From Laws to Algorithms: Detecting Unfairness in Machine Learning Models

T. Iritie[1] and D. Lenders[2]

[1] AdviceRobo, Saturnusstraat 60, 2516 AH The Hague, Netherlands
tibe@advicerobo.com
[2] Universiteit Antwerpen, Antwerp, Belgium

Abstract. As algorithms increasingly automate decisions based on potentially biased data, it's essential for these algorithms to be fairness-aware. Choosing the appropriate fairness definition is critical to ensure compliance with non-discrimination legislation, allowing it to be used as evidence in court and preventing it from being subjectively or arbitrarily chosen by system developers/controllers. Wachter, Mittelstadt, and Russell's 2021 paper advocates conditional demographic parity (CDP) as a philosophically sound and legally aligned fairness standard. We propose a fairness definition using k-NN situation testing with a custom distance function, building on Lenders and Calders' 2021 work. Our method enhances their approach by ensuring fairness comparisons are only made within similar groups and enabling the measurement of both discrimination and favouritism. Our approach offers advantages over CDP and individual fairness definitions based on k-NN.

Keywords: Individual fairness · Group fairness · Non-discrimination law · Discrimination · Favoritism · Situation testing.

1 Introduction

In today's automated decision-making, subjective human decisions are increasingly replaced by supposedly objective algorithms. However, algorithms trained on biased, human-influenced data will produce biased decisions. Relying on algorithms unaware of this bias does not bring us closer to objectivity; instead, it risks amplifying and perpetuating unfairness [8]. Unfair algorithmic outcomes reinforce existing stereotypes, creating biased datasets leading to a self-fulfilling prophecy. Addressing this requires fairness-aware algorithms (e.g., classifiers). Unlike human judgment, algorithms disclose the attributes used and their correlation to sensitive attributes like gender, ethnicity, and religion. This transparency is an advantage, as human judgment often harbors subconscious biases from group generalizations.

Integral to fairness-aware classification is the challenge of measuring and assessing fairness. Wachter, Mittelstadt, and Russell proposed conditional demographic parity (CDP) as a baseline fairness definition, aligned with the European Court of Justice (ECJ) "gold standard" [23]. The ECJ standard mandates com-

paring the composition (e.g., positive decision proportion) of deprived (e.g., females) and favored (e.g., males) groups [17]. Additionally, EU non-discrimination laws require that individuals in both groups be in similar situations, reflecting Aristotle's principle of treating similar cases similarly [2]. Discrimination may be legally justified if it serves a legitimate aim and is necessary and proportionate (e.g., a genuine job requirement). Harmonizing statistical measurements with EU non-discrimination laws is crucial to ensure fairness definitions are court-admissible and not subjectively or arbitrarily chosen by system developers/controllers. The latter would imply an unjustifiable shift in power from regulators/judges to system developers/controllers. CDP is suitable only for datasets where class labels may contain some unwanted bias, meaning labels represent human decisions prone to bias (e.g., recidivism risk). Independence criteria, such as CDP, use only features and predicted labels, excluding the class label. Without a bias-free class labels, fairness definitions using the class label should not be applied, making only independence criteria suitable [20]. For objective measurements or indisputable facts (e.g., image recognition), the bias-free class labels (i.e. correct outcome) are directly measurable or unambiguously observable.

However, Wachter, Mittelstadt, and Russell's work has a limitation in its lack of distinction between group fairness and individual fairness [23]. The key difference between CDP and individual fairness lies in how explanatory attributes define groups requiring equal treatment. Explanatory attributes can justify "unfairness" (e.g., genuine job requirements), so they must be incorporated into fairness definitions. CDP mandates equal positive rates among groups with a common explanatory attribute value or within a cluster based on multiple explanatory attributes. Individual criteria compare the positive rates of one or more groups similar to a given instance, with similarity measured by a distance function applied to all explanatory attributes. The size of the groups being compared for existing individual criteria is also restricted; for example, a deprived instance is only compared to the k nearest favored instances [5, 15].

A drawback of individual fairness definitions is the need for a defined distance function. Luong, Ruggieri, and Turini defined such a distance function to apply the legally grounded methodology of situation testing, where pairs of similar individuals—differing only in a sensitive attribute—are compared [16, 3, 19]. Building on this, Lenders and Calders developed an optimization algorithm to learn a weighted distance function from data to compare similar individuals and assess potential discrimination [15]. Their algorithm assigns high weights to attributes relevant to the decision task (e.g., explanatory attributes) and low weights to redlining attributes. Redlining attributes correlate with the class label mainly through the sensitive attribute, implying that given the sensitive attribute, the class label is only weakly correlated with the redlining attribute. For example, postal codes may correlate with loan approval decisions but cannot justify unfair decisions. Postal codes are also linked to sensitive attributes like ethnicity, making them prone to unfairness due to historical racism in housing [4, 1].

However, Lenders and Calders' approach has limitations: it does not measure reverse unfairness and favoritism, restricts the group size of "similar" instances to a constant k, does not assess the significance of unfairness results, and only accommodates interval-scaled attributes [15]. Individual fairness focuses on fairness for each instance, making it important to distinguish the specific sensitive group to which the instance belongs. In contrast, CDP assesses fairness at the group level, comparing outcomes across entire groups, and does not capture individual-level biases. Therefore, discrimination and favoritism are uniquely relevant to individual fairness, not to CDP. Individual unfairness measured from the perspective of a deprived instance is referred to as discrimination, while that measured from the perspective of a favored instance is referred to as favoritism.

