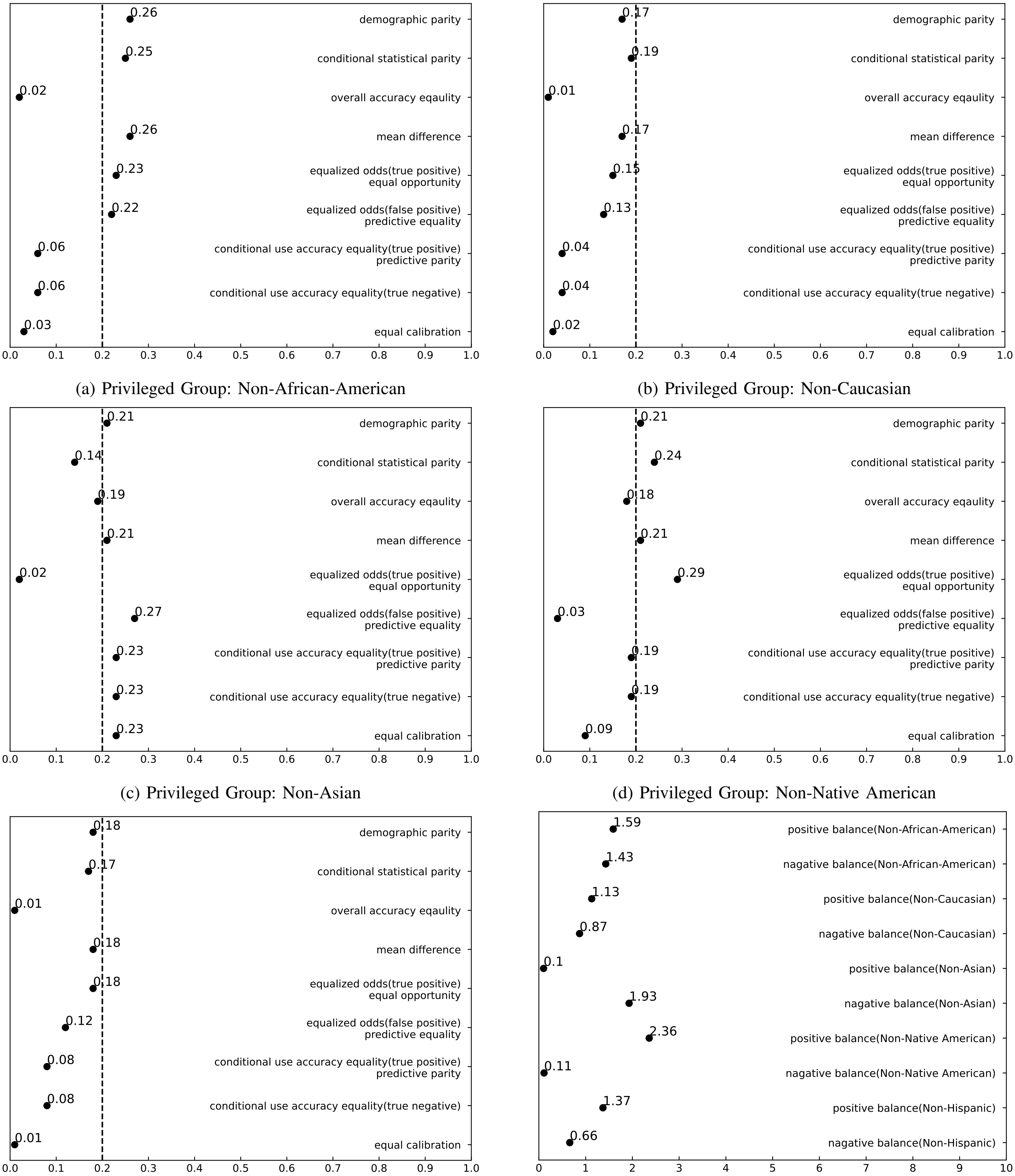
[12] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.

[13] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," arXiv preprint arXiv:1609.05807, 2016.

[14] J. Angwin, "Machine bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[15] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016.

[16] ——, "Propublica compas analysis—data and analysis for 'machine bias'," https://github.com/propublica/compas-analysis, 2016.



