IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 January 2023, accepted 22 February 2023, date of publication 3 March 2023, date of current version 22 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3252370

RESEARCH ARTICLE

Measuring Imbalance on Intersectional Protected Attributes and on Target Variable to Forecast Unfair Classifications

MARIACHIARA MECATI, (Member, IEEE), MARCO TORCHIANO, (Senior Member, IEEE), ANTONIO VETRÒ, (Member, IEEE), AND JUAN CARLOS DE MARTIN, (Member, IEEE)
Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy

Corresponding author: Mariachiara Mecati (mariachiara.mecati@polito.it)

ABSTRACT Bias in software systems is a serious threat to human rights: when software makes decisions that allocate resources or opportunities, may disparately impact people based on personal traits (e.g., gender, ethnic group, etc.), systematically (dis)advantaging certain social groups. The cause is very often the imbalance of training data, that is, unequal distribution of data between the classes of an attribute. Previous studies showed that lower levels of balance in protected attributes are related to higher levels of unfairness in the output. In this paper we contribute to the current status of knowledge on balance measures as risk indicators of systematic discriminations by studying imbalance on two further aspects: the intersectionality among the classes of protected attributes, and the combination of the target variable with protected attributes. We conduct an empirical study to verify whether: i) it is possible to infer the balance of intersectional attributes from the balance of the primary attributes, ii) measures of balance on intersectional attributes are helpful to detect unfairness in the classification outcome, iii) the computation of balance on the combination of a target variable with protected attributes improves the detection of unfairness. Overall the results reveal positive answers, but not for every combination of balance measure and fairness criterion. For this reason, we recommend selecting the fairness and balance measures that are most suitable to the application context when applying our risk approach to real cases.

INDEX TERMS Data bias, data imbalance, intersectionality, algorithmic fairness, automated decision-making, data ethics.

I. INTRODUCTION

The problem of bias in information systems, although present in the scientific literature of software systems during the past quarter century –e.g., see the pioneering work proposed in [1]– got wider attention only in the mid 2010s, in connection with the large investments in Artificial Intelligence (AI) / Machine Learning (ML), digital automation of organizational processes and, in general, automated decision-making (ADM) systems. At that time, several studies and journalistic investigations rapidly attracted the interest of the public at large by showing how software systems may perpetuate and even exacerbate existing inequalities, for example, the work presented by Barocas and Selbst [2], or the influential book by O'Neil [3]. The swiftly achieved relevance of the topic contributed to the birth of a new field of research, whose main forum today is the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT, previously ACM FAT, founded in 2018[1]). The topic has been relevant for policymakers, too: for example, *avoidance of unfair bias* is one of the key requirements listed in the Ethics Guidelines for Trustworthy AI [4], a foundational document for the European efforts to regulate AI, currently going through the last steps of the European legislative process. In the meantime, the major institutions for technology standardization are also devoting special attention to the

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco J. Garcia-Penalvo.

[1] See https://facctconference.org/2018/index.html

topic: in 2022, the US National Institute of Standards and Technology has published the draft of the future Standard for Identifying and Managing Bias in Artificial Intelligence [5]. This initiative follows the publication of the Technical Report "Bias in AI systems and AI aided decision making" by the International Standard Organization [6].

The potential danger posed by AI and, in general, by data-driven technologies to fundamental human rights, highlighted by several jurists such as in [7] and [8], is the main driver of the many initiatives in the field of ethics and governance of AI, of which the ones mentioned above are only a small fraction.[2] Bias in software systems is also a threat to human rights, as disparate software impact may discriminate against people based on their personal traits (e.g., gender, ethnic group, etc.). This can take the form of both denied opportunities and adverse decisions. Examples for this first case are a differential advertising of job ads based on gender and ethnic group [10], and an algorithmic assignment to an intensive care program skewed towards specific ethnic groups [11]; examples of the latter are the recidivism risk predictions skewed towards black people in the well known COMPAS case [12], or the impact of birthplace and gender on the automated prices quoted to drivers in the Italian car insurance industry [13]. Often the cause for the software-biased impact lies in the imbalance of training data [14], that is an unequal distribution of data between the classes of one or more attributes [15]. Unequal distribution has been recognized as a critical aspect in the machine learning domain for a long time [16] –and it is still relevant [17]– because it skews the performances of classifiers, leading to varying accuracy among the classes of given attributes in the data. This consequence has been documented in a variety of domains and technologies, for example, male dominance in training data can perpetuate such bias in the output of automatic generation of images [18] or in the selection of CV's [19], while the geographic imbalance in the content production that feeds recommender systems can generate (dis)advantage to a specific group [20].

Given this causal relationship, it is possible to detect the risk of bias in the classification output by measuring the level of (im)balance of specific attributes in a dataset. Our previous studies showed that lower levels of balance in protected attributes are related to higher levels of unfairness in the output [21], [22].

In this work we move forward on the assessment of balance measures as risk indicators of systematic discrimination by including two more aspects: i) intersectionality among the classes of protected attributes, and ii) the contribution of the target variable to the unfairness detection. The first aspect is relevant because social identities and inequality are interdependent for groups –such as *black women*– and not mutually exclusive [23]. The second aspect has been recognized as a challenge in a variety of domains, for example, fraud detection, network intrusion detection, medical diagnostics, and a number of other fields [24]: often negatively labeled instances significantly outnumber positively labeled instances, but the latter are associated with the most relevant events for end users (e.g., a fraud). To our knowledge, none of the current approaches to intersectionality and to the target variable combined with protected attributes (and to their effects on classifications) use synthetic indicators to measure balance. This is our main contribution to the state of the art.

On the basis of the above motivations, we put forward the following research questions:

> RQ 1. How do intersectional attributes relate to the corresponding primary attributes, in terms of balance and fairness?

Given the crucial importance of intersectional classes in understanding discrimination risks and inequalities that are even exacerbated in correspondence of the intersection of certain social identities, we believe that it is fundamental to better understand their nature. In particular, it is important to understand to what extent the imbalance of the primary attributes (binary or multiclass) affects the imbalance of the intersectional attribute, as well as how fairness with respect to an intersectional attribute is linked to fairness with respect to the primary attributes.