In individual fairness, an instance is compared with similar ones, typically using a k-nearest neighbors approach. When a deprived instance is compared with its k nearest favored neighbors, only the denial rate for positive outcomes in the favored group is considered [5, 15]. This approach overlooks apparent biases against favored instances by ignoring the denial rate of the nearest deprived group. Reverse unfairness from the perspective of a deprived instance measures this apparent bias against the favored group (e.g., a high denial rate). If both groups are denied the same proportion of positive decisions, no unfairness regarding the sensitive attribute exists. Like discrimination and favoritism, reverse unfairness is an individual-level bias. To properly assess individual unfairness, both biases against deprived and favored instances must be considered.

This paper introduces a new fairness definition, expanding on Lenders and Calders' work to detect unfairness in both datasets and machine learning (ML) models [15]. Our definition offers five key innovations: it measures reverse unfairness and favoritism, is based solely on individual similarity, evaluates the significance of unfairness, and supports mixed data types. Additionally, it can be integrated with unfairness prevention techniques to eliminate biases in ML models. We advocate using our definition over CDP because it ensures only similar groups are compared and provides unfairness scores that quantify discrimination or favoritism for each instance. We validate our approach using the COMPAS dataset and a dataset from the credit lender AdviceRobo, where the decision task is loan granting and the sensitive attribute is the applicant's language [14]. Although the latter results are not discussed in this paper, they are available on GitHub and support our COMPAS findings.[3]

2 Methodology and Data

In this section, we present our methodology and data, including notation, fairness definitions, a glossary, evaluation methods, and the dataset description.

[3] Source code: https://github.com/python211223/fairness-aware-classification.

2.1 Notation

We consider a two-class classification problem with class labels $c \in \{+, -\}$ over n instances $\boldsymbol{x} \in \mathcal{X}$ with p attributes in an unfair dataset $\mathcal{D} = (x_{ji}, c_j) \in \mathbb{R}^{n \times (p+1)}$. An unfair dataset implies the class labels may contain some bias (i.e., the ground truth is not available). The binary sensitive attribute S takes values s_i or $s \in \{d, f\}$, indicating whether an instance is in the deprived group ($s = d$) or the favored group ($s = f$). Superscripts of $\mathcal{X}$ denote membership of instances in the deprived ($\mathcal{X}^d$) or favored ($\mathcal{X}^f$) group, and the presence of the desired/positive ($\mathcal{X}^+$) or undesired/negative ($\mathcal{X}^-$) class label. The explanatory attribute E has h different values denoted e_i or e. Instances with the same attribute value e_i are indicated as $\mathcal{X}^{e_i}$. Cardinalities of sets are denoted by vertical bars; for example, $|\mathcal{X}^{f+}|$ represents the number of favored instances with a positive label. The fairness definitions in this paper measure fairness in a dataset but can also assess the fairness of classifier fairness using the labeled test data as dataset.

2.2 Methodological Background

Group Fairness To measure discrimination, Kamiran, Žliobaitė, and Calders used the difference in positive rates between the favored and deprived groups:

$$D_{all}(D, S) := \frac{|\mathcal{X}^{f+}|}{|\mathcal{X}^f|} - \frac{|\mathcal{X}^{d+}|}{|\mathcal{X}^d|}, \tag{1}$$

where demographic parity (DP) is satisfied if $D_{all} = 0$ [10]. In probability terms, D_{all} represents the difference in positive class probabilities between $\boldsymbol{x} \in \mathcal{X}^f$ and $\boldsymbol{x} \in \mathcal{X}^d$.[4] To determine which part of D_{all} is fair (justified due to an explanatory attribute), they defined the probability of observing $\boldsymbol{x} \in \mathcal{X}^+$ for the explanatory attribute e_i in the absence of discrimination as follows:

$$P^*\left(c^+ \mid \mathcal{X}^{e_i}\right) := \frac{P\left(c^+ \mid \mathcal{X}^{e_i}, \mathcal{X}^f\right) + P\left(c^+ \mid \mathcal{X}^{e_i}, \mathcal{X}^d\right)}{2}. \tag{2}$$

They assumed the same fraction of favored instances benefit from discrimination as the fraction of deprived instances disadvantaged. Thus, they used the average $P(c^+ \mid \mathcal{X}^{e_i}, \mathcal{X}^s)$ for $s \in \{d, f\}$.[5] To measure fair discrimination (i.e., explainable discrimination), Kamiran, Žliobaitė, and Calders (2013) used

$$D_f(D, S, E) := \sum_{i=1}^{h} \left[P(e_i \mid \mathcal{X}^f) - P(e_i \mid \mathcal{X}^d)\right] P^*\left(c^+ \mid \mathcal{X}^{e_i}\right). \tag{3}$$

D_f measures the probability difference in observing $\boldsymbol{x} \in \mathcal{X}^+$ between favored and deprived instances if each instance with a fixed explanatory attribute value e_i had the same chance of receiving a positive class label, independent of its sensitive

[4] $D_{all}(D, S) := P\left(c^+ \mid \mathcal{X}^f\right) - P\left(c^+ \mid \mathcal{X}^d\right)$

[5] $P(c^+ \mid \mathcal{X}^{e_i}, \mathcal{X}^s)$ is a short notation for $P(c = + \mid E = e_i, S = s)$.

attribute value. To measure group unfairness conditioned on one explanatory attribute, we subtract fair discrimination D_f in (3) from the total discrimination D_{all} in (1):

$$D_u(D, S, E) := D_{\text{all}}(D, S) - D_f(D, S, E) \tag{4}$$

[10]. CDP is satisfied when $D_u = 0$ and assumes that all members conditioned on e_i are similar.