(a) Privileged Group: Non-African-American

(b) Privileged Group: Non-Caucasian

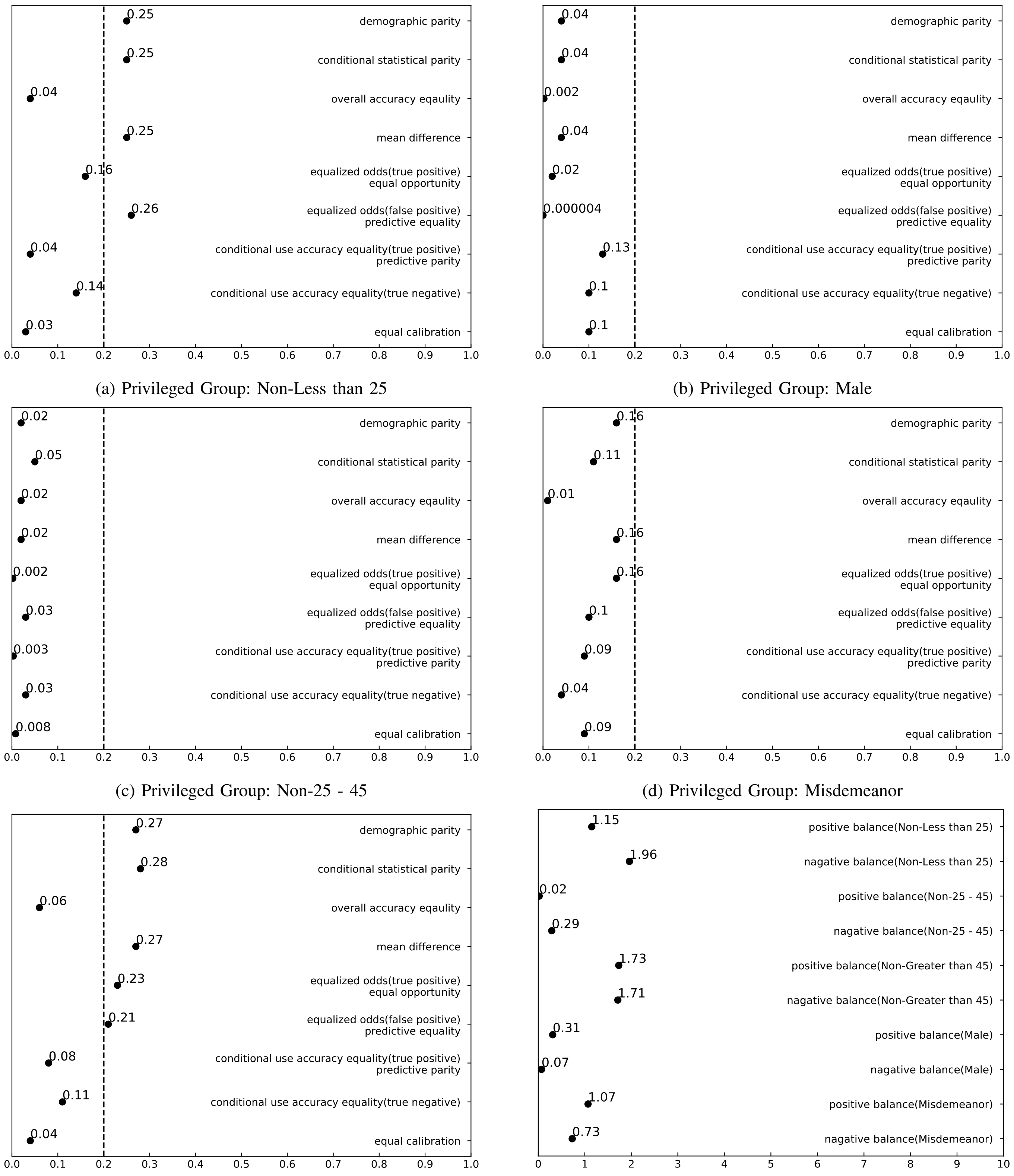
(c) Privileged Group: Non-Asian

(d) Privileged Group: Non-Native American

(e) Privileged Group: Non-Hispanic

(f) Balance

Fig. 1: Fairness measures of ProPublica COMPAS dataset grouped by race



(a) Privileged Group: Non-Less than 25

(b) Privileged Group: Male

(c) Privileged Group: Non-25 - 45

(d) Privileged Group: Misdemeanor

(e) Privileged Group: Non-Greater than 45

(f) Balance

Fig. 2: Fairness measures of ProPublica COMPAS dataset grouped by age, sex, and charged degree