Stereotype ratio as an evaluation tool for fairness measures

KIRIL ANDRUKH, ULADZISLAŬ LUKASHEVICH, XIMA RAN, and DARIUSZ BRZEZIŃSKI, Institute of Computing Science, Poznan University of Technology, Poland

The increasing reliance on predictive models in societal sectors such as criminal justice, credit risk management, and hiring has attracted a lot of attention on fairness measures. These measures play a crucial role in preventing automated systems from exhibiting discriminatory behavior towards specific groups. We expanded on existing introducing negative and positive stereotype proportions and studying general properties of fairness measures for changing negative and positive stereotype proportions. In addition we measure the probability of achieving perfect fairness changes for stereotypical ratio. We also found the correlations between negative stereotype proportion, positive stereotype proportion, stereotype proportion to group proportion and positive class proportion. Our results show that measures get more undefined values when the stereotype ratio is high and the imbalance is not major; measures highest possible values decrease when both stereotype ratio and imbalance increase. The results of this study can assist scientists and professionals in selecting the most suitable fairness metrics for their classification tasks.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

1 INTRODUCTION

In modern society, widespread use of machine learning algorithms for decision-making processes has prompted a consensus on the imperative of ensuring fairness in these models. This concern is reflected in recent guidelines, such as the proposed EU regulation on AI [6] and the UNESCO recommendation on AI ethics, underscoring the significance of fair and ethical practices in artificial intelligence [15]. The pursuit of fairness in machine learning models serves as a foundational aspect of advancing Responsible and Trustworthy AI [17].

The concept of *fairness* in machine learning embodies the principle that predictive systems should be carefully designed and operated to ensure fairness and equality for all individuals and groups [12, 13]. This implies a commitment to non-discrimination on personal grounds, including but not limited to gender, religion, nationality, or race. To methodically assess the fairness of a machine learning algorithm, the introduction of **protected attributes** that distinguish between *protected* and *unprotected* groups is key. These groups are characterized by a disproportionately lower or higher probability of a positive outcome.

In this complicated context, various ways to check fairness have been introduced. These methods particularly look at how machine learning programs treat certain groups, adding to the ongoing discussion about using artificial intelligence in an ethical way.

In this paper we analyze how a ratio

The detailed contributions of this paper are as follows:

- In Section 3 and 4 we put forward a new ratio and a new method to analyze fairness measures on the basis of their probability mass functions.
- In Section 7 we formulate guidelines on using fairness measures in different data scenarios and draw lines for future research.

Authors' address: Kiril Andrukh; Uladzislaŭ Lukashevich; Xima Ran; Dariusz Brzeziński, Institute of Computing Science, Poznan University of Technology, ul. Piotrowo 2, 60-965, Poznan, Poland.

2 RELATED WORK

In recent years, the machine learning community has gained awareness of fairness issues, especially in predictive systems affecting people [9, 14]. Despite decades of research on fairness in social sciences, the machine learning perspective has introduced new notations and approaches to prevent discriminatory biases in classifiers, leading to increased interest in the subject [18].

Fairness in machine learning lacks a single, universally accepted definition and a consistent mathematical verification method. Dablain *et al.* discuss as many as 21 fairness notions [7], categorized in different ways [9**?** , 10]. Notably, Gajane and Pechenizkiy [10] categorize fairness into various types, including group fairness (examined here), individual fairness, preference of treatment, preference of impact, equality of opportunity, counterfactuals, and unawareness.

The literature does not definitively endorse a single approach to fairness, but group and individual fairness measures are the most popular. In the machine learning fairness domain, considerable efforts focus on detecting [**?**] and correcting [5] discriminatory bias in models. Interventions are typically categorized into pre-processing, in-processing, and post-processing methods [5, 9].

Brzeziński *et al.* address the gap in systematically studying the influence of class imbalance and protected group ratios on fairness measures [1].

In summary, while fairness measures play a crucial role in machine learning, the diverse definitions and interventions, the often-overlooked impact of class imbalance, and the growing interest in the subject necessitate a comprehensive exploration of these aspects for a nuanced understanding of fairness in different scenarios.