To address the limitation of using only one explanatory attribute, Kamiran, Žliobaitė, and Calders proposed an approach using k-means clustering, where clusters are defined by all available explanatory attributes, referred to as CDP-ME [10]. CDP-ME requires equal positive rates within each cluster rather than within groups similar in just one explanatory attribute. They excluded attributes highly correlated with the sensitive attribute (e.g., redlining attributes) from clustering. A drawback of CDP(-ME) is that the selection of explanatory attributes must be determined externally by law or domain experts. Additionally, Kamiran, Žliobaitė, and Calders did not apply weights to attributes or provide guidance on selecting the number of clusters [10]. Since not all explanatory attributes may be equally relevant to the class label, weighting them by relevance is preferable. The number of clusters affects within- and between-cluster similarity and, consequently, adherence to the ECJ's principle of equal treatment.

Individual Fairness Lenders and Calders provided an optimization algorithm to learn the relevance weight $\boldsymbol{w}$ for each attribute by minimizing

$$\begin{aligned}\frac{1}{|C(\mathcal{X}^d)|}\sum_{\boldsymbol{x},\boldsymbol{x'}\in C(\mathcal{X}^d)} d^2_{\boldsymbol{w}}(\boldsymbol{x},\boldsymbol{x'}) + \frac{1}{|C(\mathcal{X}^f)|}\sum_{\boldsymbol{x},\boldsymbol{x'}\in C(\mathcal{X}^f)} d^2_{\boldsymbol{w}}(\boldsymbol{x},\boldsymbol{x'}) \\ - \frac{1}{|C'(\mathcal{X}^d)|}\sum_{\boldsymbol{x},\boldsymbol{x'}\in C'(\mathcal{X}^d)} d^2_{\boldsymbol{w}}(\boldsymbol{x},\boldsymbol{x'}) - \frac{1}{|C'(\mathcal{X}^f)|}\sum_{\boldsymbol{x},\boldsymbol{x'}\in C'(\mathcal{X}^f)} d^2_{\boldsymbol{w}}(\boldsymbol{x},\boldsymbol{x'}) \\ + \lambda\|\boldsymbol{w}\|_2^2, \end{aligned} \tag{5}$$

where C denotes the instances with the same class label, C' represents instances with a different class label, and $\lambda\|\boldsymbol{w}\|_2^2$ defines an L2 regularizer [15]. Attributes correlated with the class label only through a sensitive attribute (i.e., redlining attributes) receive a weight of 0 in the optimal solution, as they are conditionally independent of the class label and should be excluded from the distance function. Attributes that are uncorrelated with the class label also receive near-zero weights since they only increase the distances in the last two sums of (5). Sequential least squares programming (SLSQP) is used to minimize (5).[6] The optimized weights $\boldsymbol{w} = (w_1, \ldots, w_{p-1})$ are applied in a distance function to find the k most similar (nearest) instances for which equal treatment is required:

$$d^E_w(\boldsymbol{x},\boldsymbol{x'}) := \sqrt{\sum_{i=1}^{p-1} w_i(x_i - x'_i)^2}. \tag{6}$$

[6] SLSQP is a quasi-Newton method that approximates the region around the optimum of (5) with a quadratic function.

Nominal attributes must be one-hot encoded, and interval-scaled attributes should be Z-score standardized or min-max normalized to avoid dominance in the distance calculation. To measure individual fairness Lenders and Calders introduced the D_k-score:

$$D_k(\boldsymbol{x}) := \frac{|\mathcal{X}_k^{f+}|}{|\mathcal{X}_k^{f}|}, \tag{7}$$

where for each observation $\boldsymbol{x} \in \mathcal{X}^{d-}$, the positive rate among its k nearest favored neighbors is measured [15]. To quantify the unfairness of a dataset we use the average D_k-score over all deprived instances with a negative class label:

$$U_k(D, S) := \frac{\sum_x D_k(x)}{|\mathcal{X}^{d-}|}. \tag{8}$$

2.3 Methodological Contribution

Unfairness Scores beyond Discrimination A drawback of using (7) to measure individual fairness is that the maximum allowable distance for a neighbor to be considered near is not restricted. Without such a restriction, distant neighbors may be included in the k-nearest neighbors set, potentially compromising similarity. This is problematic since the legal methodology of situation testing and the ECJ's principle of equal treatment emphasize comparing similar individuals. Therefore, it may be preferable to base the nearest neighbors solely on their similarity without restricting the number of neighbors. Additionally, using k-nearest neighbors can yield unstable results if the dataset contains duplicate rows with different class labels, as choosing the k nearest neighbors among duplicates with the same distance but different labels is challenging. Another limitation of (7) is that it only measures biases against $\boldsymbol{x} \in \mathcal{X}^d$ and does not measure biases against $\boldsymbol{x} \in \mathcal{X}^f$ (reverse unfairness) or biases in favor of $\boldsymbol{x} \in \{\mathcal{X}^d, \mathcal{X}^f\}$ (favoritism). To address these issues, we introduce new unfairness scores:

$$D_m(\boldsymbol{x}) := \frac{|\mathcal{X}_m^{f+}|}{|\mathcal{X}_m^{f}|} - \frac{|\mathcal{X}_m^{d+}|}{|\mathcal{X}_m^{d}|}, \tag{9}$$

$$F_m(\boldsymbol{x}) := \frac{|\mathcal{X}_m^{d-}|}{|\mathcal{X}_m^{d}|} - \frac{|\mathcal{X}_m^{f-}|}{|\mathcal{X}_m^{f}|}. \tag{10}$$