> RQ 2. Can the measure of balance on intersectional attributes detect unfairness risks?

There is evidence that working at the level of protected primary attributes, the balance of the attribute classes can detect the risk of classification unfairness with respect to such attributes. Our goal here is to understand whether this capability extends to intersectional attributes too.

> RQ 3. Does the combination of the target variable with protected attributes improve the detection of unfair classification risks?

The imbalanced distribution of target classes can be taken into consideration by looking at their combination with protected attributes (both primary and intersectional) and assessing whether the combined balance can detect the risk of unfair classification. Note that in the following we call *combined attributes* those attributes given by the combination of the target variable with protected attributes (primary or intersectional).

The paper is organized as follows: in this Section, we outlined the research context and we discussed the theoretical foundations of our proposal, while in Section II we position our work in relation to the existing literature by showing how it is linked to several existing research strands. In Section III we present our experimental design:

[2] As a matter of fact, consider the dozens of principles and guidelines for ethical artificial intelligence (AI) issued by private companies, research institutions and public sector organizations [9].

first we describe the research method we followed, the dataset we analyzed, and the mutation techniques that we adopted to create a large number of synthetic datasets starting from the original dataset, then we show the balance measures and the fairness criteria we employed to conduct our experiment. In Section IV we report the analysis of our results with the related discussion and explanation, as well as an overview of the limitations to be addressed in future lines of research (Subsection IV-D). Finally, in Section V we briefly summarize the whole experiment, draw our conclusions, and outline possible future work.

II. RELATED WORK

Our research can be located in the area of algorithmic bias and fairness: as summarized in the introduction, in the last few years an important collective effort has been devoted to this field of research for the purpose of exploring and improving novel strategies to make outcomes from automated systems as equitable and unbiased as possible. For better positioning our study in this large area of inquiry, we identify the following characterizing features:

- our study focuses on inputs and processes, to contribute to filling a gap in the literature as identified by Firmani et . [25], by Hutchinson and Mitchell "Returning to the idea of unfairness suggests several new areas of inquiry [...] a shift in focus from outcomes to inputs and processes" [26], and by Pitoura *"There is a need to consider social-minded measures along the whole data pipeline"* [27].
- The balance measures can be incorporated into existing data labeling schemes (e.g., the Dataset Nutrition Label [28]) or toolkits for bias detection and mitigation (see the landscape synthesis in [29] that neglect balance or the FairMask algorithm proposed in [30] for bias mitigation).
- The proposed methodology address the need to better document the AI pipeline, particularly relevant in the algorithmic fairness community as shown in the exhaustive work of Fabris et al. [31]. Reporting imbalances in a synthetic and meaningful way is part of the necessary further efforts of the AI/ML community in devoting more attention to the dataset documentation, as acknowledged by Königstorfer and Thalmann: "one should also record whether there were imbalances in the training data with regards to the target categories or how these imbalances were corrected" [32].
- The whole approach is coherent with the ISO standards on data quality and risk management, as analytically described in [33] that originated the series of studies to which this paper belongs.

Regarding the latter point, we highlight that our previous studies tested the reliability of the balance measures only towards single protected attributes: in the first one [21], we tested the measures on a few hypothetical exemplar distributions; then, we run more exhaustive tests by applying mutation techniques to generate a number of derived synthetic datasets having different levels of balance, in one case to binary attributes [22] and in the other case to multiclass attributes [34]. Two fundamental recurring elements between these studies and the current one are i) the experimental procedure, since the method that we adopted to collect synthetic data remains unchanged (see Section III for the detailed description of the experimental design), and ii) the computation of the relationship between imbalance and unfairness, in accordance with the usage of balance measures as indicators of the risk of systematic discrimination. Instead, the marginal difference and novel contribution of this paper are given by i) the integration of the concept of intersectionality between the classes of two or more single protected attributes, and ii) the consideration of the distribution of protected attributes in the target variable.

Intersectionality was introduced in the late 80s in the Black Feminist literature in relation to the intersection of gender and race [35] and it has been successively extended to embrace other traits such as disability status, socioeconomic class, sexual orientation, etc. The concept has recently appeared in the context of fairness and machine learning, related to issues of intersectional discrimination in different domains: Buolamwini and Gebru studied the impact of the intersection of gender and skin color on computer vision performance [36]; Holman et al. explored intersectionality in the medical field [37]; whereas Subramanian et al. advocated for the use of intersectional groups in the validation of NLP models to better intercepts the social and cultural biases reflected in the corpus of training data [38]. Other works present attempts of introducing intersectionality in fairness measures [39] and in causal models [40]: however, up to our knowledge, none of these and other studies in the AI/ML fairness literature constructed and applied synthetic measures of (im)balance to intersectional protected attributes.

As far as the imbalance of the target variable is concerned, a comprehensive survey has been conducted by Branco et al., who collected existing techniques for handling the problem for both classification and regression tasks [24]. The same authors examined more in-depth the context of regression tasks [41], where the target variable is continuous: they presented three new pre-processing approaches to tackle the problem of forecasting rare values of a continuous target variable. Other works concern the mitigation of the imbalance issue of the target variable, and they have been developed with the aim of improving the predictive accuracy of rare cases in forecasting tasks through the adoption of different resampling methods (e.g., see [42], [43]). The closest work to ours is the one by Thabtah et al. [44], who studied the impact of varying class imbalance ratios on classifier accuracy: they identify nine different imbalance ratios (from 10%:90% to 90%:10%, with steps of 10% increase/decrease) and compute their effect on standard measures of classifier performance (error rate, predictive accuracy, recall and precision). Thus, they focus on the nature of the relationship between the degree of class imbalance and the corresponding classifier performance, but they neither use specific and synthetic measures of balance nor consider multilevel attributes. The same consideration applies to the other studies mentioned above.

III. EXPERIMENTAL DESIGN

With the goal of investigating the research questions outlined above, we create a number of synthetic datasets by aggregating sensitive attributes (also with the target variable) and mutating the distributions of the occurrences between their classes. Then, we chose four indexes that can evaluate balance in the data –and thus the lack of balance, i.e. imbalance– as well as a set of fairness criteria to measure unfairness occurring in the classification outcomes. Note that both the selection of the dataset and the experimental procedure, with related mutation techniques, balance measures, and fairness criteria, reflect the same pattern and definitions already analyzed in our previous works [22], [34], since the method that we adopted to collect synthetic data – on which we conducted our analysis to answer the different research questions – remains unchanged.

