Properties of fairness measures in the context of varying class imbalance and protected group ratios

DARIUSZ BRZEZINSKI, JULIA STACHOWIAK, JERZY STEFANOWSKI, IZABELA SZCZECH, ROBERT SUSMAGA, SOFYA AKSENYUK, ULADZIMIR IVASHKA, and OLEKSANDR YASINSKYI, Institute of Computing Science, Poznan University of Technology, Poznan, Poland

Society is increasingly relying on predictive models in fields like criminal justice, credit risk management, or hiring. To prevent such automated systems from discriminating against people belonging to certain groups, fairness measures have become a crucial component in socially relevant applications of machine learning. However, existing fairness measures have been designed to assess the bias between predictions for protected groups without considering the imbalance in the classes of the target variable. Current research on the potential effect of class imbalance on fairness focuses on practical applications rather than dataset-independent measure properties. In this paper, we study the general properties of fairness measures for changing class and protected group proportions. For this purpose, we analyze the probability mass functions of six of the most popular group fairness measures. We also measure how the probability of achieving perfect fairness changes for varying class imbalance ratios. Moreover, we relate the dataset-independent properties of fairness measures described in this paper to classifier fairness in real-life tasks. Our results show that measures such as Equal Opportunity and Positive Predictive Parity are more sensitive to changes in class imbalance than Accuracy Equality. These findings can help guide researchers and practitioners in choosing the most appropriate fairness measures for their classification problems.

CCS Concepts: • **Computing methodologies** → **Machine learning**; *Artificial intelligence*; **Supervised learning by classification**; • **Social and professional topics** → *Computing / technology policy*.

Additional Key Words and Phrases: group fairness, class imbalance, protected group imbalance

1 INTRODUCTION

Machine learning systems are increasingly being used to make decisions that affect people. Despite the benefits often associated with improving or speeding up tasks, there is also an awareness of the risks associated with such systems. In particular, there is a general agreement that machine learning models need to be controlled to maintain fairness and avoid biased decisions. These issues are reflected in recent AI guidelines, such as the proposed EU regulation on AI [11] and the recently accepted UNESCO recommendation on ethics in AI [42]. The fairness of machine learning models is also a key research driver toward Responsible and Trustworthy AI [45].

Fairness in machine learning refers to the idea that predictive systems should be designed and operated in a way that is fair and just to all individuals and groups [36]. This means that machine learning models should not discriminate against people based on their race, gender, age, or other personal characteristics. Therefore, approaches that mitigate unfairness are based on the notion of *protected attributes* (sometimes also called sensitive attributes) that define *protected and unprotected groups*, i.e., groups that are disproportionately less or more likely

Authors' address: Dariusz Brzezinski; Julia Stachowiak; Jerzy Stefanowski; Izabela Szczech; Robert Susmaga; Sofya Aksenyuk; Uladzimir Ivashka; Oleksandr Yasinskyi, Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60-965, Poznan, Poland, dariusz.brzezinski@cs.put.poznan.pl.

ACM 1556-4681/2024/3-ART
https://doi.org/10.1145/3654659

to be positively classified. Practical machine learning applications with protected attributes include automated hiring procedures [40], credit scoring [34], and criminal justice [4].

In this context, several measures that assess fairness towards protected groups have been put forward [36]. However, current research focuses mainly on proposing methods that improve the fairness of machine learning models [25] or studying correlations between fairness measures in practical applications [1], rather than investigating general dataset-independent properties. Therefore, providing general advice on which fairness measures are best suited for a given case study is hard. Moreover, by focusing on dataset-oriented evaluations, current studies do not answer questions concerning the behavior of fairness measures in the presence of different types of bias in the data (representation bias, skewed distributions, feature bias). In particular, there are no studies connecting theoretical properties of fairness measures with different levels of class imbalance, i.e., situations where at least one of the target classes contains a much smaller number of examples than the other classes [29]. Such a study would be valuable in guiding machine learning practitioners on the use of fairness measures, as many real-world datasets are naturally imbalanced [5]. Finally, there has been no investigation into how interactions between class imbalance and disproportions between protected and unprotected groups might affect the properties of fairness measures.

In this paper, we define general dataset-independent properties of fairness measures that assess their behavior for varying levels of class imbalance and protected group bias. To identify these properties in popular group fairness measures, we analyze their distributions in the context of class and protected group imbalance. As a result, we provide guidelines on which fairness measures are applicable to which types of datasets. The detailed contributions of this paper are as follows:

- In Sections 3 and 4, we recall six popular group fairness measures and put forward a **method for analyzing fairness measures** on the basis of their probability mass functions.
- In Section 5, we propose **a set of general (dataset-independent) fairness measure properties** related to their behavior in the presence of different levels of class imbalance and protected group ratios. We then verify whether the studied six fairness measures possess the proposed properties.
- In Section 6, by training six different classifiers in a controlled experiment **case study using real-world data with varying class and protected group ratios**, we verify whether the identified dataset-independent properties apply to practical classification scenarios.
- In Section 8, we formulate **guidelines on using fairness measures** in different data scenarios and draw lines for future research.

2 RELATED WORK

The concepts discussed in this paper, fall into the field of machine learning fairness [9, 20, 39]. Although the general notion of fairness has been thoroughly studied for several decades, e.g., in social sciences, the awareness of fairness in the machine learning community has developed only in the last few years. Since then, there has been growing interest in the topic, which has led to proposing different fairness notations and constructing many new approaches to prevent classifiers from discriminating against certain groups of people [47].

It is worth noting that there is no single definition of fairness and no single way of verifying it mathematically. For instance, Dablain *et al.* mention as many as 21 notions of fairness [14]. Moreover, these definitions of fairness can be categorized in different ways [8, 20, 25, 26]. For example, according to the categorization by Gajane and Pechenizkiy [26], one can formalize group fairness (considered in this paper), individual fairness, preference of treatment, preference of impact, equality of opportunity, counterfactuals, and unawareness. Among all these options, the literature does not determine which approach to fairness is the best. The two most popular categories are group fairness and individual fairness measures. Group fairness can be quantified by means of differences or ratios. In this study, we follow Žliobaitė [48], who recommends using difference-based rather than ratio-based

definitions. Finally, several papers discuss the orthogonality of various fairness measures [8]. It has been shown that it is impossible to satisfy all fairness measures simultaneously since many of them are mutually exclusive [12].

Most research efforts in machine learning fairness have been directed at proposing methods for detecting [2] and correcting [9] discriminatory bias of machine learning models. Such fairness interventions to models are usually divided into three categories: (1) *pre-processing* (performing specific transformations of data to remove discrimination bias from the training data); (2) *in-processing* (modifying the model's training to adhere to some fairness criteria or constraints); (3) *post-processing* (correcting the biased classifier with respect to the protected attribute or selected measures). For more details on classifier intervention methods, the reader is referred to surveys of Caton and Haas [9] and Dunkelau [20].

Studies on machine learning fairness usually do not explicitly take into account the relative sizes and imbalances of classes and protected groups despite the fact that such characteristics were often present in the data under consideration. On the other hand, it has been observed that the presence of class imbalance may not only lead to poor recognition of minority classes but also to the worsening of fairness measures [14]. Also, Deng *et al.* [15] have reported experiments with imbalanced datasets confirming the potential of class imbalance impacting the fairness of classifier predictions.

Works that tie fairness with the large body of work on class imbalance [23] mainly introduce new specialized methods for improving the selected fairness measure while simultaneously dealing with class imbalances when training the classification models. Kamiran and Calders [33] proposed a pre-processing method, which makes the least intrusive modifications to the proportions of protected and unprotected examples that lead to an unbiased dataset and non-discriminating classifier. Iosifidis and Ntoutsi [31] studied online learning of classifiers in the presence of both biases and proposed to modify boosting ensembles. Their proposal involves adjusting the weights of new training examples according to the monitored changes of the imbalance ratio and adaption of the final decision threshold with respect to one of two fairness measures. Another approach is presented by Ferrari and Bacciu [24], where the authors introduce a new loss function that takes into account both fairness and class imbalance. Similarly, a logit-based loss function that incorporates class imbalance-dependent margins was studied in [15]. In parallel, fair empirical risk minimization provides a framework for optimizing classifiers to conform with a given fairness-oriented loss function [18, 37]. Finally, a resampling-based pre-processing approach, named FairOversampling, has been recently proposed in [14].