In (9), for each $\boldsymbol{x} \in \mathcal{X}^{d-}$, the positive rate among the deprived neighbors within a distance m is subtracted from the positive rate among the favored neighbors within a distance m. The vice-versa case applies to each $\boldsymbol{x} \in \mathcal{X}^{f+}$ in (10), which measures favoritism. The subtraction in (9) and (10) ensures that reverse unfairness is also accounted for.

Setting m For this study, we calculate each instance's distance to its nearest neighbor and set m (9) and (10) equal to $Q3 + 1.5 \times \text{IQR}$, where $Q3$ is the third quartile and IQR is the interquartile range of these distances. We use the

maximum non-outlier value (i.e., $Q3 + 1.5 \times \text{IQR}$) to exclude distant neighbors ("outliers") when calculating unfairness scores [21]. Thereby, we assume that all neighbors within a distance m are similar. If an instance has no neighbor with a different sensitive attribute within a distance m, it receives no unfairness score due to lack of statistical evidence for unfairness. Nevertheless, human auditors can still provide proof using non-statistical evidence for unfairness; however, these types of evidence are beyond the scope of this paper [23]. When no neighbor with the same sensitive attribute is found within a distance m, the subtraction in (9) and (10) will be equal to zero, ensuring that unfairness scores are not based on distant (dissimilar) neighbors.

Significance of Unfairness Score Lenders and Calders only consider $D_k > t$ as unfair, as large values for t will not consider instances with small D_k-scores as unfair [15]. Ideally, t should be based on existing discrimination laws or clear guidelines, but these are not always available. They therefore set t equal to the maximum non-outlier value of D_k-scores for $\boldsymbol{x} \in \mathcal{X}^{f-}$, defined as the largest value within $Q3 + 1.5 \cdot \text{IQR}$. This approach assumes that the class labels of favored instances are fair (i.e., favoritism does not exist), but since favoritism might also be present, t should not be set based on favored instances' class labels. Lenders and Calders and Kamiran, Žliobaitė, and Calders also did not address the statistical significance of their fairness results [15, 10]. Assessing statistical significance is crucial to determine whether the observed unfairness is not due to random chance. Rather than setting a specific threshold value for t, we consider an unfairness score as unfair if it passes a statistical significance test, specifically the two-sided two-proportion Z-test. We use D_m defined in (9) to explain the confidence intervals (CI's), but the same methodology can be applied to F_m in (10) by replacing all references to (9) with (10). The CI for the two-proportion Z-test of the unfairness score in (9) is calculated as follows:

$$CI_{2p} := \hat{p}_1 - \hat{p}_2 \pm z \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \tag{11}$$

where $\hat{p}_1$ and $\hat{p}_2$ represent the first and second terms in (9), respectively, and $\hat{p}_1 - \hat{p}_2$ equals the unfairness score. z is the z-score for the desired confidence level, and n_1 and n_2 are the denominators of the first and second terms in (9). The large sample size assumptions for the two-proportion Z-test are $n_1 \times \hat{p}_1 \geq 5$, $n_1 \times (1-\hat{p}_1) \geq 5$, $n_2 \times \hat{p}_2 \geq 5$, and $n_2 \times (1-\hat{p}_2) \geq 5$ [24]. If these assumptions are not met, a one-proportion Z-test is used. Since one proportion is nonnegative, a one-sided confidence interval is used with the alternative hypothesis that the proportion is greater than zero. The lower bound of the CI for the one-proportion Z-test is:

$$CI_{1p} := \hat{p} - z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{12}$$

with $\hat{p}$ as the unfairness score, z as the z-score, and n as the denominator of the first term in (9). The large sample size assumptions for the one-proportion Z-test

are $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$ [24]. We use the commonly used 5% significance level, with z-scores of ∼1.96 for the two-sided Z-test and ∼1.645 for the one-sided Z-test. If the assumptions are unmet or the unfairness score is insignificant, the instance does not receive an unfairness score due to insufficient evidence for unfairness.

Aggregating Unfairness Scores To express the unfairness of a dataset calculated with (9) and (10), we introduce the fairness definition:

$$U_m(D,S) := \frac{\sum_{\boldsymbol{x}} D_m(\boldsymbol{x}) + \sum_{\boldsymbol{x}} F_m(\boldsymbol{x})}{|D_m(\mathcal{X}^{d-})| + |F_m(\mathcal{X}^{f+})|}, \tag{13}$$

where $D_m(\mathcal{X}^{d-})$ and $F_m(\mathcal{X}^{f+})$ denote the sets of given unfairness scores in a dataset. In (13), the unfairness scores of all instances that received an unfairness score are summed and divided by the total number of instances that received a score. This ensures that $-1 \leq U_m \leq 1$, facilitating comparison with group unfairness measured by (1) and (4). U_m represents the average unfairness score of all instances that received a score, with $U_m > 0$ indicating unfairness against $\boldsymbol{x} \in \mathcal{X}^d$ and $U_m < 0$ indicating reverse unfairness against $\boldsymbol{x} \in \mathcal{X}^f$. The proportion of instances receiving an unfairness score is defined as

$$U_n(D,S) := \frac{|D_m(\mathcal{X}^{d-})| + |F_m(\mathcal{X}^{f+})|}{|\mathcal{X}^{d-}| + |\mathcal{X}^{f+}|}. \tag{14}$$