After that, we assessed whether the balance/unfairness of intersectional attributes can be inferred from the balance/unfairness of the primary attributes, and finally, whether the combination of a protected attribute with the target variable improves identifying the unfairness.

We defined an intersectional attribute as a multiclass attribute whose classes are given by the combination –in all the possible ways– of the classes of (single) primary attributes that can be either binary or multiclass. Similarly to the definition of imbalance already stated in the previous Section, intersectionality is between-attributes when only two attributes are taken into consideration, or multiattribute when the intersectionality involves multiple attributes. In this article, we will explore the concept of multi attributes intersectionality in greater detail.

In general, data is imbalanced with respect to the target variable if at least one of the target variable values has a significantly smaller number of instances when compared to the other values.

Specifically, we set up the following procedure:

1) we chose a sizable *dataset* (as described in Section III-A) that includes two *protected attributes*[3]: the multiclass attribute "education" with cardinality m equal to 4, and the binary attribute "sex" (with m=2);
2) several derived synthetic datasets with different levels of balance have been generated by means of two suitable *mutation techniques*: specifically, we adopted two processing methods, one specific for multiclass attributes and one for binary attributes. We adjusted the parameters of the two methods to alter the distribution of occurrences among the classes –and consequently the balance– of the two protected attributes under analysis (see Section III-B);

TABLE 1. Balance measurements with the respective unfairness measurements for each protected attribute.

Balance measurement	Unfairness measurement
$\mathfrak{B}$(sex)	$\mathfrak{U}$(sex)
$\mathfrak{B}$(education)	$\mathfrak{U}$(education)
$\mathfrak{B}$(sex_education)	$\mathfrak{U}$(sex_education)
$\mathfrak{B}$(sex_target)	$\mathfrak{U}$(sex)
$\mathfrak{B}$(education_target)	$\mathfrak{U}$(education)
$\mathfrak{B}$(sex_education_target)	$\mathfrak{U}$(sex_education)

3) we aggregate the two primary protected attributes in one intersectional attribute "sex_education" by combining the classes in all the possible ways, thus creating an intersectional multiclass attribute of cardinality m equal to 8 (= 2 "sex" × 4 "education"); likewise we aggregate the three previous attributes with the target variable, obtaining three combined multiclass attributes, i.e. "sex_target" (m=4), "education_target" (m=8) and "sex_education_target" (m=16);
4) we used four different *balance measures* $\mathfrak{B}$ (as outlined in Section III-C) to compute the level of (im)balance of both the primary protected attributes and the intersectional attribute in the training set;
5) we built a *binomial logistic regression* model in order to forecast the *score variable* for each synthetic dataset: we trained a binary classifier on a training set composed of the 70% (chosen randomly) of the data, and then tested it on the remaining 30% (which represents the test set);
6) we applied three different *fairness criteria* $\mathfrak{U}$ (see Section III-D) to both the primary protected attributes and the intersectional attribute in the test set –i.e. to the classifications obtained from the model– for a total of five distinct unfairness measures, following the pattern described in Table 1; note that for the protected attributes combined with the target variable we compute the unfairness on the corresponding protected attribute *not* combined with the target in the test set;
7) we analyzed the collected results in order to answer the research questions.

A. DATASET SELECTION

We selected a dataset from the financial services context, as it is one of the most considerable application domains of ADM systems: **Default of Credit Card Clients**, which has been retrieved by the Kaggle platform.[4]

The dataset has been chosen because of the high impact of using ADM systems in the financial domain and for its popularity at the time of the research, when it was ranked as the fourth most voted dataset on credit cards on Kaggle.[5]

Moreover, as we are interested in datasets that collect data on

[3] For identifying an attribute as *protected*, we considered as reference the definition provided in "Article 21 - Non-discrimination" of the EU Charter of Fundamental Rights [45].

[4] https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

[5] https://www.kaggle.com/datasets?search=credit+card&sort=votes last visited on January 13, 2023.

persons, Dccc fits better our research work than the dataset "Credit Card Fraud Detection" ranked first, which is based on transactions.

The properties of Dccc have been summarized in Table 2: it is composed of 25 variables and contains information on demographic factors, history of payment, credit data, default payments and bill statements of credit card clients in Taiwan from April 2005 to September 2005. In particular, we took into account two protected attributes: the first one is "education", which is composed of six classes in the original dataset, but two of the classes –i.e., *NA* and *unknown*– do not represent an actual category of individuals, therefore we exclude such unknown and missing values (NA) from the analysis; thus, the resulting dataset is composed of 29655 rows, where the classes of the protected attribute "education" are composed as follows: 10585 *graduate school*, 14030 *university*, 4917 *high school* and 123 *others*. The second protected attribute is the binary attribute "sex", which is composed of 11760 instances of the class *male* and 17895 instances of the class *female*.

In addition, note that this dataset does *not* contain a pre-computed classification, thus we implemented a *binomial logistic regression* model in order to foresee the *score* variable: specifically, we trained a binary classifier on a *training* set represented by the 70% (randomly selected) of the original dataset and we ran it on the *test* set, composed of the remaining 30% of the data. Moreover, note that in real datasets we can often find missing values (NA): as we were interested in examining existing intersectional classes of protected attributes, we choose to keep out missing values from the analysis.

B. MUTATION TECHNIQUES

Two distinct *pre-processing* methods were employed as mutation techniques to generate multiple variations of the distribution of the occurrences between the classes of the protected attributes taken into account. Specifically, we chose these methods as they are two widely used re-balancing techniques whose we analyzed –in our previous works [22], [34]– the effects on different datasets, other than analyzing the response of the balance measures to such variations in the protected attributes: as a matter of fact, we obtained a positive response both in the case of binary and multiclass attributes by observing the behavior of the balance measures as the relevant mutation parameter varies, with an increase in the balance measures as the distribution of the occurrences between the classes become more and more balanced.

a: MULTICLASS ATTRIBUTE

To mutate the classes of the multiclass protected attribute "education", we adopted the R `UBL-package`,[6] which offers a variety of pre-processing functions to address both classification (binary and multi-class) and regression issues that include non-uniform costs and/or benefits. Specifically, we employed the `SmoteClassif` function[7] as mutation technique, which deals with imbalanced classification problems by means of the SMOTE method, thus creating a new "smoted" dataset that resolves the class imbalance problem.