Nevertheless, to the best of our knowledge, there is still a lack of systematic research on the impact of different levels of class imbalance and protected group ratios on the properties of selected fairness measures. Aiming to fill this gap, our study builds on:

(A) works defining fairness measures for different applications [8, 9, 20, 48],
(B) studies on class imbalance and fairness [14, 15],
(C) our previous works on analyzing classification measures using confusion matrices [6, 7].

In this paper, we investigate fairness measures defined in (A) in the context of problems discussed in (B) using methods inspired by (C). The novelty of our approach comes from using probability mass functions (C) in a new setting (A+B) for pairs of confusion matrices within subgroups defined by imbalance and group ratios. We deviate from existing analyses on group fairness measures in (B) because we are independent of concrete datasets and classifiers and because we focus on general properties. We believe that the analysis of confusion matrices provides a broader view compared to results based on individual datasets.

Our work asks the question of *which* fairness measure should be optimized and *when*, whereas most existing works focus on *how* to enforce selected measures. Therefore, our work complements existing research on fairness intervention methods, showing the (statistical) suitability of measures for particular types of datasets and classification scenarios. That being said, it is important to acknowledge that fairness is a sensitive topic. Therefore, in real-world applications, several additional aspects (e.g., legal, social, or cultural) and biases should

be very carefully considered when choosing appropriate metrics. Critical discussions of the differences between the technical approach to machine learning fairness and the legal aspects of regulations can be found in recent works by Wachter et al. [43] and Kirat et al. [35].

3 GROUP FAIRNESS MEASURES

To tackle the problem of potentially discriminatory behavior of machine learning models, researchers have put forward many ways of quantifying fairness [8, 9, 20]. Among many different notions of fairness, the most commonly used metrics emphasize either *individual* (similarity-based) fairness or *group* (statistical) fairness. In this paper, we will focus on group fairness, which underlines equal treatment of various groups identified by protected attributes [8]. Without loss of generality, for the sake of simplicity, we will consider problems with one protected categorical attribute that divides examples in a classification dataset into two groups: the *protected group* ($_p$) and the *unprotected group* ($_{up}$). Moreover, we will focus on fairness measures that can be defined for binary classification problems, where the two considered classes will be referred to as the *postive class* and the *negative class*. The numbers of positive and negative examples in a dataset will be denoted as P and N, respectively. The total number of examples in a dataset will be denoted as $n = P + N$.

Following the above, group fairness measures can be defined using entries from a two-class confusion matrix presented in Figure 1. The TP (*True Positive*) and TN (*True Negative*) entries denote the number of examples classified correctly by the classifier as positive and negative, whereas the FN (*False Negative*) and FP (*False Positive*) indicate the number of misclassified positive and negative examples, respectively [32]. In the context of group fairness, this definition of a confusion matrix can be considered as a sum of confusion matrices for the protected and unprotected group, with entries indexed with $_p$ and $_{up}$, respectively (Figure 1). For the purpose of quantifying fairness, when referring to a confusion matrix, we will imply a tuple of eight values: $\{TP_p, FN_p, FP_p, TN_p, TP_{up}, FN_{up}, FP_{up}, TN_{up}\}$.

Entire dataset

Actual \ Predicted	Positive	Negative	total
Positive	TP	FN	P
Negative	FP	TN	N
total	$\widehat{P}$	$\widehat{N}$	n

$=$

Protected group

Actual \ Predicted	Positive	Negative	total
Positive	TP_p	FN_p	P_p
Negative	FP_p	TN_p	N_p
total	$\widehat{P}_p$	$\widehat{N}_p$	n_p

$+$

Unprotected group

Actual \ Predicted	Positive	Negative	total
Positive	TP_{up}	FN_{up}	P_{up}
Negative	FP_{up}	TN_{up}	N_{up}
total	$\widehat{P}_{up}$	$\widehat{N}_{up}$	n_{up}

Fig. 1. Confusion matrix for two-class classification. The diagram shows how, in the context of quantifying fairness, a confusion matrix can be viewed as the sum of two confusion matrices—one for the protected group and one for the unprotected group.

A dataset can have different proportions of examples belonging to positive/negative classes and to protected/unprotected groups. To quantify the class imbalance and group imbalance in a given dataset consisting of n examples, we will use the notions of *imbalance ratio* (IR) and *group ratio* (GR) defined as:

$$IR = \frac{P}{n} \qquad\qquad GR = \frac{n_p}{n}$$

Using the above notation, a dataset with perfect class balance will have $IR = 0.5$. Independently, a dataset with an even number of examples in the protected and unprotected groups will have $GR = 0.5$.

In our work, we focused on fairness measures that can be expressed by the eight entries in the confusion matrices defined above. Barocas et al. denote such measures as *observational* [3]. More formally, we will analyze measures that depend on:

- the protected (sensitive) attribute S that defines the groups for which we want to measure fairness (g_p, g_{up});

- the target attribute Y, which in binary classification represents two classes that we can predict ($Y = 0$ or $Y = 1$);
- the classification score C, which represents the predicted score within $[0, 1]$.

Within observational group fairness measures, we selected six that are traditionally defined as equalities and cover all three categories of non-discrimination criteria: independence ($C \perp S$), separation ($C \perp S|Y$), sufficiency ($Y \perp S|C$) [3, 9]. From the *independence* category, we chose *Accuracy Equality* (Eq. 1) [4] and *Statistical Parity* (Eq. 2) [46].

$$P(C = Y|S = g_p) = P(C = Y|S = g_{up}) \tag{1}$$

$$P(C = 1|S = g_p) = P(C = 1|S = g_{up}) \tag{2}$$

From the *separation* category, we chose *Equal Opportunity* (Eq. 3) [28] and *Predictive Equality* (Eq. 4) [13].

$$P(C = 1|Y = 1, S = g_p) = P(C = 1|Y = 1, S = g_{up}) \tag{3}$$

$$P(C = 1|Y = 0, S = g_p) = P(C = 1|Y = 0, S = g_{up}) \tag{4}$$

Finally, from the *sufficiency* category we chose *Positive Predictive Parity* (Eq. 5) and *Negative Predictive Parity* (Eq. 6) [10].

$$P(Y = 1|C = 1, S = g_p) = P(Y = 1|C = 1, S = g_{up}) \tag{5}$$

$$P(Y = 0|C = 0, S = g_p) = P(Y = 0|Y = 0, S = g_{up}) \tag{6}$$

Even though the notions of group fairness are traditionally discussed in terms of probabilities, for concrete classification scenarios these probabilities are usually estimated using the entries from the confusion matrix. Moreover, in order to quantify fairness, these probability equations are transformed into either ratios or differences. In this study, we have chosen the latter form of quantifying fairness as the values of differences are constrained to a $[-1, 1]$ range, less prone to division by zero problems, and easier to interpret [48]. Moreover, definitions based on differences are more general than definitions based on equations. In fact, the satisfaction of a fairness equation corresponds to a difference being equal to zero; we will refer to this special case as *perfect fairness*.