When two nearby instances with different sensitive attributes both receive positive unfairness scores, there is a high likelihood that at least some neighboring instances influence the unfairness scores of both instances. Thus, relabeling one instance is likely to reduce the unfairness score of the other. The decision on which instance to relabel involves balancing fairness, as measured by (13), with the performance (e.g., accuracy) of a ML model post-relabeling. Ideally, fairness should be improved with minimal performance loss [7]. Conversely, distant instances with different sensitive attributes and positive unfairness scores do not impact each other's unfairness. For nearby instances with different sensitive attributes and negative unfairness scores, there is a high likelihood that at least some neighboring instances contribute to the reverse unfairness for both instances. The same principle applies to distant instances with different sensitive attributes and negative unfairness scores.

Redefined Distance Function The weighted Euclidean distance in (6) is applicable only to interval-scaled attributes. To address this, we use a new distance function based on Podani's distance to calculate U_k in (8) and U_m in (13) [18]). Podani's distance extends Gower's distance by also incorporating ordinal attributes. We further adapt Podani's distance by including weights, which we refer to as the weighted Podani's distance [18]:

$$d^P_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x'}) := \sqrt{\sum_{i=1}^{p-1} w_i \left(\frac{x_i - x'_i}{s_{x_i, x'_i}} \right)^2}. \tag{15}$$

For a binary attribute i, $s_{x_i x'_i} := 1$. For a nominal attribute i, $s_{x_i x'_i} := x_i - x'_i$ if $x_i \neq x'_i$, and $s_{x_i x'_i} := 1$ if $x_i = x'_i$. For an interval-scaled or ordinal attribute i, $s_{x_i x'_i} := \max_i - \min_i$ if $\max_i \neq \min_i$, and $s_{x_i x'_i} := 1$ if $\max_i = \min_i$. We compute the weights $\boldsymbol{w}$ similarly to Lenders and Calders [15]. Ordinal attributes are label-encoded. (15) avoids the need for one-hot encoding of nominal attributes, thus mitigating the curse of dimensionality. We include all attributes except sensitive ones to evaluate the effectiveness of minimizing (5) in filtering out irrelevant attributes. When referring to an instance being near/similar to instance $\boldsymbol{x}$, we use (15), indicating proximity based only on relevant attributes.

2.4 Glossary

In Table 1, the definitions of the most important terms are provided.

Table 1: Definitions of Terms Used

Term	**Definition**
D_k-score	The positive decision rate among the k nearest favored neighbors for a deprived instance; see (7)
D_m-score	The difference in positive decision rates between favored and deprived neighbors within a distance m for a deprived instance; see (9)
F_m-score	The difference in negative decision between favored and deprived neighbors within a distance m for a favored instance; see (10)
U_k	Average D_k-score; see (8)
U_m	Average unfairness score (D_m-/F_m-score); see (13)
U_n	The proportion of instances that received an unfairness score; see (14)
Unfairness score	D_m- or F_m-score

2.5 Experiments

We begin by addressing the attribute weights obtained by minimizing equation (5), using $\lambda = 0.09$, consistent with the approach of Lenders and Calders [15]. We also adopt their starting values and bounds for weights: $w = \begin{bmatrix} 0.1 \cdots 0.1 \end{bmatrix}$ and $1 \cdot 10^{-14} \leq w_i < \infty$ [15]. Next, we analyze group similarity under our fairness definition in (13) compared to that of Lenders and Calders in (7). We set the distance threshold m in (13) equal to the maximum non-outlier value distance from instances to their nearest neighbor; however, it remains crucial to manually examine if the value of m is appropriate. Thus, we begin our analysis by examining the histograms of distances from instances to their nearest neighbor. We also use histograms of distances to neighbors with different sensitive attributes to provide insight into evidence of unfairness. If the nearest neighbor is far, other neighbors will also be distant, and using these neighbors in (8) will therefore violate the ECJ's principle of equal treatment. Finally, we provide histograms of the number of neighbors used for each unfairness score, as the number of neighbors influences the significance of these scores.

Unlike Lenders and Calders' fairness definition, our unfairness scores measure discrimination and favoritism [15]. Consequently, we present histograms and means of the unfairness scores to assess (reverse) discrimination and favoritism. Additionally, we show histograms after relabeling all deprived instances with $D_m > 0$ to identify which D_m- and F_m-scores are influenced by the same neighboring instances. If no (reverse) favoritism remains after relabeling, it implies that all unfair deprived instances were near unfair favored ones. Furthermore, we compare our fairness definition with Lenders and Calders' by providing histograms of D_k-scores in (7) [15]. Lenders and Calders tuned k in (7) to maximize accuracy when predicting class labels of favored instances. However, because favoritism can exist in a dataset, k should not be tuned for predicting "unfair" class labels. Instead, we use the median number of near favored neighbors for $\boldsymbol{x} \in \mathcal{X}^{d-}$ in U_m in (13).