This method has been applied with the following settings:

- "`education∼`" is the multi-class protected attribute –composed of four distinct categories– which is used as a formula;
- "`C.perc`" is a list that holds the percentages of under-sampling or/and over-sampling to apply to each class of the protected attribute selected as a formula: thus, a class remains unchanged if the number 1 is provided for that class, while an under-sampling percentage is a number below 1, and an over-sampling percentage should be a number above 1; this also means that there exists an infinite number of possible combinations of the percentages for each class. Otherwise, `C.perc` may be set to "balance" (the default value), which represents a case where the sampling percentages are automatically estimated to balance the examples between the minority and majority classes, or to "extreme", through which the distribution of examples across the existing classes is inverted by changing the majority classes into the minority, and vice-versa;
- "`repl=FALSE`" is a boolean value that can be used to prevent the repetition of examples when conducting an under-sampling of the majority class(es).

In our study, we chose to analyze five distinct cases for the parameter `C.perc`. We first set the parameter to the default value "balance" –i.e., the perfect uniform distribution–, then we adopted four different lists of percentages for the categories of the protected attribute, previously defined and studied in [21] as "exemplar distributions":

- *Quasi Balance*: half of the classes are 10% lower than the percentage of the max balance case, while the other half of the classes are 10% higher;
- *One Off*: occurrences are equally distributed between all the classes except for one, which is empty;
- *Half High*: the majority of occurrences are found in one-half of the classes, while the other half has a much lower frequency; in particular, we defined the frequencies of the two halves by setting a ratio of 1:9;
- *Power 2*: the number of occurrences between the different classes increases exponentially according to a power law with base 2.

Finally, for each exemplar distribution we looked at 4 different permutations of the values of the percentages assigned to the various classes of the protected attribute. For instance, in the *One Off* configuration the four different permutations have each a different class with zero occurrences.

[6] https://rdocumentation.org/packages/UBL/versions/0.0.6/topics/UBL-package, last visited on January 13, 2023

[7] https://www.rdocumentation.org/packages/UBL/versions/0.0.6/topics/SmoteClassif, last visited on January 13, 2023

TABLE 2. Summary of the most important properties of the dataset.

Dataset	Size	Domain	Target variable	Protected attributes	Cardinality (m)
Default of credit cards clients (Dccc)	30000×25	Financial	Default payment next month	Education	4
				Sex	2

b: BINARY ATTRIBUTE

To mutate the classes of the binary protected attribute "sex", we adopted the R `ROSE-package`[8] [46], which provides functions to deal with binary classification problems in the presence of imbalanced classes. Specifically, we applied the `ovun.sample` function[9] as mutation technique, which creates possibly balanced samples by random over-sampling minority examples, under-sampling majority examples or combination of over- and under-sampling.

This technique has been implemented using the following settings:

- "`sex~`" is the binary protected attribute chosen as formula, since it is one of the most common sources of imbalance and consequent discrimination;
- "`both`" as method, which indicates a combination of over-sampling minority examples and under-sampling majority examples to perform the random sampling;
- "`N`" equal to the same number of rows of the dataset under analysis as the desired sample size of the resulting dataset;
- "`p`" represents "*the probability of resampling from the rare class*" and it has been set to 17 different values in order to vary as much as possible the distribution of the occurrences between the two categories of the attribute "sex": 0.01 (corresponding to the case of minimum balance), 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5 (maximum balance), 0.6, 0.7, 0.8, 0.9, 0.925, 0.95, 0.975, 0.99. When the value of `p` is set to 0.5, it indicates a uniform distribution between the two classes; lower values of `p` will result in a less balanced distribution, while increasing the value from 0.5 to 1 will lead to a more balanced distribution, but with inverted proportions;
- "`seed`" is "*a single value, interpreted as an integer, recommended to specify seeds and keep track of the sample*", therefore we decided to vary such value by randomly selecting 50 values between 1 and 1000, in order to enhance the variability and consequently the reliability of our approach.

Note that in both cases –multiclass attributes and binary attributes– the generated mutated datasets have the same number of rows as the original ones, and the distribution of the other variables in the dataset remains unchanged.

Finally, with a view to increasing the variability and the reliability of our approach, given the random nature of the resampling we decided to vary a `seed` (an integer value used to ensure reproducibility and keep track of the samples) by randomly generating 50 different values between 1 and 1000. This means that for the analysis and discussion of the results we always kept track of the outputs for each value of the `seed` and for each value of the parameter `p`, then: for the mutations obtained by setting `C.perc`="balance" we collected a total of 6 (attributes) × 50 (seed) × 17 (levels of `p`) = 5100 values for the *balance measures* and 5100 values for the *fairness criteria*. For the mutations obtained through the four different lists of percentages instead, we gathered a total of 6 (attributes) × 50 (seed) × 17 (levels of `p`) × 4 (exemplar distributions) × 4 (permutations) = 81600 values for both the *balance measures* and the *fairness criteria*; the sum of all these elements adds up to
5100 + 81600 = 86700 values for the *balance measures* and 86700 values for the *unfairness measures*. Note that the 6 attributes are those summarized in Table 1.