By redefining the selected measures as differences based on entries of confusion matrices we get:

$$\text{Accuracy Equality} = \frac{TP_p + TN_p}{n_p} - \frac{TP_{up} + TN_{up}}{n_{up}} \tag{7}$$

$$\text{Statistical Parity} = \frac{TP_p + FP_p}{n_p} - \frac{TP_{up} + FP_{up}}{n_{up}} \tag{8}$$

$$\text{Equal Opportunity} = \frac{TP_p}{FN_p + TP_p} - \frac{TP_{up}}{FN_{up} + TP_{up}} \tag{9}$$

$$\text{Predictive Equality} = \frac{FP_p}{FP_p + TN_p} - \frac{FP_{up}}{FP_{up} + TN_{up}} \tag{10}$$

$$\text{Positive Predictive Parity} = \frac{TP_p}{FP_p + TP_p} - \frac{TP_{up}}{FP_{up} + TP_{up}} \tag{11}$$

$$\text{Negative Predictive Parity} = \frac{TN_p}{FN_p + TN_p} - \frac{TN_{up}}{FN_{up} + TN_{up}} \tag{12}$$

4 DISTRIBUTION-BASED ANALYSES OF FAIRNESS MEASURES

In this study, our goal is to examine the behavior of fairness measures in the context of varying class imbalance and protected group ratios. As shown in the previous section, the values of the analyzed fairness measures are derived from confusion matrices, which represent the results of classification on experimental data. By considering the training data as an outcome of a random process, we can provide a probabilistic view of fairness measure values. Specifically, measures based on confusion matrices can be regarded as discrete random variables that map a confusion matrix into a numerical value. Discrete random variables are typically characterized by their *probability mass functions* (*pmfs*), which denote the probability that a discrete random variable precisely equals a specific value [41]. Probability mass functions are frequently represented as histograms, with the x-axis indicating measure values and the y-axis signifying the probability of obtaining a particular value. We will employ such visualizations to scrutinize the fairness measures under consideration.

We will use probability mass functions depicted as histograms to analyze the effects different imbalance ratios and group ratios can have on the six fairness measures (Eq. 7–12). In our analysis, we abstract from concrete classifiers or datasets and, therefore, assume that each possible confusion matrix is equally probable. We are aware that this is a strong assumption, however, in this way, we offer the most inclusive view of what may happen in any classification task. For example, although confusion matrices with mostly incorrect predictions seem less probable than those with mostly correct ones, classifiers tackling data streams will often temporarily make incorrect predictions due to concept drift [7]. Consequently, we will follow an approach similar to that presented in [6, 7] and generate all possible confusion matrices for a dataset size n, and calculate the measure's value for each matrix (Figure 2). Note that, for a dataset size n, the number of all possible confusion matrices defined by $k = 8$ values (TP_p, FN_p, FP_p, TN_p, TP_{up}, FN_{up}, FP_{up}, TN_{up}) equals $c = \binom{n+k-1}{k-1}$. This formula is taken from the 'stars and bars' theorem [22], which shows how to calculate the number of possible k-tuples of non-negative integers summing up to n.

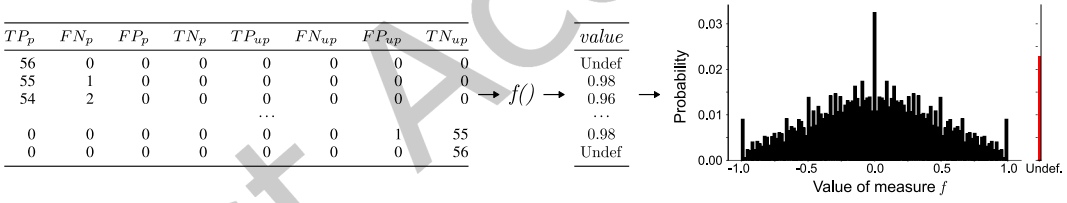

Fig. 2. Process of creating a histogram of a fairness measure probability mass function by generating all possible confusion matrices. Each table row represents an entry of a single confusion matrix (left). These confusion matrices map to fairness measure values (middle). By counting each measure value, we construct a histogram (right). The probability of an undefined measure value (Undef.) is represented as a separate red bar next to the histogram.

Using the calculated measure values, we analyze the *pmf*-based histograms of each measure for varying *IR* and *GR*. In our visualizations, we use $n = 56$ with $IR \in \{\frac{1}{28}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{27}{28}\}$ and $GR \in \{\frac{1}{28}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{27}{28}\}$. We chose $n = 56$ for two reasons: we wanted the value to be divisible by $k = 8$ (the number of analyzed confusion matrix entries), and we needed to be able to compute all the possible $c = \binom{n+k-1}{k-1}$ confusion matrices using 32 GB of RAM. Although $n = 56$ may seem like a small number, it corresponds to c=553,270,671 possible confusion matrices, which provides a respectable sample of classifier outputs.[1] The considered class and group proportions were selected to represent class balance ($\frac{1}{2}$), low imbalance ($\frac{1}{4}$ and $\frac{3}{4}$), and high imbalance ($\frac{1}{28}$ and $\frac{27}{28}$) [44]. Notice that the high imbalance ratios ($\frac{1}{28}$ and $\frac{27}{28}$) correspond to an extreme case where there are only two examples from the minority class,

[1]It took 82 minutes to generate all the possible confusion matrices using an Apple M2 Pro with 32 GB of RAM.

one for each protected/unprotected group.[2] We also note that fairness measure values may be undefined for certain confusion matrices. This is the case when denominators in Eq. 7–12 are equal to zero. The probability of obtaining such undefined values for a given *IR*/*GR* will be visualized as a red bar on the right of the *pmf* histogram (Figure 2).

Figures 3–5 present histograms for Accuracy Equality, Equal Opportunity, and Positive Predictive Parity. Statistical Parity (Supplementary Figure S2) has probability mass functions identical to Accuracy Equality, whereas Predictive Equality and Negative Predictive Parity are top-down mirror images of Equal Opportunity and Positive Predictive Parity, respectively (Supplementary Figures S4 and S6). The code for generating all the possible confusion matrices and their corresponding fairness measure values is available in the repository accompanying this paper.[3]

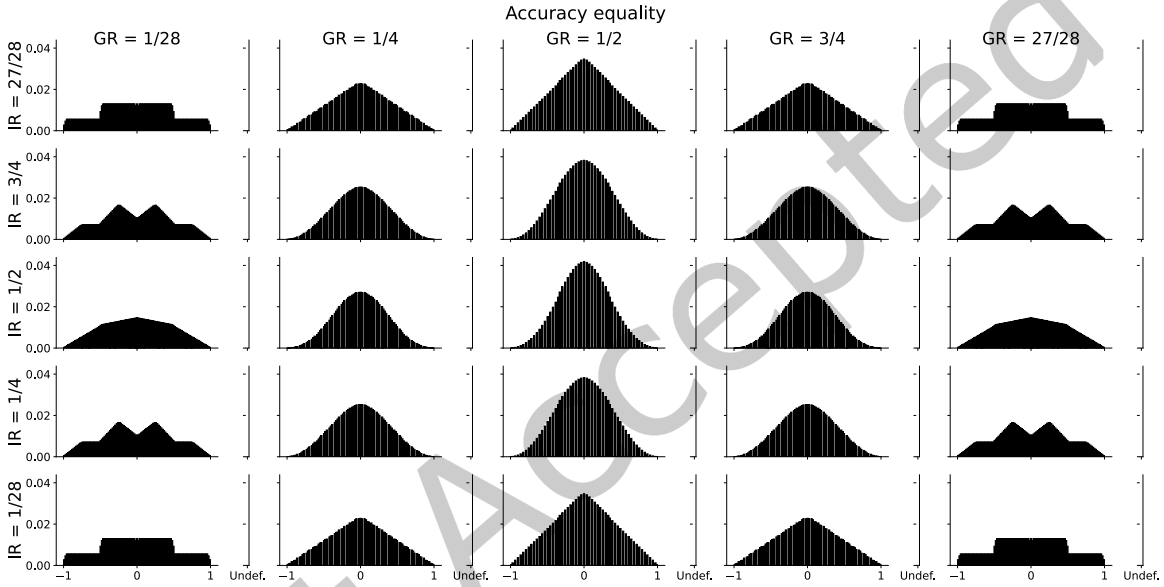

Fig. 3. Histograms of **accuracy equality** values for selected imbalance and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying imbalance ratios *IR* (top-bottom) and group ratios *GR* (left-right). The probability of undefined values (Undef.) is represented as red bars next to each histogram.