After evaluating group similarity for individual fairness, we assess group similarity and the number of groups (clusters) for CDP-ME using the silhouette score [11]. The silhouette score, ranging from [-1,1], measures how similar instances are to their own cluster compared to other clusters. A score near 1 indicates strong clustering (i.e., the instance is far from neighboring clusters), a score near 0 implies that an instance lies close to the decision boundary between two neighboring clusters, and a score below 0 indicates that an instance is assigned to the wrong cluster. An average silhouette score above 0.7 implies a strong cluster structure, 0.51-0.7 indicates reasonable clustering, and below 0.51 indicates poor clustering. Poor clustering implies low within-cluster similarity and/or high between-cluster similarity. Requiring equal treatment within a cluster with dissimilar instances, or allowing different treatments between clusters with similar instances, violates the ECJ's principle of equal treatment. Therefore, it is crucial that instances within a cluster are similar and that those from different clusters are dissimilar. We only use relevant attributes ($w_i > 0.01$), determined by minimizing (5). Finally, we compare group unfairness using D_{all} in (1) with unfairness measured by U_m in (13) for each cluster.

2.6 Data

We evaluate our methods using the COMPAS dataset, excluding missing values [14]. This dataset was chosen for its inherent unfairness, measured using D_{all} from (1), and its size ensures adequately sized bins for CDP in (4), avoiding misleading results [9]. The COMPAS dataset features criminal defendants from Broward County, Florida, assessed for recidivism risk with scores categorized into "low" (1–4) and "medium + high" (5–10) [13]. As recidivism risk is a proxy rather than an objective measure, we consider the class labels potentially biased.[7] We use all attributes except *sex*, IDs, case numbers, names, dates, *age_cat*, non-recidivism COMPAS scores, attributes with many missing values,

[7] Actual recidivism within two years (ground truth) in the COMPAS dataset is excluded to enable its use for our fairness definition. Assessing the fairness of COMPAS scores with all available information is beyond this paper's scope.

and charge descriptions. *Race*, the sensitive attribute, is binary (white or black) with other races excluded [12]. *Priors_ count* is used for CDP as it accounts for most discrimination ($D_{all} \approx 0.24$, $D_u \approx 0.19$).

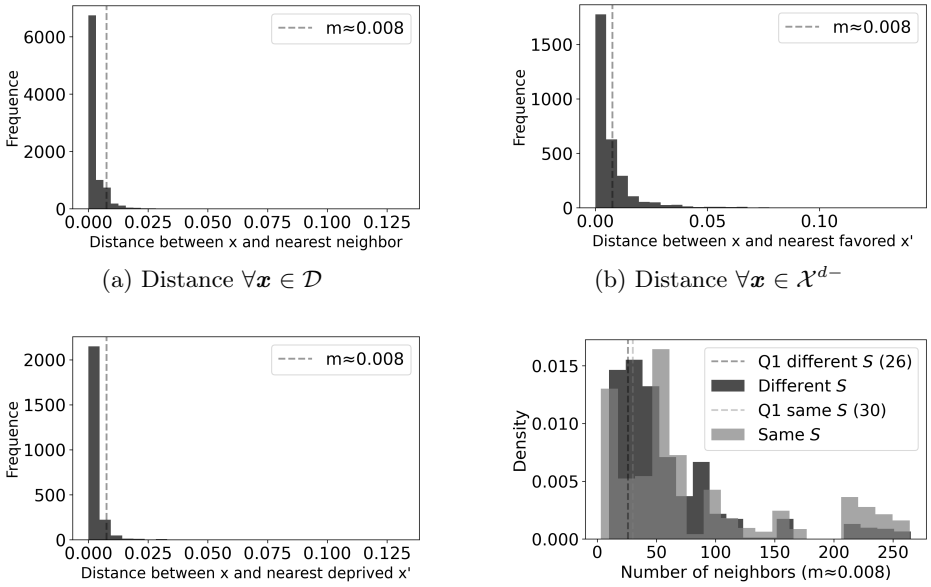
3 Results

3.1 Individual Fairness

Distribution of Minimum Distances The most relevant attributes in the COMPAS dataset are *is_ recid* and *r_ charge_ degree*, while five attributes have low weights (exact weight values are on GitHub). As expected, an individual's prior offense history (*is_ recid*) and the severity of that offense (*r_ charge_ degree*) are key factors in determining recidivism risk. Since multiple relevant attributes exist, fairness definitions that cannot incorporate all relevant attributes (e.g., CDP) are unable to accurately assess unfairness. Figure 1a shows the frequency distribution of the distances from all the instances to their nearest neighbor. For the COMPAS dataset, m in (13) is set to ∼0.008, the maximum non-outlier value in Figure 1a. Most instances have a near neighbor at a distance of ∼0.0. Some instances do not have any near neighbors, and the most isolated instance lies at a distance of ∼0.13 from its nearest neighbor. Isolated instances should not be included in group comparisons, as they lack comparable (i.e., similar) neighbors.