C. BALANCE MEASURES

This research was limited to *categorical* attributes and the same four indexes of data balance were chosen from our prior studies (see Table 3). The measures were normalized to meet two criteria: (i) range in the interval [0, 1]; (ii) share the same interpretation: the closer the measure is to 1, the more balanced the distribution of frequencies across categories; conversely, values closer to 0 indicate that the frequencies are concentrated in fewer categories, resulting in an imbalanced distribution.

a: GINI INDEX

The Gini index is a commonly used measure of heterogeneity that shows the number of different types present. It is used in various fields, such as market competition, political polarization, ecological diversity, and even racial discrimination. In terms of statistics, the heterogeneity of a discrete random variable can range from a degenerate case (the lowest level of heterogeneity) to an equiprobable case (the highest level of heterogeneity, where all categories are equally represented). Thus, given a certain number of categories, the heterogeneity increases when the probabilities of each class become more equal, meaning that each class has a similar representation.

b: SHANNON INDEX

Species diversity in a community is a widely accepted concept in ecology, biology, and phylogenetics. By taking into account the relative amounts of different species

[8] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ROSE-package, last visited on January 13, 2023

[9] https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample, last visited on January 13, 2023

TABLE 3. The *balance measures* with the related formula: given a discrete random variable with m classes, we define as f_i the proportion of the class i w.r.t. the total, where $i = 1, \ldots, m$:.

Gini	$\mathrm{G} = \frac{\mathrm{m}}{\mathrm{m}-1} \cdot (1 - \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{f}_\mathrm{i}^2)$
Simpson	$\mathrm{D} = \frac{1}{\mathrm{m}-1} \cdot \left(\frac{1}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{f}_\mathrm{i}^2} - 1\right)$
Shannon	$\mathrm{S} = -\left(\frac{1}{\ln \mathrm{m}}\right) \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{f}_\mathrm{i} \ln \mathrm{f}_\mathrm{i}$
Imbalance Ratio	$\mathrm{IR} = \frac{\min(\{\mathrm{f}_{1..\mathrm{m}}\})}{\max(\{\mathrm{f}_{1..\mathrm{m}}\})}$

(categories), the indexes of diversity can be very useful to measure the imbalance of a community's composition.

c: SIMPSON INDEX

This index is another measure of diversity: it calculates the probability that two randomly chosen individuals from a sample belong to the same species, i.e., the same category. It is employed in economics and social sciences to measure equity, uniformity, and wealth, and in ecology to assess the diversity of living organisms in a given area.

d: IMBALANCE RATIO

The Imbalance Ratio (IR) is a commonly used metric that is computed by dividing the highest frequency by the lowest frequency. To make it comparable to the other balance measures, we take the inverse in order to normalize it to a range of $[0, 1]$, where lower values indicate a higher imbalance.

D. FAIRNESS CRITERIA

We evaluated the *unfairness* of automated classification outputs on the basis of three criteria formalized in [47] in chapter 3 "Classification". Note that to assess the unfairness of a classification outcome, we will refer to *"Unfairness measures"* and *"Fairness criteria"* interchangeably, as we consider the Fairness criteria to be indicators of unfairness.

Generally, to evaluate the unfairness we take into consideration a protected categorical attribute A that can assume various values $(a_1, a_2, \ldots)$, a target variable Y, and a predicted class R where Y is binary (i.e., $Y = 0$ or $Y = 1$ and thus $R = 0$ or $R = 1$). In practice, the aim of our study is to assess the fairness of an ADM system in relation to the various values of a protected attribute when assigning a predicted class.

a: INDEPENDENCE CRITERION

To determine if the acceptance rate is the same across all groups, we can use the concept of demographic parity or statistical parity, which requires the probability of acceptance (i.e. $R = 1$) to be equal for all groups. This means that the independence criterion enforces groups to have equal selection rates. In terms of probability, it is represented by the following condition:

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) = \ldots$$

If A is binary (that is, $A = a_1$ or a_2), then we can compute the Independence unfairness measure as:

$$\mathfrak{U}_I(a_1, a_2) = |P(R = 1 \mid A = a_1) - P(R = 1 \mid A = a_2)|$$

b: SEPARATION CRITERION

In simple terms, when the protected characteristic is linked to the target variable –as it happens in many contexts– the separation criterion allows correlation between the score and the sensitive attribute as long as it is justified by the target variable. Indeed, this criterion is also known as equalized odds, equality of opportunity, or even conditional procedure accuracy. Specifically, the separation criterion requires the true positive rate and false positive rate to be equivalent for each level of the protected attributed being examined:

- $P(R = 1 \mid Y = 1, A = a_1) =$
  $= P(R = 1 \mid Y = 1, A = a_2) = \ldots$

- $P(R = 1 \mid Y = 0, A = a_1) =$
  $= P(R = 1 \mid Y = 0, A = a_2) = \ldots$

Therefore, if A is binary we can calculate two Separation unfairness measures ($\mathfrak{U}$) in the following ways:

- $\mathfrak{U}_{Sep\_TP}(a_1, a_2) =$
  $|P(R = 1 \mid Y = 1, A = a_1) - P(R = 1 \mid Y = 1, A = a_2)|$

- $\mathfrak{U}_{Sep\_FP}(a_1, a_2) =$
  $|P(R = 1 \mid Y = 0, A = a_1) - P(R = 1 \mid Y = 0, A = a_2)|$

c: SUFFICIENCY CRITERION

Given a certain protected attribute, this criterion implies the calibration of the model for the different categories, that is, Parity of Positive/Negative predictive values (respectively $R = 1$ or 0) for each level of the protected attribute:

- $P(Y = 1 \mid R = 1, A = a_1) =$
  $= P(Y = 1 \mid R = 1, A = a_2) = \ldots$

- $P(Y = 1 \mid R = 0, A = a_1) =$
  $= P(Y = 1 \mid R = 0, A = a_2) = \ldots$

As before, if A is binary we can compute two Sufficiency unfairness measures ($\mathfrak{U}$) as follows:

- $\mathfrak{U}_{Suf\_PP}(a_1, a_2) =$
  $|P(Y = 1 | R = 1 \wedge A = a_1) - P(Y = 1 \mid R = 1 \wedge A = a_2)|$

- $\mathfrak{U}_{Suf\_PN}(a_1, a_2) =$
  $|P(Y = 1 | R = 0 \wedge A = a_1) - P(Y = 1 \mid R = 0 \wedge A = a_2)|$

When dealing with non-binary attributes, that is $m > 2$, All the definitions above can be extended by considering the mean of indexes can be computed by taking all the possible pairs of levels in A:

$$\mathfrak{U}(a_1, \ldots, a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathfrak{U}(a_i, a_j)$$

Finally, we remind that all the unfairness measures previously described range in the interval [0, 1]: the higher the values the higher the unfairness, thus a value equal to zero indicates a perfectly fair classification, while a value close to 1 means unfair behavior.