[2]This is the smallest number of examples needed to analyze binary class imbalance and fairness between two groups.
[3]Source codes available at: https://github.com/Rasalrai/analysis-of-fairness-measures/.

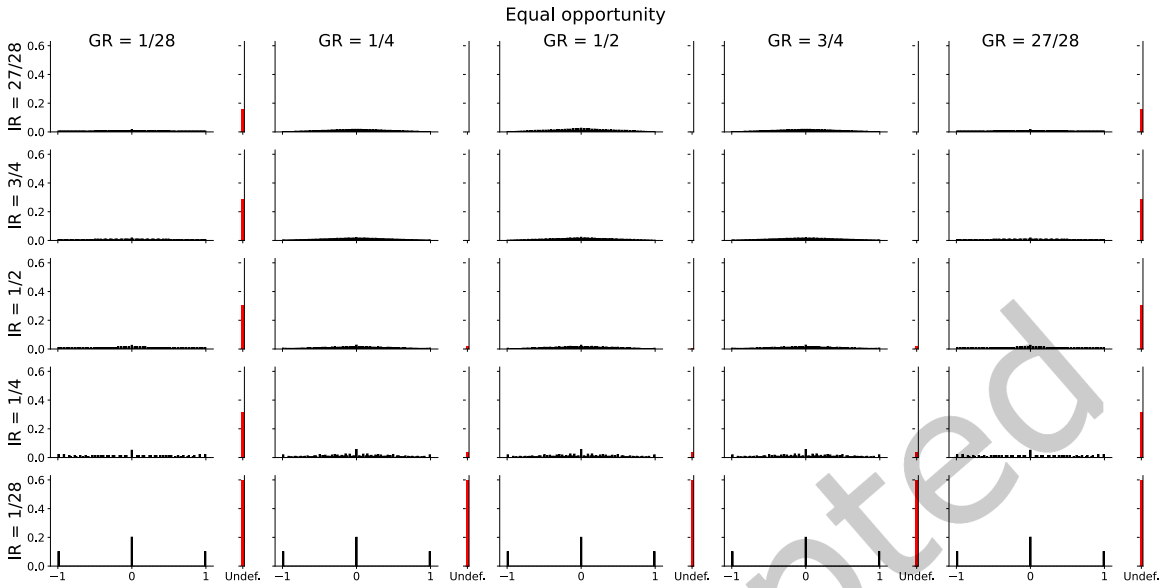

Fig. 4. Histograms of **equal opportunity** values for selected imbalance and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying imbalance ratios *IR* (top-bottom) and group ratios *GR* (left-right). The probability of undefined values (Undef.) is represented as red bars next to each histogram.

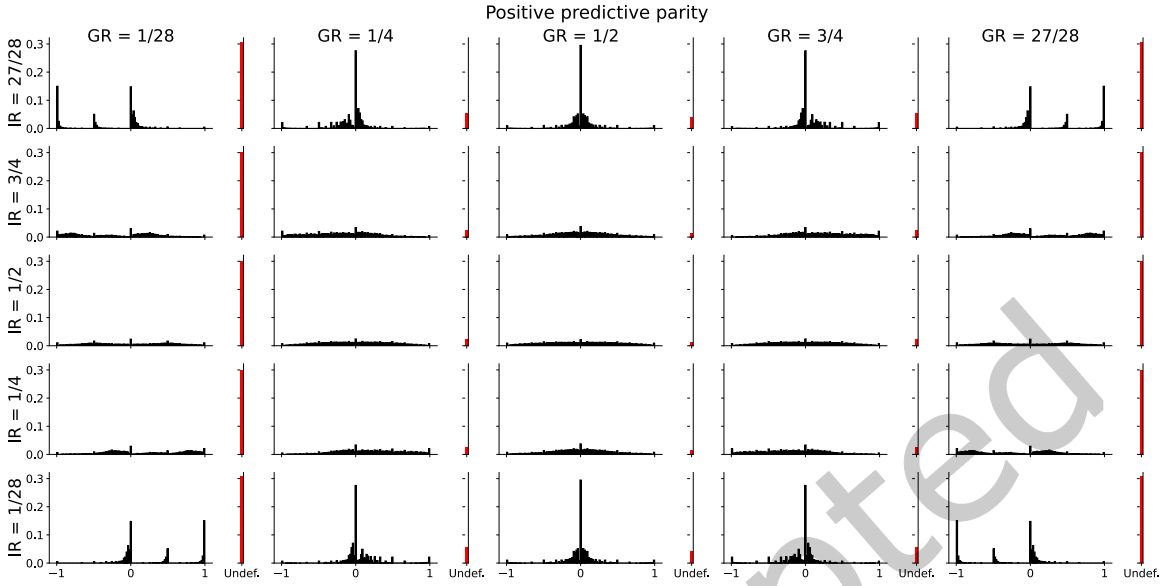

Fig. 5. Histograms of **positive predictive parity** values for selected imbalance and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying imbalance ratios *IR* (top-bottom) and group ratios *GR* (left-right). The probability of undefined values (Undef.) is represented as red bars next to each histogram.

To complement the analysis, we also created line plots depicting the probability of achieving *perfect fairness*, i.e., a fairness measure value equal to 0. Figure 6 shows the fraction of confusion matrices corresponding to perfect fairness (y-axis) for varying imbalance and group ratios (x-axis). As can be noticed, extreme imbalance ratios can make it much more probable to achieve perfect fairness for some measures, whereas group ratio has a negligible effect (zoomed-in versions of the plots can be found in Supplementary Figures S7 and S8). A similar plot was created to show the probabilities of obtaining undefined values of fairness measures (Figures 7, S9, S10). Based on the presented histograms, perfect fairness plots, and undefined value plots, in the following section, we will analyze how the measure's probability mass functions change with varying *IR* and *GR* and propose a set of dataset-independent fairness measure properties.

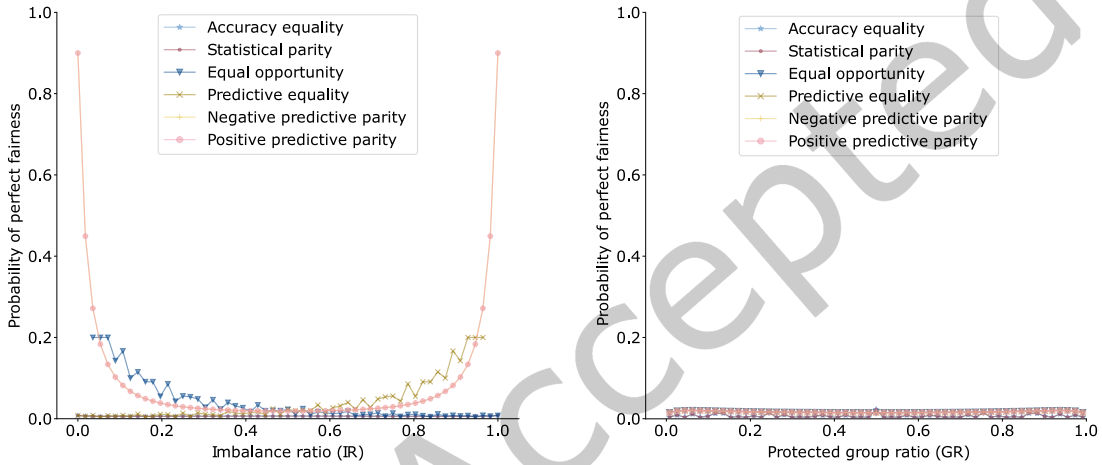

Fig. 6. Probability of achieving perfect fairness using different measures for varying class imbalance ratios (left) and varying group ratios (right).