3 STEREOTYPE FAIRNESS MEASURES

Addressing the potential issue of biased behavior in machine learning models, various methods for measuring fairness have been proposed by researchers [4, 5, 9]. These metrics often revolve around two main concepts: *individual* (similarity-based) fairness and *group* (statistical) fairness. This paper focuses specifically on group fairness, emphasizing equal treatment across distinct groups identified by protected attributes [4]. For simplicity, we consider problems with one protected categorical attribute dividing a classification dataset into two groups: the *protected group* (p) and the *unprotected group* (up). Moreover, our attention centers on fairness measures applicable to binary classification problems, with the two classes referred to as the *positive class* and the *negative class*. The number of positive and negative examples in a dataset is denoted as P and N, respectively, with the total number of examples denoted as $n = P + N$.

In line with the above, group fairness measures are defined using entries from a two-class confusion matrix illustrated in Figure 1. The *TP* (*True Positive*) and *TN* (*True Negative*) entries represent the number of correctly classified positive and negative examples by the classifier. Conversely, the *FN* (*False Negative*) and *FP* (*False Positive*) entries indicate the misclassified positive and negative examples [11]. In the context of group fairness, this confusion matrix definition can be seen as a summation of matrices for the protected and unprotected groups, with entries indexed as p and up, respectively (Figure 1). When discussing a confusion matrix for fairness quantification, we refer to a tuple of eight values: TPp, FNp, FPp, TNp, $TPup$, $FNup$, $FPup$, $TNup$.

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.

In a dataset, the distribution of examples across positive/negative classes and protected/unprotected groups may vary. To assess these imbalances in a dataset with n examples, we introduce the concepts of *imbalance ratio* (IR) and *group ratio* (GR), defined as:

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

In this notation, a dataset with perfect class balance achieves $IR = 0.5$. Similarly, a dataset evenly split between protected and unprotected groups attains $GR = 0.5$.

To quantify the correlation of what is the fraction of positive examples to the number of all positive examples in a dataset the term *stereotype ratio positive*(SR_p) was introduced. To quantify the correlation of what is the fraction of negative examples to the number of all negative examples in a dataset the term *stereotype ratio negative*(SR_n) was introduced.

$$SR_p = \frac{P_p}{P} \qquad\qquad SR_n = \frac{N_p}{N}$$

Stereotypical ratio is a geometric mean of SR_p and SR_n.

$$SR = \sqrt{SR_p * SR_n}$$

We will examine six widely recognized metrics for evaluating group fairness [5, 12], defined as:

$$\text{Accuracy Equality} = \frac{TP_p + TN_p}{n_p} - \frac{TP_{up} + TN_{up}}{n_{up}} \tag{1}$$

$$\text{Statistical Parity} = \frac{TP_p + FP_p}{n_p} - \frac{TP_{up} + FP_{up}}{n_{up}} \tag{2}$$

$$\text{Equal Opportunity} = \frac{TP_p}{FN_p + TP_p} - \frac{TP_{up}}{FN_{up} + TP_{up}} \tag{3}$$

$$\text{Predictive Equality} = \frac{FP_p}{FP_p + TN_p} - \frac{FP_{up}}{FP_{up} + TN_{up}} \tag{4}$$

$$\text{Positive Predictive Parity} = \frac{TP_p}{FP_p + TP_p} - \frac{TP_{up}}{FP_{up} + TP_{up}} \tag{5}$$

$$\text{Negative Predictive Parity} = \frac{TN_p}{FN_p + TN_p} - \frac{TN_{up}}{FN_{up} + TN_{up}} \tag{6}$$

To measure fairness, we transform probability equations into differences (Eq. 1 - 6). This is because difference values are bounded within the range of [-1,1], which makes them less susceptible to division by zero issues and more straightforward to interpret [19]. Furthermore, definitions based on differences are more versatile than those based on equations. In fact, when a fairness equation is satisfied, it means that the difference equals zero; we refer to this particular scenario as *perfect fairness*.