IV. RESULTS AND DISCUSSION

In this Section we examine and discuss the results of our investigation, according to the three research questions we formulated above.

A. RQ 1 - INTERSECTIONAL VS. PRIMARY ATTRIBUTES

1) METHOD

In order to investigate the relationship between intersectional and primary attributes we observe the results of an ANOVA on two linear regression models, one for the balance measures and the other for the unfairness measures.

$$\begin{aligned}&\mathfrak{B}(sex\_education) =\\&= c_{sex} \cdot \mathfrak{B}(sex) + c_{education} \cdot \mathfrak{B}(education) + c_0\end{aligned}$$

$$\begin{aligned}&\mathfrak{U}(sex\_education) =\\&= c_{sex} \cdot \mathfrak{U}(sex) + c_{education} \cdot \mathfrak{U}(education) + c_0\end{aligned}$$

The first model was applied with all the four distinct balance measures reported in Section III-C and the second model was evaluated using all the five unfairness measures described in Section III-D. To answer our research question we look at two results from the analysis: adjusted R^2 and p-value. The adjusted R^2 is a goodness-of-fit measure for linear models and it is an indicator of the model accuracy, as it identifies the percentage of variance in the output variable that is explained by the input variables. In fact, R^2 tends to optimistically estimate the fit of the linear regression: a value of 1 indicates a model that perfectly predicts dependent values, whereas a value closer to 0 means that the model has no predictive capability. Thus, in our specific case, values of R^2 close to 1 mean that the measure related to the intersectional attribute can be explained by those related to the primary attributes. Smaller values indicate that the intersectional attribute cannot be explained by primary attributes alone. To assess the statistical significance of the results, we observe the p-value and consider significant a relationship whose p-value is lower than 5%. In addition, looking at the coefficients c_{sex} and $c_{education}$, we evaluate whether the two primary attributes provide an equal contribution.

2) BALANCE

The results of the regression for the balance measures are reported in Table 4. We observe that in the cases of the Gini index the R^2 is very close to 1 (0.941), and for the Shannon and Simpson indexes the R^2 is around 0.86; while it is much smaller (0.540) for the Imbalance Ratio index. For all the cases the p-value is $< 2.2 \cdot 10^{-16}$, indicating statistically significant results. This means that in three cases out of four, the balance measures related to the intersectional attribute can be explained by those related to the primary attributes, thus we can accurately infer the balance of the multiclass intersectional attribute from the balance of the primary attributes which compose the intersectional attribute itself; in the case of IR we have a smaller correlation, probably due to the fact that for many data points the IR assumes values close to zero more frequently than the other measures.

TABLE 4. Balance measures: evaluation of the linear regression model sex_education~sex+education.

Balance measure	Adjusted R^2	p-value	Coefficients		
			c_0	c_{sex}	$c_{education}$
Gini	0.941	$< 2.2 \cdot 10^{-16}$	17.308	0.195	0.671
IR	0.540	$< 2.2 \cdot 10^{-16}$	-2.908	0.475	0.107
Shannon	0.877	$< 2.2 \cdot 10^{-16}$	13.296	0.345	0.533
Simpson	0.850	$< 2.2 \cdot 10^{-16}$	-6.652	0.476	0.564

In addition, we computed the regression coefficients, reported in the three rightmost columns of Table 4. Indeed, looking at the coefficients c_{sex} and $c_{education}$, we observed that overall the balance measurements of the primary attributes have a high positive correlation with the intersectional attribute: in particular, such a positive correlation is higher in correspondence of the primary multiclass attribute "education", except for the IR index, which presents a coefficient $c_{education}$ much smaller with respect to the other coefficients.

Therefore, as concerns the balance measures, we positively answer the first research question:

> The measures of the Gini, Shannon and Simpson indexes related to the intersectional attribute can be explained by those related to the primary attributes; while the measure of the IR index related to the intersectional attribute is explained by the measures of the IR index related to primary attributes alone to a smaller extent with respect to other indexes.

3) UNFAIRNESS

The results of the regression for the unfairness measures are reported in Table 5. Differently from the Balance measures, for the fairness criteria we observe overall lower values of the adjusted R^2: in particular, we found values of R^2 around 0.6 for the independence, separation-TP and sufficiency-PN criteria, and even lower values – around 0.4 – in the case of the separation-FP and sufficiency-PP criteria. For all the cases

TABLE 5. Unfairness measures: evaluation of the linear regression model sex_education∼sex+education.

Unfairness measure	Adjusted R^2	p-value	Coefficients		
			c_0	c_{sex}	$c_{education}$
Independence	0.624	$< 2.2 \cdot 10^{-16}$	0.858	0.395	0.567
Separation – TP	0.625	$< 2.2 \cdot 10^{-16}$	1.971	0.472	0.622
Separation – FP	0.395	$< 2.2 \cdot 10^{-16}$	0.525	0.610	0.547
Sufficiency – PP	0.393	$< 2.2 \cdot 10^{-16}$	5.432	0.285	0.430
Sufficiency – PN	0.549	$< 2.2 \cdot 10^{-16}$	1.896	0.249	0.646

the p-value is $<2.2 \cdot 10^{-16}$, indicating statistically significant results. In addition, we computed the regression coefficients, reported in the three rightmost columns of Table 5. As before for the balance measures, overall the unfairness measurements of the primary attributes have a positive correlation with the intersectional attribute: specifically, the coefficient $c_{education}$ –which is between 0.430 and 0.622– assumes higher values than the coefficient c_{sex} for all the fairness criteria except for the separation-FP criterion, indicating overall a higher positive correlation in correspondence of the primary multiclass attribute ''education''. Overall –in four cases out of five– there exists a higher positive correlation between the unfairness measurements of the intersectional attribute and those of the primary attribute ''education'', with respect to the correlation between the intersectional attribute and the primary attribute ''sex''. On the basis of this analysis, we can integrate our answer to RQ1 with the following observations on unfairness:

> - There exists a correlation between the unfairness measurements of the intersectional attribute and the primary attributes, but the former is only partly determined by the latter.
> - The unfairness measures related to the intersectional attribute can be explained by those related to the primary attributes, but to a lower extent with respect to the balance measures.