Finally, the proposed method of analyzing fairness measures can also be applied to classification measures based on confusion matrices. In particular, one can analyze whether classifiers with high accuracy are less likely to be fair, simply due to the fact how fairness is measured. Figure 8 tries to answer this question by showing the relation between different fairness measures (columns) and predictive performance measured by Accuracy $= \frac{TP+TN}{n}$ and G-mean $= \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{FP+TN}}$ (rows). As can be noticed, perfect fairness can be achieved for any value of Accuracy or G-mean. Also, when values of Accuracy depart from 0.5, there is a larger proportion of confusion matrices corresponding to 'close-to-ideal' than 'unfair' confusion matrices, especially in the case of Equal Opportunity, Predictive Equality, Positive Predictive Parity, and Negative Predictive Parity (the two rightmost panels in the first row). Moreover, the asymmetry in G-mean heatmaps (lower vs upper parts of all the panels in the second row) is due to the asymmetry of the G-mean distribution itself, which is much more likely to achieve lower values than higher ones. Additionally, as was seen in Figures 3–5, for particular high imbalance and group ratios the number of possible values of fairness can decrease rapidly. Therefore, although the analyzed definitions of fairness do not discriminate accurate classifiers in general, for datasets with particular example proportions oftentimes there is a fairness-accuracy tradeoff [13].

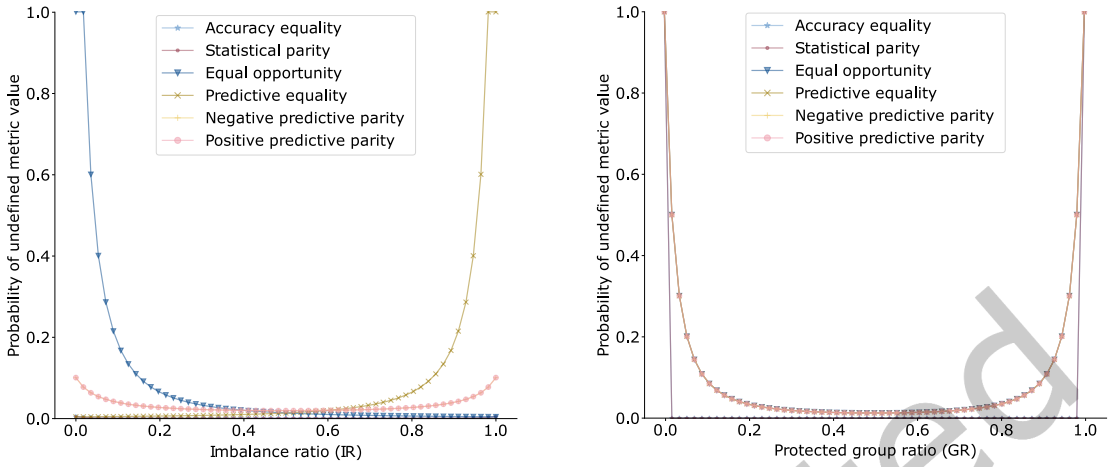

Fig. 7. Probability of achieving an undefined value for different measures under varying class imbalance ratios (left) and varying group ratios (right). Notice the abrupt growth in the number of undefined values for Accuracy Equality and Statistical Parity, and the gradual growth of the remaining four measures for varying group ratios (right).

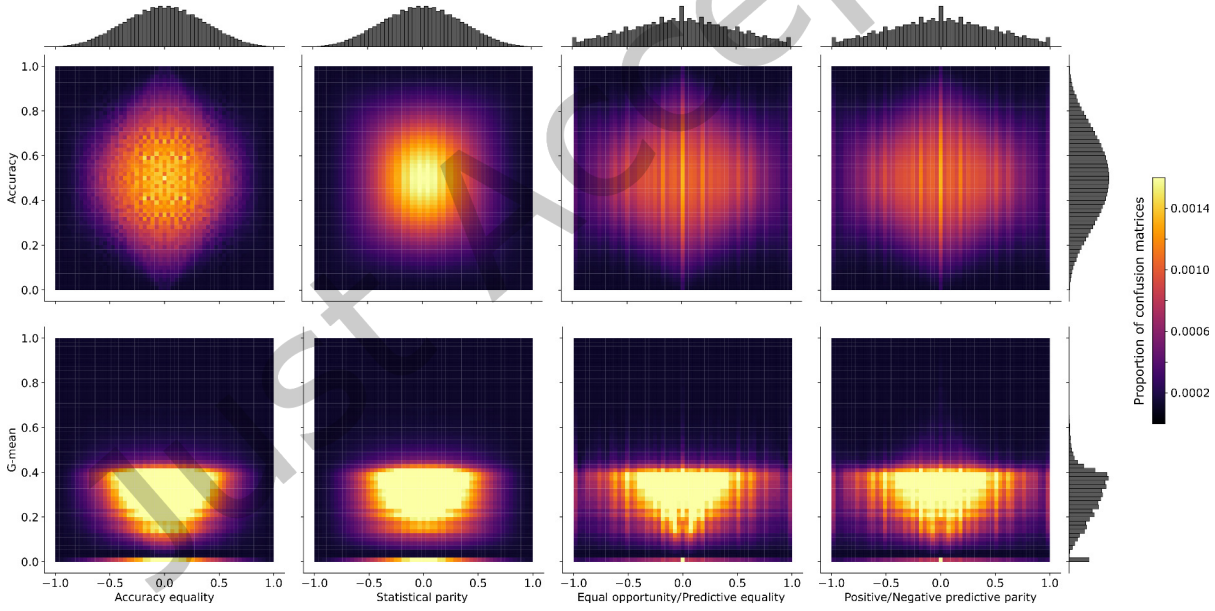

Fig. 8. Heatmaps showing the relation between different fairness measures (columns) and classification accuracy or G-mean (rows). Each heatmap shows the proportion of confusion matrices (color) that correspond to a given value of fairness (x-axis) and predictive performance (y-axis). Fairness measures with identical heatmaps have been grouped together.

5 PROPERTIES OF FAIRNESS MEASURES

With a method for analyzing fairness measures at hand, in this section, we will define and interpret the potentially desirable properties of these measures. The properties should aid researchers in comparing various measures, examining differences in their distributions, noticing unusual or unexpected values, and selecting measures suitable for a given classification task. Defining such a set of properties in the context of class imbalance and protected group bias is vital for several reasons. First, it helps understand the performance and limitations of these measures under varying dataset characteristics. As real-world datasets often have imbalanced classes and bias toward certain groups, understanding how these factors impact the fairness measures can guide their use. Second, defining these properties provides a theoretical framework that can help design novel fairness measures. Finally, it supports the development of more robust and fair machine learning models, as the proposed properties may help in the development of new fairness intervention methods for classifiers.

In this context, we postulate to analyze group fairness measures with respect to eight properties:

Immunity to IR changes: the distribution of the measure's values does not change for varying imbalance ratios. Measures with this property should perform the same way for datasets with different imbalance ratios.

Immunity to GR changes: the distribution of the measure's values does not change for varying protected group ratios. Measures with this property should perform the same way for datasets with different group ratios.

Resolution Stability: the number of unique measure values is large regardless of the imbalance and group ratios. Measures with this property will always be able to provide a wide range of different values, rather than degrading to signaling only a few values, e.g., perfect fairness (0) and perfect unfairness (1/-1).

Fairness Symmetry: all the distributions of the measure values for varying *IR* and *GR* levels should be symmetrical around zero (i.e., around perfect fairness). Measures with this property will not promote one (protected/unprotected) group over the other for any class or group ratio.