Histograms of the distances from $\boldsymbol{x} \in \{\mathcal{X}^d, \mathcal{X}^f\}$ to their nearest neighbor with a different sensitive attribute are shown in Figure 1b and 1c. Most instances have a nearest neighbor with a different sensitive attribute at a distance of ∼0.0, with maximum distances of ∼0.14 and ∼0.13, respectively. In Figure 1b, for example, the relevant attributes causing the distance of 0.14 are *priors_ count* and *age*, implying that the two instances are equal in terms of all other relevant attributes. The deprived instance in question is 45 years old with 38 prior crimes, while the favored instance is 56 with 17 priors. Since m is based on dataset-specific weights and attribute values, it can vary between datasets. For the COMPAS dataset, a distance of 0.14 already indicates a large difference between instances. Distances in Figures 1b and 1c are between nearest neighbors, implying that a fairness definition based on k nearest neighbors (e.g., $k = 1$) would already require equal treatment of dissimilar instances. It is even possible that more instances exist that are (more) isolated, as only histograms for the nearest neighbors are plotted. Figure 1d displays the distribution of the number of neighbors for significant unfairness scores, with Q1 indicating the first quartile. It reveals that significant scores are based on relatively many neighbors. Using all near neighbors within a distance m for an instance would be more accurate and fairer than using only the k nearest neighbors with the fairness definition of Lenders and Calders [15].

Distribution of Unfairness Scores For ∼17% of $\boldsymbol{x} \in \{\mathcal{X}^{d-}, \mathcal{X}^{f+}\}$, no near neighbors exist; about 44% of these instances fail to meet the assumptions of the Z-test, and ∼27% have insignificant unfairness scores. Consequently, the propor-



(a) Distance $\forall \boldsymbol{x} \in \mathcal{D}$

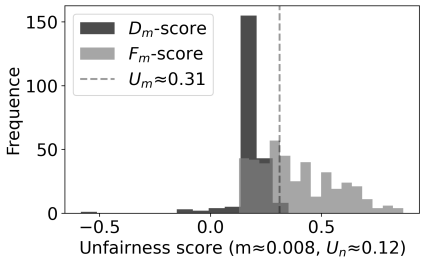
(b) Distance $\forall \boldsymbol{x} \in \mathcal{X}^{d-}$

(c) Distance $\forall \boldsymbol{x} \in \mathcal{X}^{f+}$

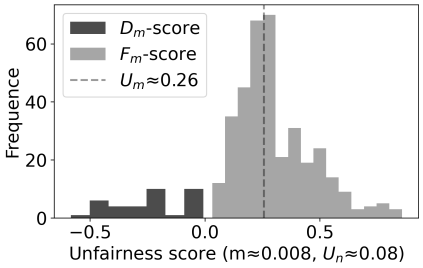
(d) Distribution of number of neighbors

Fig. 1: Histograms determining the unfairness scores

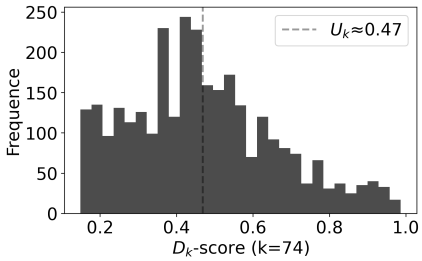
tion of instances that received an unfairness score (U_n) is ∼0.12. Since many instances have few near neighbors and/or low unfairness scores, it is crucial to exclude distant neighbors and assess the significance of unfairness scores. Histograms of p-values for all scores can be found in our code on GitHub. Figure 2a gives the frequency distribution of all the significant unfairness scores. Most D_m-scores are around 0.17, while most F_m-scores are around 0.35, resulting in an average unfairness score of $U_m \approx 0.31$. Some D_m-scores are negative, while all F_m-scores are nonnegative, indicating reverse discrimination but no reverse favoritism. Figure 2b shows the distribution of the unfairness scores after relabeling all deprived instances with $D_k > 0$. Favoritism persists post-relabeling, indicating that most unfair favored instances are not near unfair deprived instances. Thus, examining only deprived instances is insufficient, as favoritism can occur without discrimination. Figure 2c presents the D_k-scores for deprived instances. According to Lenders and Calders, $D_k > t$ is considered unfair, where t is set equal to $Q3 + 1.5 \cdot \text{IQR} \approx 1.12$ of the D_k-scores of favored instances [15]. Since $Q3 \approx 0.59$ is relatively high due to the prominent discrimination against favored instances, $t > 1$, meaning no deprived instance is deemed unfair. Conversely, D_m in (9) identifies 221 deprived instances as unfair. The discrepancies between the unfair instances identified by D_m and D_k highlight the need to compare only similar instances, consider reverse unfairness, and assess score significance—factors that D_k does not address.



(a) Distribution of D_m-/F_m-scores



(b) Distribution of D_m-/F_m-scores after relabeling



(c) Distribution of D_k-scores

Fig. 2: Histograms of different unfairness scores

3.2 Group Fairness

For CDP-ME, eight clusters provided the best solution, with an average silhouette score of 0.46. Silhouette scores below 0.51 indicate poor clustering, which implies similar clusters and/or clusters with dissimilar instances. The most dissimilar instances within a cluster are at a distance of $0.33 > m$, due to the relevant attribute *r_ charge_ degree*, implying similarity in all other relevant attributes. The deprived instance in question has no charge, whereas the favored instance has a charge degree of 1. CDP-ME requires equal treatment within each cluster. To ensure this, all similar instances must be in the same cluster (i.e., no similar clusters should exist), while dissimilar instances should be in different clusters. Poor clustering thus violates the ECJ's principle of equal treatment. The group unfairness using D_{all} in (1) for each cluster is 0.25, 0.13, 0.14, 0.21, 0.14, 0.21, 0.12, and 0.17. However, when using U_m in (13), three clusters have no unfair instances as the proportion Z-test assumptions are not met (see GitHub for detailed cluster information). This indicates that the unfairness measured for these instances is too small and/or the number of neighbors is insufficient. The differences between unfairness measured with D_{all} and U_m highlight the need to assess the significance of unfairness in clusters.