B. RQ 2 - BALANCE AS INTERSECTIONAL UNFAIRNESS PREDICTOR

1) METHOD

In order to answer the second research question, we analyze the relationships between balance measures and fairness criteria for the intersectional multiclass attribute sex_education. We compute the correlation between the balance and the unfairness measure; for each balance measure and each unfairness indicator. We use the Spearman correlation coefficient since we do not expect a linear relationship. A negative and statistically significant correlation coefficient –i.e. lower balance corresponding to higher unfairness– suggests a positive answer to the research question.

However, we remind from our previous studies on primary protected attributes [22], [34] that the balance measures properly detect unfairness of software output, however their effectiveness in identifying unfairness is dependent on the chosen metric, which has a relevant impact on the threshold to consider as risky, and thus on the detection of discriminatory outcomes. As we are investigating the balance measures as unfairness predictors when specifically applied to intersectional attributes, we plot LOESS curve to better understand the relationship between balance and unfairness in the case of intersectional protected attributes.

2) CORRELATION

The correlation coefficients are reported in Table 6. We observe that they are all negative and the corresponding p-values are all smaller than $2.2 \cdot 10^{-16}$. We thus can answer positively to the research question:

> We observe a moderate negative correlation between balance measures and fairness criteria, indicating that intersectional protected attributes can be taken into account to identify unfairness risks.

3) RELATIONSHIP

Fig. 1 reports the trend lines –as smoothed regression– of the five fairness criteria (along the Y-axis) with respect to the increase in balance measures (along the X-axis), in percentage values. It has to be noted that maximum levels of unfairness are higher for Sufficiency (PP and PN) and Separation-TP (more than 10% in correspondence of the lowest values of balance) and less than 5% in the other cases. Overall we observe decreasing trends, in accordance with negative correlation values, however often *not* monotonic, which explains why correlation values were not high. In general, the trends are consistent with our previous studies on primary protected attributes, with most irregular patterns related to Sufficiency. Since the specific unfairness criteria reflect different levels of balance in slightly different ways, we recommend choosing distinct thresholds of risks for the four balance measures: the specific application context might suggest using more sensitive balance measures – IR and Simpson – for cases where unfairness tolerance is low, and the less sensitive Gini and Shannon when higher levels of unfairness can be socially accepted. We complete our answer to RQ2 as follows:

> The behavior of the fairness criteria in response to the balance measures results to have a decreasing trend, even though the distinct fairness criteria reflect different levels of balance in slightly different ways.

C. RQ 3 - CONTRIBUTION OF COMBINED TARGET

1) METHOD

Before looking into the contribution of the target variable combined with protected attributes to the detection of

TABLE 6. Correlation between balance and unfairness for the intersectionl attribute *sex_education*: $\mathfrak{B}$(sex_education) ∼ $\mathfrak{U}$(sex_education).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.1614	-0.1652	-0.1687	-0.1706
Separation – TP	-0.2583	-0.2799	-0.2696	-0.2902
Separation – FP	-0.2130	-0.2244	-0.2203	-0.2340
Sufficiency – PP	-0.1836	-0.1842	-0.1905	-0.1862
Sufficiency – PN	-0.1487	-0.1654	-0.1425	-0.1631