IR Symmetry: the measure's distribution is the same for counterpart ratios of positive and negative class examples (same distribution, e.g., for $IR = 0.01$ and $IR = 0.99$). Measures with this property focus on both classes and will behave the same way for a given proportion of classes, regardless of whether the positive or negative class is underrepresented.

GR Symmetry: the measure's distribution is the same for counterpart ratios of protected and unprotected group examples (same distribution, e.g., for $GR = 0.01$ and $GR = 0.99$). Measures with this property focus on both protected and unprotected groups and will behave the same way for a given proportion of groups, regardless of whether the protected or unprotected group is underrepresented.

Perfect Fairness Stability: the probability of achieving perfect fairness stays almost constant for different imbalance and group ratios. Measures with this property will make the task of achieving perfect fairness (e.g., by classifier interventions) comparable for different class and group proportions.

Undefined Values: the existence of undefined values. Measures with fewer undefined values will make quantifying fairness less prone to numerical problems.

Having presented the visualization technique in Section 4 and having defined the properties above, we will now use the proposed tools to analyze the fairness measures. Table 1 summarizes the results of the verification of the properties for each of the examined measures. Below we compare these outcomes, providing our observations.

Let us start by noticing that the six analyzed measures form three groups: i) Accuracy Equality and Statistical Parity, ii) Equal Opportunity and Predictive Equality, and iii) Positive Predictive Parity and Negative Predictive Parity. Indeed, the measures within the pairs have similar definitions (see Eq. 7–12), e.g., the TN_p and TN_{up} in Accuracy Equality are simply substituted by FP_p and FP_{up} in Statistical Parity. Additionally, these measure pairs either have the same value distributions (see, e.g., Figure 3 and Figure S2 in the supplementary materials) or

Table 1. Properties of selected fairness measures; †: extreme group ratios introduce undefined values, but the shape of the distribution does not change.

	Accuracy Equality	Statistical Parity	Equal Opportunity	Predictive Equality	Positive Predictive Parity	Negative Predictive Parity
Immunity to IR changes	×	×	×	×	×	×
Immunity to GR changes	×	×	✓†	✓†	×	×
Resolution Stability	✓	✓	×	×	×	×
Fairness Symmetry	✓	✓	✓	✓	×	×
IR Symmetry	✓	✓	×	×	×	×
GR Symmetry	✓	✓	✓	✓	×	×
Perfect Fairness Stability	✓	✓	×	×	×	×
Undefined Values	when $n_p = 0$ or $n_{up} = 0$	when $n_p = 0$ or $n_{up} = 0$	low/high GR, low IR	low/high GR, high IR	low/high GR	low/high GR

their distributions are symmetrical, e.g., Equal Opportunity focuses on the positive class, whereas Predictive Equality focuses on the negative class (see Figure 4 and Figure S4 in the supplementary materials). The same can be noticed on Figures 6 and 7. As a result, the columns in Table 1 representing the measures from each pair have identical entries.

Now, let us consider the first three properties. It can be noticed that each measure is susceptible to changes in class and group ratios. For varying imbalance ratios (examine the panels in Figures 3–5 from top to bottom), each measure changes its distribution, with Equal Opportunity/Predictive Equality having very few unique values (low resolution) for low/high *IR*, and Positive and Negative Predictive Parity also having fewer unique values and much bigger chances of perfect fairness for extreme imbalance ratios (see also Figure 6). Although Accuracy Equality and Statistical Parity also change their probability mass functions with class imbalance, the differences are much less drastic and do not affect the number of unique measure values. Similarly, for varying group ratios (examine the panels in Figures 3–5 from left to right), the measures change their distributions, but these changes do not influence the chances of achieving perfect fairness that much, with Equal Opportunity and Predictive Equality practically having the same distributions, only with a higher chance of obtaining an undefined value (Figures 6 and 7).

The next three properties ascertain the symmetry of measure behavior. Practically all measure distributions (Figures 3-5) are always centered around zero, i.e. perfect fairness. Only Positive and Negative Predictive Parity become asymmetrical for imbalanced group proportions. Moreover, Accuracy Equality and Statistical Parity additionally have the same distributions for counterpart group and class proportions, making these two measures the most symmetrical in terms of treating positive/negative and protected/unprotected examples. On the other hand, Equal Opportunity and Predictive Equality are only symmetrical in treating protected/unprotected groups, whereas Positive and Negative Predictive Parity favor protected or unprotected groups depending on *IR* and *GR*.

Taking a closer look at the chances of achieving perfect fairness (Figure 6), we see that only Accuracy Equality and Statistical Parity offer consistency. The remaining measures, on the other hand, have substantially higher theoretical chances of signaling perfect fairness depending on the imbalance ratio and group ratio. In particular, it is interesting to note that for Equal Opportunity, there is a bigger chance of achieving perfect fairness when

there are few positive examples in the dataset. In contrast, Predictive Equality makes it easier to achieve perfect fairness when there are few negative examples.

Finally, the chances of obtaining an undefined value differ between measures (Figure 7). We see that only Accuracy Equality and Statistical Parity are almost free of undefined values—they only occur when one of the groups is missing ($n_p = 0$ or $n_{up} = 0$). The remaining measures, on the other hand, can have high probabilities of undefined values depending on the imbalance ratio and group ratio. In particular, for very high and very low values of *GR*, undefined values are the dominant values for Equal Opportunity, Predictive Equality, Positive Predictive Parity, and Negative Predictive Parity (Figures 4 and 5). Additionally, *IR* has a strong effect on the number of undefined values for Equal Opportunity and Predictive Equality. Although undefined values can be considered merely a numerical problem, not being able to measure fairness can impact the ability to act upon unfair predictions.

In the following section, we will verify whether these dataset-independent properties can affect the measured fairness of machine learning models in a practical classification scenario.

6 CASE STUDY ON THE EFFECT OF MEASURE PROPERTIES ON CLASSIFIER FAIRNESS

6.1 Experimental setup

To verify how the proposed properties apply to practical problems involving real classifiers, we performed an experiment using the UCI Adult dataset [19]. We chose UCI Adult as it is one of the most popular binary datasets in the context of both imbalanced learning and model fairness [21]. The prediction task is binary classification: the positive class '> 50*K*' (also referred to as the *rich*) indicates people with over 50,000 USD of yearly income, and the negative class '<= 50*K*' (*poor*) denotes a yearly income not higher than 50,000 USD. We selected *sex* as the protected attribute, with *Female* considered as the protected group and *Male* as the unprotected group.

Using the UCI Adult dataset, our goal was to see whether the fairness of classifiers (as quantified by different measures) will be affected by different levels of protected group ratios (*GR*) and imbalance ratios (*IR*). Therefore, we decided to sample subsets from this dataset with controlled values of *IR* and *GR*. Each subset had a specified value for one of the ratios (the selected ratios {0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.98, 0.99}), while the other ratio was set to 0.5. All the sampled subsets of the dataset were equally sized to $n = 1100$ in order to avoid the impact of differently-sized data on the results. Special care was taken to ensure that the proportions of groups were in accordance with the given *GR* within the entire dataset as well as within each class. For instance, for a data subset with $IR = 0.1$ and $GR = 0.5$, there were 55 poor women, 55 rich women, 495 poor men, and 495 rich men in this subset. All these conditions were set to make the classification problems for each data subset as similar as possible, with only *IR* or only *GR* changing.

Using the prepared data subsets with given *IR* and *GR*, for each subset, we performed 50 repetitions of randomly stratified holdout evaluations (67% train, 33% test). In each evaluation, we assessed the fairness of six types of popular learning algorithms chosen for their diversity: k-Nearest Neighbors (k-NN), Naive Bayes, Decision Tree, Logistic Regression, Random Forest, a Multilayer Perceptron with a hidden layer of 100 neurons (MLP). Since our goal is only to illustrate the effect of varying imbalance and group ratios, we left the classifiers with default parameters in their Python implementation in the scikit-learn library [38]. The experiments were conducted on a machine equipped with Intel© Core™ i7-1260P 4.7GHz processor and 48 GB of RAM, using Python 3.10.10 and scikit-learn version 1.2.2. Reproducible scripts for data preparation and all experiments are available at: https://github.com/Rasalrai/analysis-of-fairness-measures/.