4 Discussion and Conclusion

This study addresses the limitations of existing fairness definitions, particularly their inability to ensure comparisons among similar groups and account for both discrimination and favoritism. Our fairness definition offers a holistic assessment by incorporating individual similarity, reverse unfairness, and favoritism, improving fairness evaluations in datasets and ML models. One key finding is that our method improves on Lenders and Calders' k-nearest neighbors approach, which risks comparing dissimilar instances and violates the principle of treating like cases alike [15]. By introducing a threshold distance m to ensure only similar instances are compared, our approach enables more suitable fairness evaluations, preventing misleading fairness assessments from inappropriate comparisons. Another key insight is that our method addresses both discrimination and favoritism. This dual focus is crucial for comprehensive fairness evaluation, as fairness involves not only mitigating harm to deprived groups but also preventing undue benefits to favored ones. Our approach supports more balanced fairness-aware algorithm design and opens the door to flexible unfairness prevention by guiding the relabeling of both deprived and favored instances based on their impact on overall model performance. Our approach ensures more reliable between- and within-group similarity compared to clustering-based methods like CDP-ME. CDP-ME's suboptimal silhouette scores suggest poor clustering, potentially violating equal treatment within groups. Unlike CDP(-ME), our method identifies which group instances have the most unfair class labels, aiding unfairness prevention.

One limitation of our method, common to many existing approaches, is the requirement for a binary sensitive attribute. This poses challenges when the sensitive attribute is non-binary, unobservable, or restricted by GDPR. For non-binary attributes, values can be assigned to represent the deprived group; however, assessing unfairness becomes challenging when the sensitive attribute is unobservable or prohibited [10, 22]. Future research could investigate how to evaluate fairness without sensitive attributes. Furthermore, all fairness definitions discussed in this paper lack insights into missing explanatory attributes, and quantifying some moral values complicates their integration into a distance function or CDP [6]. While similar treatment is essential for fairness, a trade-off between a fairness definition and potentially biased human judgment is necessary to ensure that no critical information is overlooked in decision-making.

References

1. Alexander, M.: The new Jim Crow. Ohio St. J. Crim. L. **9**, 7 (2011)
2. Barnes, J., et al.: The complete works of Aristotle, volumes I and II. Princeton: Princeton University (1984)
3. Bendick, M.: Situation testing for employment discrimination in the United States of America. Horizons Stratégiques (3), 17–39 (2007)
4. Dedman, B., et al.: The color of money. Atlanta Journal-Constitution **1** (1988)

5. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
6. Fleisher, W.: What's fair about individual fairness? In: Proc. 2021 AAAI/ACM Conf. AI, Ethics, and Society. pp. 480–490 (2021)
7. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
8. Kamiran, F., Karim, A., Verwer, S., Goudriaan, H.: Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. In: 2012 IEEE 12th Int. Conf. Data Mining Workshops. pp. 370–377. IEEE (2012)
9. Kamiran, F., Mansha, S., Karim, A., Zhang, X.: Exploiting reject option in classification for social discrimination control. Inf. Sci. **425**, 18–33 (2018)
10. Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl. Inf. Syst. **35**, 613–644 (2013)
11. Kaufman, L., Rousseeuw, P.: Finding groups in data: An introduction to cluster analysis. John Wiley & Sons (2009)
12. Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: Proc. World Wide Web Conf. pp. 853–862 (2018)
13. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: How we analyzed the COMPAS recidivism algorithm. ProPublica (5 2016) **9**(1), 3–3 (2016)
14. Larson, J., Mattu, S., Kirchner, L., Angwin, J.: ProPublica COMPAS analysis repository (2016), https://github.com/propublica/compas-analysis, accessed: 2023-08-17
15. Lenders, D., Calders, T.: Learning a fair distance function for situation testing. In: Joint European Conf. Machine Learning and Knowledge Discovery in Databases. pp. 631–646. Springer (2021)
16. Luong, B.T., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 502–510 (2011)
17. Moore, S.: Case C-167/97, R v. Secretary of State for Employment, ex parte Nicole Seymour-Smith and Laura Perez. judgment of 9 february 1999, NYR. Common Market Law Rev. **37**(1) (2000)
18. Podani, J.: Extending Gower's general coefficient of similarity to ordinal characters. Taxon **48**(2), 331–340 (1999)
19. Rorive, I.: Proving discrimination cases: The role of situation testing (2009), https://migrant-integration.ec.europa.eu/library-document/proving-discrimination-cases-role-situation-testing$_en, accessed : 2023 - 08 - 17$
20. Ruf, B., Detyniecki, M.: Towards the right kind of fairness in AI. arXiv preprint arXiv:2102.08453 (2021)
21. Tukey, J., Mosteller, F., Hoaglin, D.: Fundamentals of exploratory analysis of variance. Wiley Online Library (1991)

22. Van Bekkum, M., Borgesius, F.: Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? Comput. Law Secur. Rev. **48**, 105770 (2023)
23. Wachter, S., Mittelstadt, B., Russell, C.: Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Comput. Law Secur. Rev. **41**, 105567 (2021)
24. Weiss, N., Weiss, C., Griffey, L.: Introductory statistics. Pearson Education London (2012)