4 DISTRIBUTION-BASED ANALYSIS OF FAIRNESS MEASURES

In this research, we aim to explore how fairness measures behave when there are changes in class imbalance, protected group ratios, and stereotype ratios. As demonstrated in the previous section, the values of the fairness measures we are analyzing are obtained from confusion matrices, which depict the outcomes of classification on our experimental data.

By treating the training data as the result of a random process, we can offer a probabilistic perspective on the values of fairness measures. In particular, measures derived from confusion matrices can be seen as discrete random variables that convert a confusion matrix into a numerical value. Discrete random variables are usually defined by their probability mass functions, which indicate the likelihood that a discrete random variable exactly equals a certain value. These probability mass functions are often shown as histograms, with the x-axis representing measure values and the y-axis indicating the probability of achieving a specific value. We will use these visualizations to closely examine the fairness measures we are considering.

We will utilize probability mass functions represented as histograms to study the impact of different imbalance ratios and group ratios on the six fairness measures (Eq. 1–6). In our analysis, we distance ourselves from specific classifiers or datasets and thus presume that every possible confusion matrix is equally likely. We acknowledge that this is a significant assumption, but it allows us to provide the most comprehensive view of what could occur in any classification task. For instance, while confusion matrices with predominantly incorrect predictions may seem less likely than those with mostly correct ones, classifiers dealing with data streams will often make temporary incorrect predictions due to concept drift [3]. As a result, we will adopt an approach similar to that outlined in [2, 3], generate all possible confusion matrices for a dataset size n, and compute the measure's value for each matrix (Figure 2). It's important to note that for a dataset size n, the total number of all possible confusion matrices defined by eight values ($TP_p, FN_p, FP_p, TN_p, TP_{up}, FN_{up}, FP_{up}, TN_{up}$) equals: $\frac{\prod_{i=1}^{7}(n+i)}{7!}$.

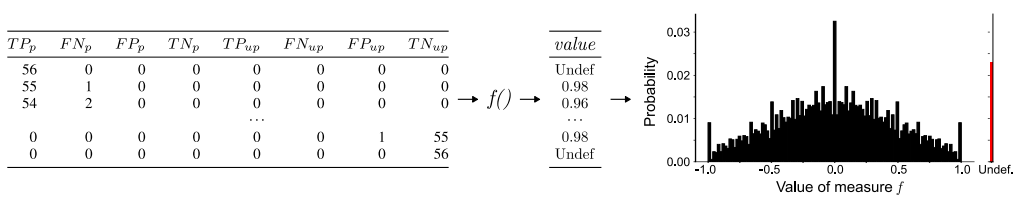

Fig. 2. The procedure involves creating a histogram that represents a fairness measure's probability mass function. This is achieved by generating all potential confusion matrices. In each confusion matrix, a row corresponds to a single entry (as shown on the left). These matrices are then used to map to specific fairness measure values (depicted in the middle). By tallying each of these measure values, we are able to construct a histogram (illustrated on the right). Any undefined measure value (marked as Undef.) is shown as a distinct red bar adjacent to the histogram.

We utilize the computed measure values to examine the histograms of each measure, which are based on the probability mass function. These histograms vary according to the *IR*, *GR* and]*mathitSR*. In our visualizations, we use

a dataset size of n=24 due to lack of machine power, with *IR* and *GR* and $SR \in \{\frac{1}{12}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{11}{12}\}$. Despite n=24 appearing to be a small dataset, it equates to 2629575 potential confusion matrices, offering a substantial sample of classifier outputs. The chosen class, group and stereotype proportions aim to depict class balance ($\frac{1}{2}$), minor imbalance ($\frac{1}{4}$ and $\frac{3}{4}$), and major imbalance ($\frac{1}{12}$ and $\frac{11}{12}$) [16]. It's worth noting that the extreme imbalance ratios ($\frac{1}{12}$ and $\frac{11}{12}$) represent a scenario where there are only two instances from the minority class, one for each protected/unprotected group.[1] We also observe that certain confusion matrices may yield undefined fairness measure values. This occurs when the denominators in Eq. 1–6 are zero. The likelihood of encountering such undefined values for a specific IR/GR/SR will be depicted as a red bar to the right of the probability mass function histogram (Figure 2).