6.2 Results

Figure 9 presents the means and standard deviations (y-axis) of the analyzed fairness measures for varying imbalance ratios *IR* (x-axis). An analogous plot for varying group ratios *GR* can be found in Supplementary Figure S8.

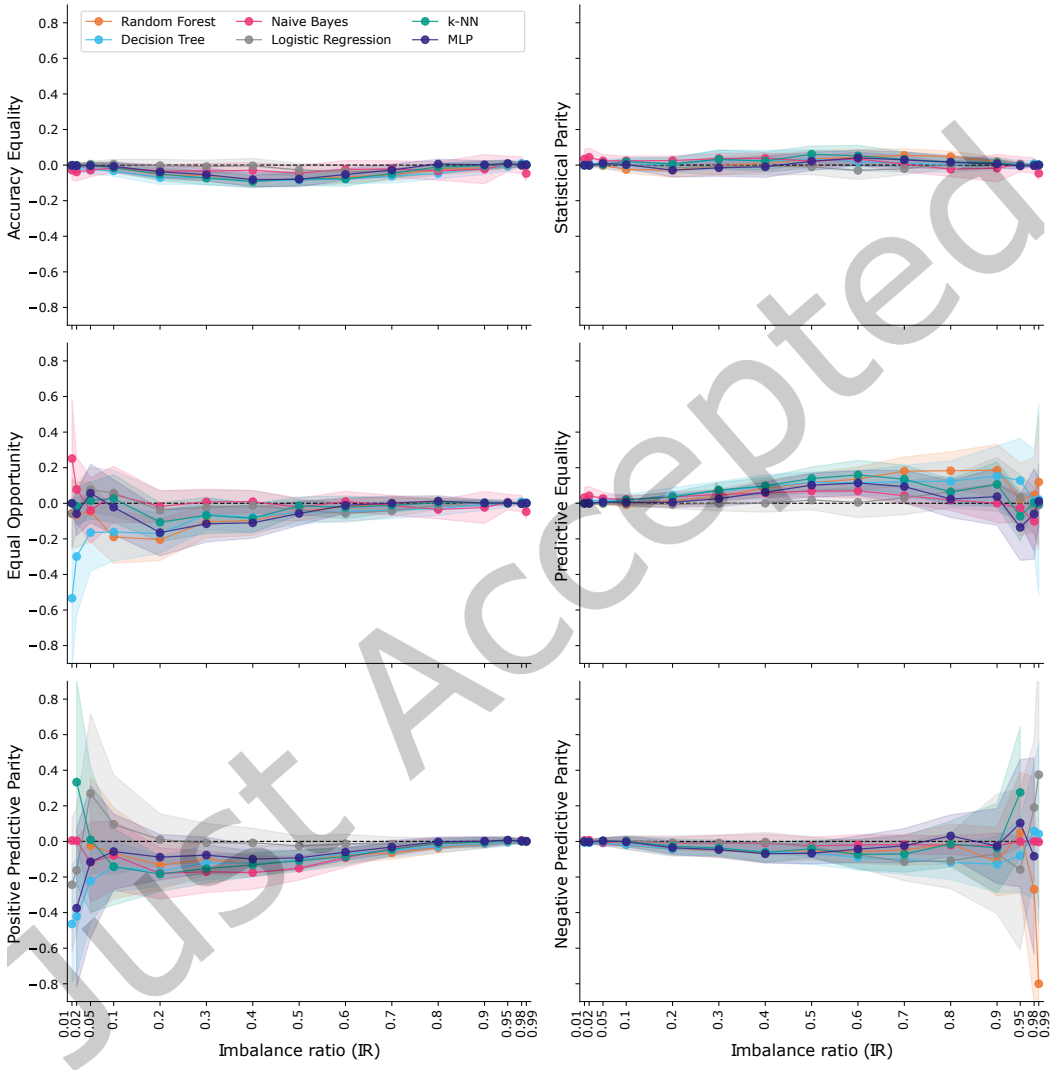

Fig. 9. Fairness values achieved by the analyzed classifiers for different fairness measures, as a function of imbalance ratio *IR* with *GR* set to 0.5. Points represent mean classifier fairness for the 50 repetitions for the given *IR* and the filled area shows the standard deviation.

As the results show, the fairness of the classifiers depends on the measure used and the imbalance ratio in the dataset. The most stable fairness is achieved according to Accuracy Equality and Statistical Parity. These

results are in line with our histogram-based analysis and the properties assigned to these two measures (Table 1). Moreover, one can notice that fairness assessments according to Equal Opportunity and Predictive Equality become unstable for imbalance ratios corresponding to very few unique values (Resolution Stability property). Therefore, Equal Opportunity may be less reliable for low values of *IR* (few positives), whereas Predictive Equality might not be the best choice for datasets with high values of *IR* (few negatives). Finally, Positive and Negative Predictive Parity are even more unstable than the previous two measures. This is the combined effect of very few unique values (Resolution Stability) and the measures relying on predicted positives (*FP* and *TP*) or predicted negatives (*FN* and *TN*) rather than actual positives and negatives as was the case for Equal Opportunity and Predictive Equality. Therefore, the measures are more prone to biased classifier predictions. For varying *GR*, all the measures performed very similarly (Supplementary Figure S12). Indeed, as it was mentioned in Section 5, the analyzed measures are much less prone to changes in protected group ratios.

7 DISCUSSION

The presented study shows that data imbalance can have a non-negligible effect on the behavior of group fairness measures. In particular, positive predictive parity and negative predictive parity behave asymmetrically depending on the imbalance ratio, which makes it easier to achieve high fairness using these measures simply due to class imbalance. Predictive equality and equal opportunity showcase similar problems, albeit to a lesser extent. This highlights the need for careful selection of fairness measures depending on the analyzed data and the potential need for new measures that are more immune to class imbalance.

Apart from restating the main findings of this study, it is worth listing its limitations. First, we note that the presented analysis focused on group fairness measures defined on the basis of entries of confusion matrices. Several other measures quantify algorithmic fairness through a different perspective, e.g., individual fairness [9]. To study the effect of class imbalance on those measures, methods different than the one proposed in this paper are required. Similarly, we focused on problems with two classes and two groups. The analysis of multiple classes and multiple groups using our approach would be possible but would require analyzing the data using a one-vs-all strategy. Also, even though the number of simulated confusion matrices is very large (c=553,270,671), the fact that the sum of the entries of these confusion matrices is $n = 56$ may lead to an exaggeration of undefined and perfect fairness situations compared to datasets with much larger n. That being said, all the properties in Table 1 hold true regardless of the dataset size. Finally, the experimental case study uses only one popular fairness benchmark, which has its limitations [16]. Therefore, more real-world experiments, in particular in the domains of image and text analysis, are still required to assess the practical impact of the characteristics of fairness measures discussed in this paper. Nevertheless, the properties defined in Section 5 are independent of the dataset and classifier type and can be supportive for practitioners regardless of their problem setting.

Our claim that the discussed properties of fairness measures are independent of any dataset and classifier relies on the fact that we exhaustively analyze all possible confusion matrices. However, one may argue that not all confusion matrices are equally probable. Indeed, classifiers that perform worse than a majority stub are not likely to be used in practice [30]. Nevertheless, we decided to analyze all possible confusion matrices because this approach relies on fewer assumptions. What constitutes a 'good enough' classifier differs from application to application. Especially in the domain of class-imbalanced learning, practitioners are willing to trade overall accuracy in favor of correct detection of the minority class, often expressed by means of specialized measures of predictive performance [27, 29]. Moreover, as mentioned earlier, classifiers tackling concept-drifting data streams will have periods of incorrect predictions after sudden changes in the data generating distribution [7]. For these reasons, we decided to treat each confusion matrix as equally probable; nevertheless, future studies may focus on more specific cases with more assumptions about the data distribution and classifier performance.

8 CONCLUSIONS

In this paper, we have analyzed the behavior of six popular group fairness measures in the context of varying class imbalance and protected group bias. For this purpose, we have defined eight dataset-independent properties that helped us characterize the studied measures using their probability mass functions. We further verified the proposed general properties through a controlled experiment using real-world data and six different classifiers.

Our results show that all the analyzed measures change their behavior in the presence of class imbalance and, to a lesser extent, in the presence of protected group bias. In particular, we have shown that measures that take into account the entire confusion matrix, such as Accuracy Equality and Statistical Parity, have the most stable value distributions under varying class and group proportions, treat both classes and groups symmetrically, have hardly any undefined values, and the chance of achieving perfect fairness stays close to constant for all imbalance and group ratios. Therefore, these measures can be considered the most reliable for imbalanced datasets. We have also highlighted that Equal Opportunity and Predictive Equality are complementary, as the first one becomes less stable with few positive examples in a dataset, and the latter performs worse with few negative examples in a dataset. Therefore, depending on the type of class imbalance, one will work better than the other. Finally, Positive and Negative Predictive Parity were found to be the least stable and most asymmetric in their distributions. That is why these two measures should be used mainly for datasets with relatively balanced classes and protected groups.

The findings of this study can be directly used to improve the fairness of machine learning models in two ways. First, our study highlights the need to select fairness measures while taking into account data characteristics. Secondly, we show which measures are more suitable for which types of datasets, guiding fairness measure selection in practical settings. Moreover, our study opens several avenues for further research. Future work could extend the current analysis to properties of fairness measures under more complex data scenarios. In particular, we would be interested in investigating other types of uneven data distributions, such as stereotypical bias. The term stereotypical bias refers to the situation when protected groups are underrepresented within certain classes, even though they are evenly distributed in the entire dataset. Recent works have shown that, indeed, stereotypical bias occurs in popular computer vision datasets [17]. Moreover, one potential limitation of this study is the assumption that all possible confusion matrices are equally probable. Further studies could implement the probability of achieving particular confusion matrices as a parameter that could be used to analyze the properties of measures under different priors. Additionally, the properties put forward in this paper could be used to design novel fairness measures that are more robust and better suited to handle varying class imbalance and protected group bias. Finally, these properties may aid in the development of new fairness intervention methods for classifiers, ultimately contributing to the creation of more responsible AI systems.

ACKNOWLEDGMENTS

This research was partly funded by the National Science Centre, Poland, grant number 2022/47/D/ST6/01770.

REFERENCES

[1] Hadis Anahideh, Nazanin Nezami, and Abolfazl Asudeh. 2021. On the choice of fairness: Finding representative fairness metrics for a given context. *arXiv:2109.05697* (2021), 1–25.

[2] Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. 2021. Dalex: responsible machine learning with interactive explainability and fairness in python. *The Journal of Machine Learning Research* 22, 1 (2021), 9759–9765.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities.* MIT Press.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[5] Paula Branco, Luís Torgo, and Rita P Ribeiro. 2016. A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)* 49, 2 (2016), 1–50.

[6] Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczech. 2018. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Information Sciences* 462 (2018), 242–261.
[7] Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczech. 2019. On the dynamics of classification measures for imbalanced and streaming data. *IEEE transactions on neural networks and learning systems* 31, 8 (2019), 2868–2878.
[8] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 4209.
[9] Simon Caton and Christian Haas. 2023. Fairness in machine learning: A survey. *Comput. Surveys* (2023), 1–33. https://doi.org/10.1145/3616865
[10] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[11] European Commission. 2021. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act).* Technical Report. European Union. Procedure number 2021/0106/COD.
[12] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023* (2018), 1–117.
[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 797–806.
[14] Damien Dablain, Bartosz Krawczyk, and Nitesh Chawla. 2022. Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *arXiv:2207.06084* (2022), 1–27.
[15] Zhun Deng, Jiayao Zhang, Linjun Zhang, Ting Ye, Yates Coley, Weijie J. Su, and James Zou. 2022. FIFA: Making Fairness More Generalizable in Classifiers Trained on Imbalanced Data. *arXiv:2206.02792* (2022), 1–23.
[16] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490.
[17] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. 2023. Gender Stereotyping Impact in Facial Expression Recognition. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*. Springer, Cham, 9–22.
[18] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical Risk Minimization Under Fairness Constraints. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., Montreal, Canada, 1–11.
[19] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[20] Jannik Dunkelau and Michael Leuschel. 2019. *Fairness-aware machine learning: An extensive overview.* Technical Report. Universität Düsseldorf. 1–60 pages.
[21] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic Fairness Datasets: The Story so Far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074—-2152.
[22] William Feller. 1968. *An introduction to probability theory and its applications. Vol. I.* John Wiley & Sons Inc., New York.
[23] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from imbalanced data sets*. Vol. 10. Springer, Switzerland.
[24] Elisa Ferrari and Davide Bacciu. 2021. Addressing Fairness, Bias and Class Imbalance in Machine Learning: the FBI-loss. *arXiv:2105.06345* (2021), 1–23.
[25] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. Cambridge University Press, Cambridge, UK, 329–338.
[26] Pratik Gajane and Mykola Pechenizkiy. 2018. On Formalizing Fairness in Prediction with Machine Learning. *arXiv 1710.03184* (2018), 1–6.
[27] Qiong Gu, Li Zhu, and Zhihua Cai. 2009. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*. Springer, 461–471.
[28] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).
[29] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.
[30] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A Model Behaviour Mutation Approach to Benchmarking Bias Mitigation Methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, New York, NY, USA, 994–1006.

[31] Vasileios Iosifidis and Eirini Ntoutsi. 2020. FABBOO - Online Fairness-Aware Learning Under Class Imbalance. In *IFIP Working Conference on Database Semantics*. Springer International Publishing, Cham, 159–174.
[32] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, Cambridge, UK.
[33] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication* (Karachi, Pakistan). IEEE, 1–6.
[34] Amir E Khandani, Adlar J Kim, and Andrew W Lo. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34, 11 (2010), 2767–2787.
[35] Thierry Kirat, Olivia Tambou, Virginie Do, and Alexis Tsoukiàs. 2022. Fairness and Explainability in Automatic Decision-Making Systems. A challenge for computer science and law. *arXiv preprint arXiv:2206.03226* (2022).
[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (2021), 1–35.
[37] Luca Oneto, Michele Donini, and Massimiliano Pontil. 2020. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, NY, USA, 1–8.
[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[39] Andrea Romei and Salvatore Ruggieri. 2013. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29 (2013), 582–638.
[40] Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. 2020. We need fairness and explainability in algorithmic hiring. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1716–1720.
[41] T. T. Soong. 2004. *Fundamentals of Probability and Statistics for Engineers*. Wiley, Chichsters, UK.
[42] UNESCO. 2021. *Draft Recommendation on the Ethics of Artificial Intelligence*. Technical Report. UNESCO. Document code SHS/BIO/REC-AIETHICS/2021.
[43] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2020. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.* 123 (2020), 735.
[44] Gary Weiss. 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, Hoboken, NJ, USA, Chapter Foundations of Imbalanced Learning, 13–43.
[45] Jeannette M. Wing. 2021. Trustworthy AI. *Commun. ACM* 64, 10 (2021), 64—-71.
[46] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[47] Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, Jie Chen, and Agus Sudjianto. 2022. Bias, Fairness, and Accountability with AI and ML Algorithms. *International Statistical Review* 90 (2022), 468–480.
[48] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31 (2017), 1060–1089.