Figures 3- 8 present histograms for correlations of different ratios for Accuracy Equality. Accuracy Equality and Statistical Parity(Figures S7-S12) change in a similar way. Equal Opportunity(Figures S13-S18) and Predictive Parity(Figures S19-S24) are changing in the way opposite to each other, the same way as Negative Predictive Parity(Figures S-31S36) and Positive Predictive Parity(Figures S25-S30).

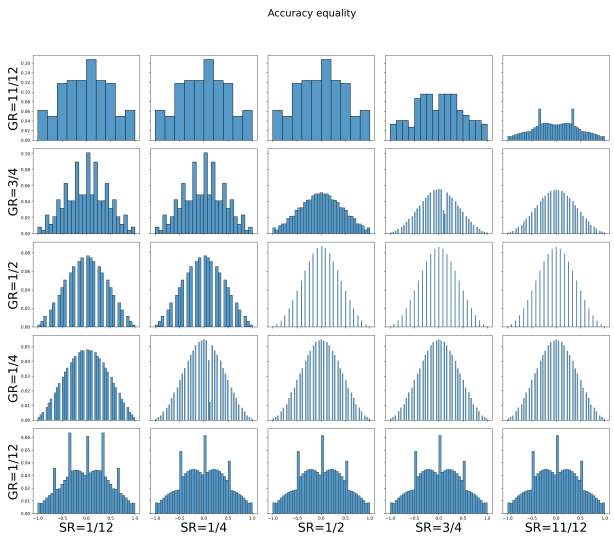

Fig. 3. Histograms of accuracy equality values for selected stereotype and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying *SR*(top-bottom) and *GR*(left-right).

[1]This is the minimum number of instances required to analyze binary class imbalance and fairness between two groups.

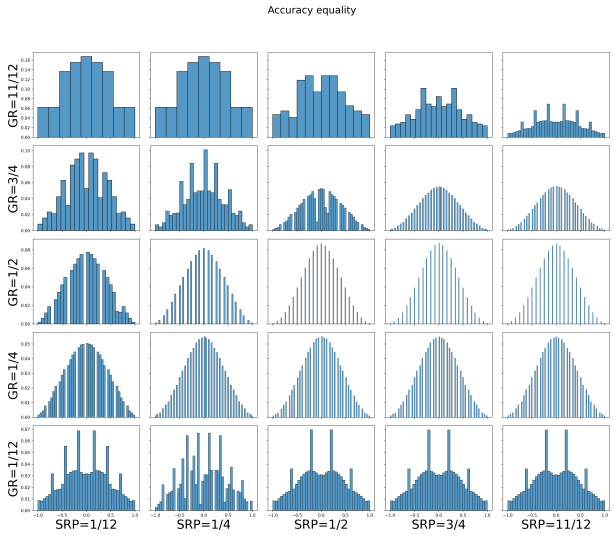

Fig. 4. Histograms of accuracy equality values for selected positive stereotype and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying SR_p(top-bottom) and GR(left-right).

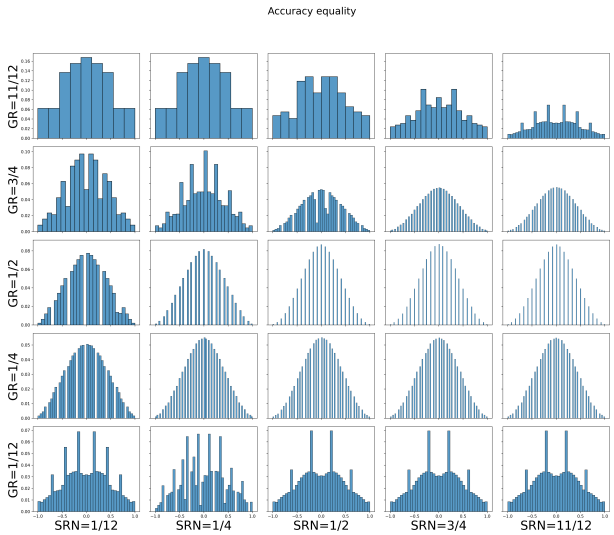

Fig. 5. Histograms of accuracy equality values for selected negative stereotype and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying SR_n(top-bottom) and GR(left-right).

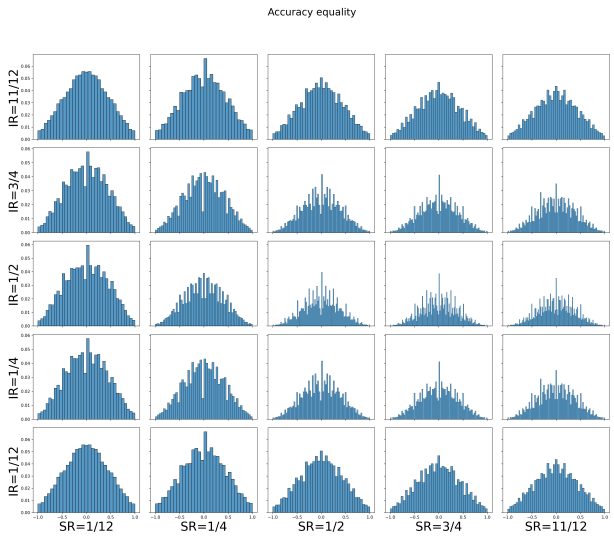

Fig. 6. Histograms of accuracy equality values for selected stereotype and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying *SR*(top-bottom) and *IR*(left-right).

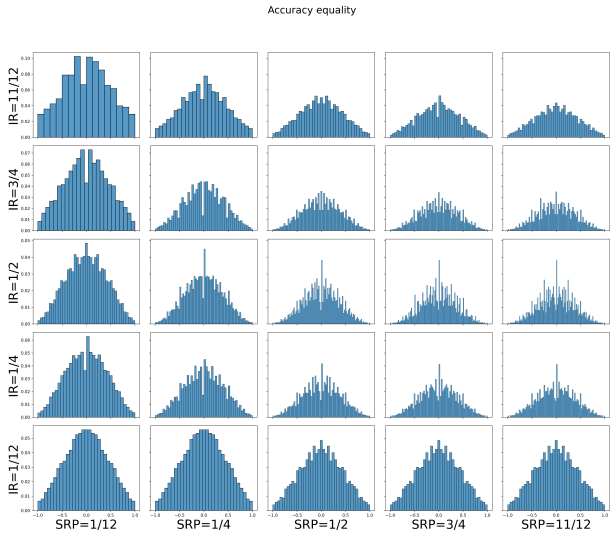

Fig. 7. Histograms of accuracy equality values for selected positive stereotype and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying SR_p(top-bottom) and *IR*(left-right).

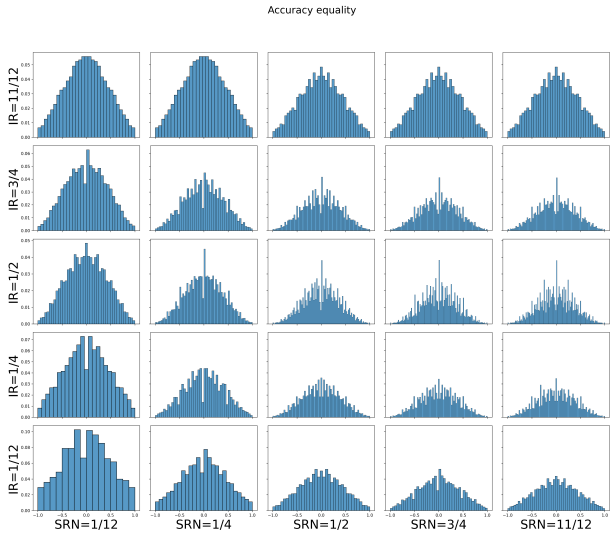

Fig. 8. Histograms of accuracy equality values for selected negative stereotype and group ratios. The x-axis shows possible measure values, whereas the y-axis shows the probability of obtaining a given value. Panels represent varying SR_n(top-bottom) and IR(left-right).

In addition to the analysis, we have also constructed line plots that illustrate the likelihood of attaining perfect fairness for different SR, which is when a fairness measure value equals 0. Figure 9 presents the proportion of confusion matrices that correspond to perfect fairness (represented on the y-axis) against varying imbalance and group ratios (represented on the x-axis) on the right and the probabilities of encountering undefined values of fairness measures on the left. It can be observed that the likelihood of obtaining an undefined value escalates with a significant imbalance ratio of $\frac{1}{12}$ to $\frac{12}{12}$. The probability of reaching perfect fairness for various SR values is generally consistent, with the exception of **negative predictive parity** and **positive predictive parity** when the values are close to 1.

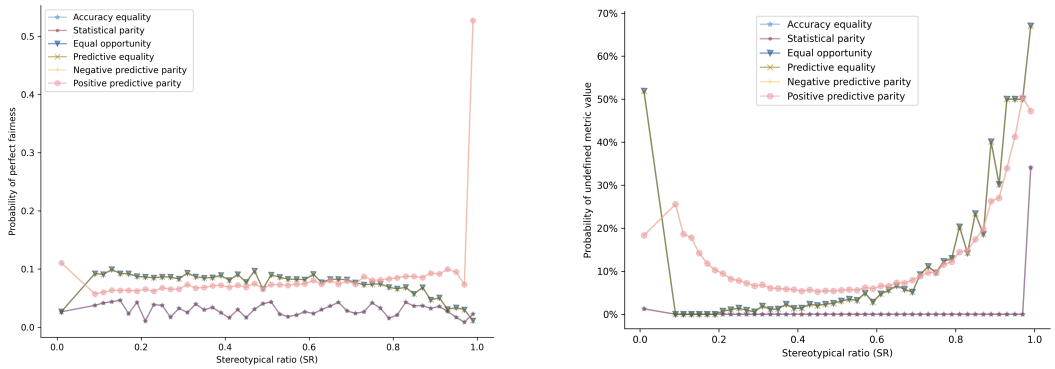

Fig. 9. Probability of achieving perfect fairness(left) and an undefined value(right) for different measures under varying stereotypical ratios

5 CONCLUSIONS

In this paper, we have analyzed the behavior of six popular group fairness measures in the context of varying class imbalance, protected group and stereotype ratio bias.

Our results show that all the analyzed measures change their behavior in the presence of class imbalance and stereotype ratio combined and in the presence of protected group bias and stereotype ratio combined. 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 [8]. 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

REFERENCES

[1] Dariusz Brzezinski, Julia Stachowiak, Jerzy Stefanowski, Izabela Szczech, Robert Susmaga, Sofya Aksenyuk, Uladzimir Ivashka, and Oleksandr Yasinski. 2023. Properties of fairness measures in the context of varying class imbalance and protected group ratios. (2023).

[2] 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.

[3] 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.

[4] 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.

[5] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv:2010.04053* (2020), 1–33.

[6] 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.

[7] 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.

[8] 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.

[9] Jannik Dunkelau and Michael Leuschel. 2019. *Fairness-aware machine learning: An extensive overview.* Technical Report. Universität Düsseldorf. 1–60 pages.

[10] Pratik Gajane and Mykola Pechenizkiy. 2018. On Formalizing Fairness in Prediction with Machine Learning. *arXiv 1710.03184* (2018), 1–6.

[11] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective.* Cambridge University Press, Cambridge, UK.

[12] 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.

[13] Luca Oneto and Silvia Chiappa. 2020. *Fairness in Machine Learning.* Springer International Publishing, Cham, 155–196. https://doi.org/10.1007/978-3-030-43883-8_7

[14] Andrea Romei and Salvatore Ruggieri. 2013. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29 (2013), 582–638.

[15] UNESCO. 2021. *Draft Recommendation on the Ethics of Artificial Intelligence.* Technical Report. UNESCO. Document code SHS/BIO/REC-AIETHICS/2021.

[16] Gary Weiss. 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications.* Wiley-IEEE Press, Hoboken, NJ, USA, Chapter Foundations of Imbalanced Learning, 13–43.

[17] Jeannette M. Wing. 2021. Trustworthy AI. *Commun. ACM* 64, 10 (2021), 64—-71.

[18] 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.

[19] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31 (2017), 1060–1089.