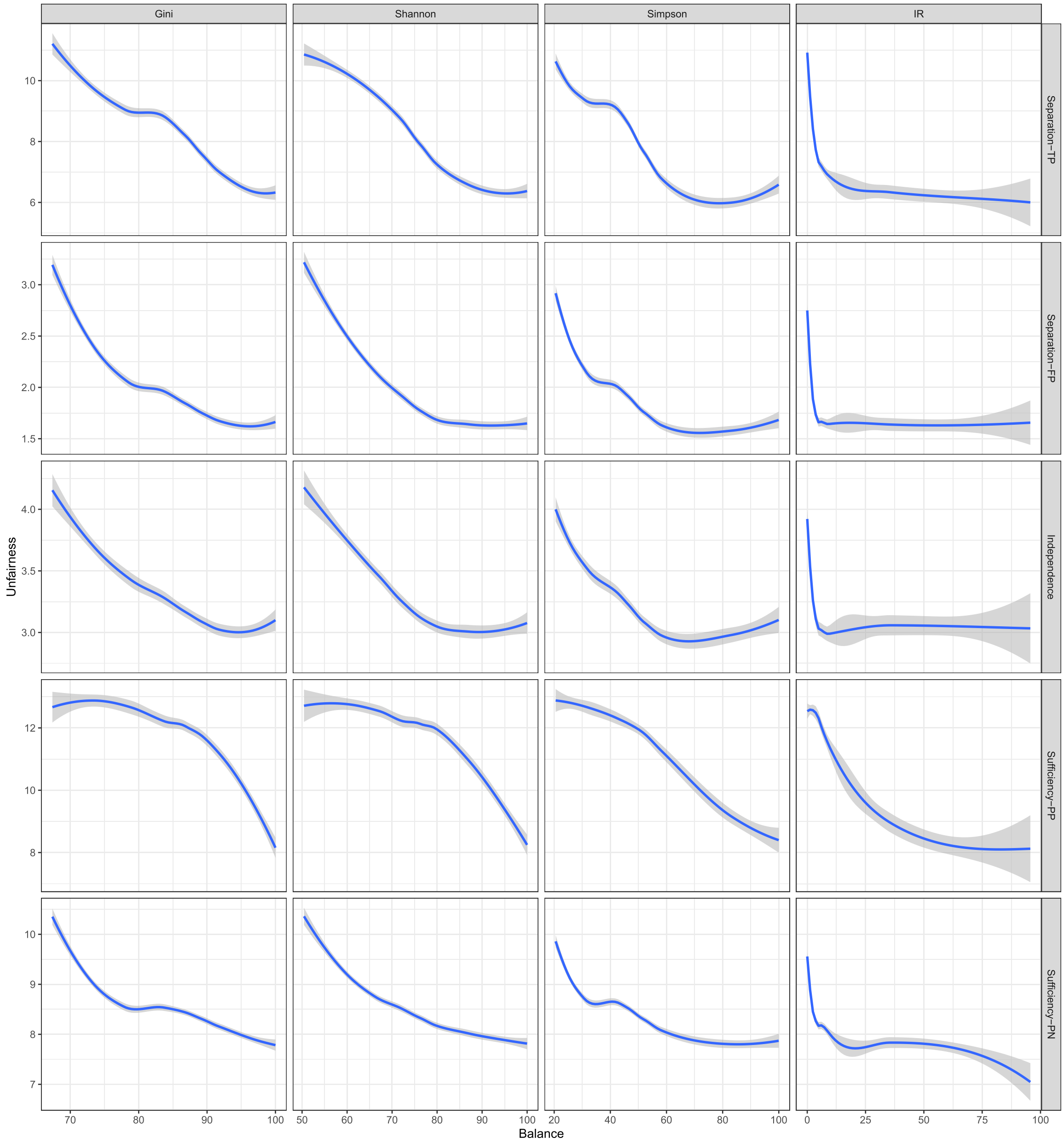


FIGURE 1. Trends of the *fairness criteria* as a response to the *balance measures* for the intersectional protected attribute *sex_education*.

unfairness, we consider the relationship between the balance values of the protected attributes (primary or intersectional) by themselves and when considered in combination with the target variable.

To answer the third research question, we computed the Spearman correlation coefficients of unfairness measures vs. balance measures, and compared the coefficients for the attributes with and without the combination with the

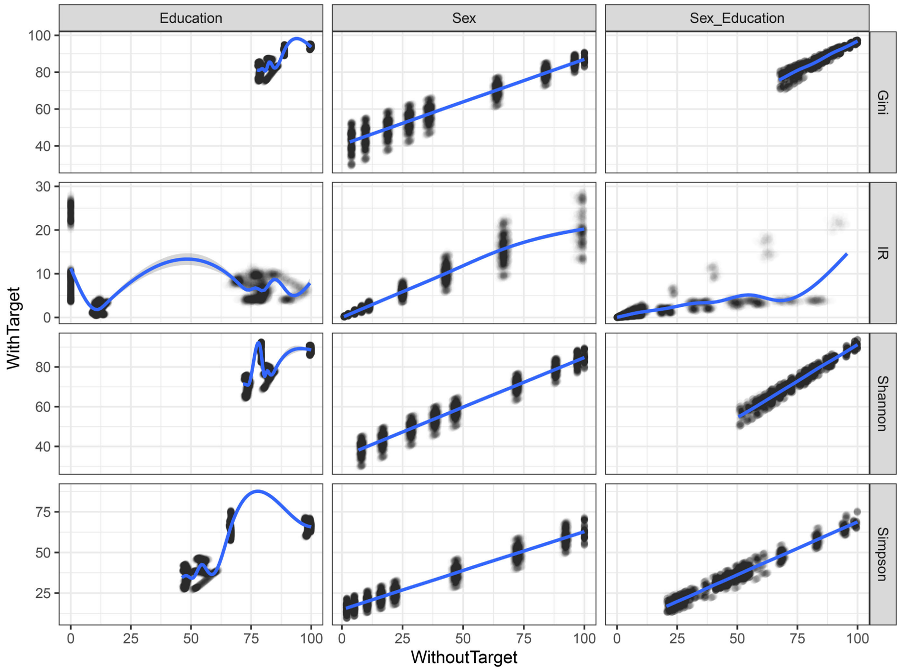


FIGURE 2. Balance measures of protected attributes combined *with target* vs. protected attributes *without target*.

target variable, with a view to investigating whether the combination of a protected attribute with the target variable improves the detection of the unfairness. Then, to examine in-detail our findings, we computed the difference between the correlation of a protected attribute (primary or intersectional) and the correlation of the same attribute combined with the target, for three different cases:

- $\text{diff}_{sex} = \text{cor}(sex) - \text{cor}(sex\_target)$
- $\text{diff}_{education} = \text{cor}(education) - \text{cor}(education\_target)$
- $\text{diff}_{sex\_education} =$
  $= \text{cor}(sex\_education) - \text{cor}(sex\_education\_target)$

where the expression "cor(*protected attribute*)" indicates the correlation between balance and unfairness measures for a given *protected attribute* (primary or intersectional). We remind that we expect the correlations to be negative, which would mean that high balance is associated with low unfairness values, and vice-versa.

2) COMBINATION WITH TARGET VARIABLE

Fig. 2 reports the scatter plot of the corresponding values with a smoothed interpolation curve. We can observe very different patterns. The "sex" primary attribute shows a relationship to its combination with the target that is close to linear for all the balance measures. As far as the "education" attribute is concerned, we observe an irregular relationship that changes among the different balance measures. The intersectional attribute encompassing both the former attributes exhibits a close to linear relationship for three balance measures except for the IR index, which presents a different more irregular pattern.

3) DIFFERENCES IN CORRELATION

We report all the numerical results in the Appendix, while here we provide only a synthetic and more readable overview in Fig. 3, where we report the correlation values for all combinations of balance and unfairness measures divided by attribute. The diagram can be interpreted as follows: the farther left to the zero (represented by the dashed black line) the points, the better they are; if the circle marker is left of the triangle then the combination with the target variable improves the correlation, i.e. the capability of detecting unfairness risk.

As concerns the binary attribute "sex", we observe that all the correlation values are negative (the data points are to the left of the dashed line), but we note a small improvement of the correlation only for the sufficiency criterion –both for Parity of Positives and Parity of Negatives–, whereas we observe a worsening in correspondence of the independence and separation criteria.

A similar pattern can be observed for the multiclass attribute "education", although with much larger differences. In particular, the deterioration of the independence and the separation criteria is so significant that the combination with the target variable makes all correlations positive, indicating that the combination of the target with multiclass protected attributes worsens the unfairness detection even more than binary attributes. Conversely, in the case of the sufficiency criterion, the combined attributes improve the identification of the unfairness to a greater extent with respect to the previous case of the binary attribute.

Finally, for the intersectional protected attribute "sex _education" we notice the same strengthening/weakening

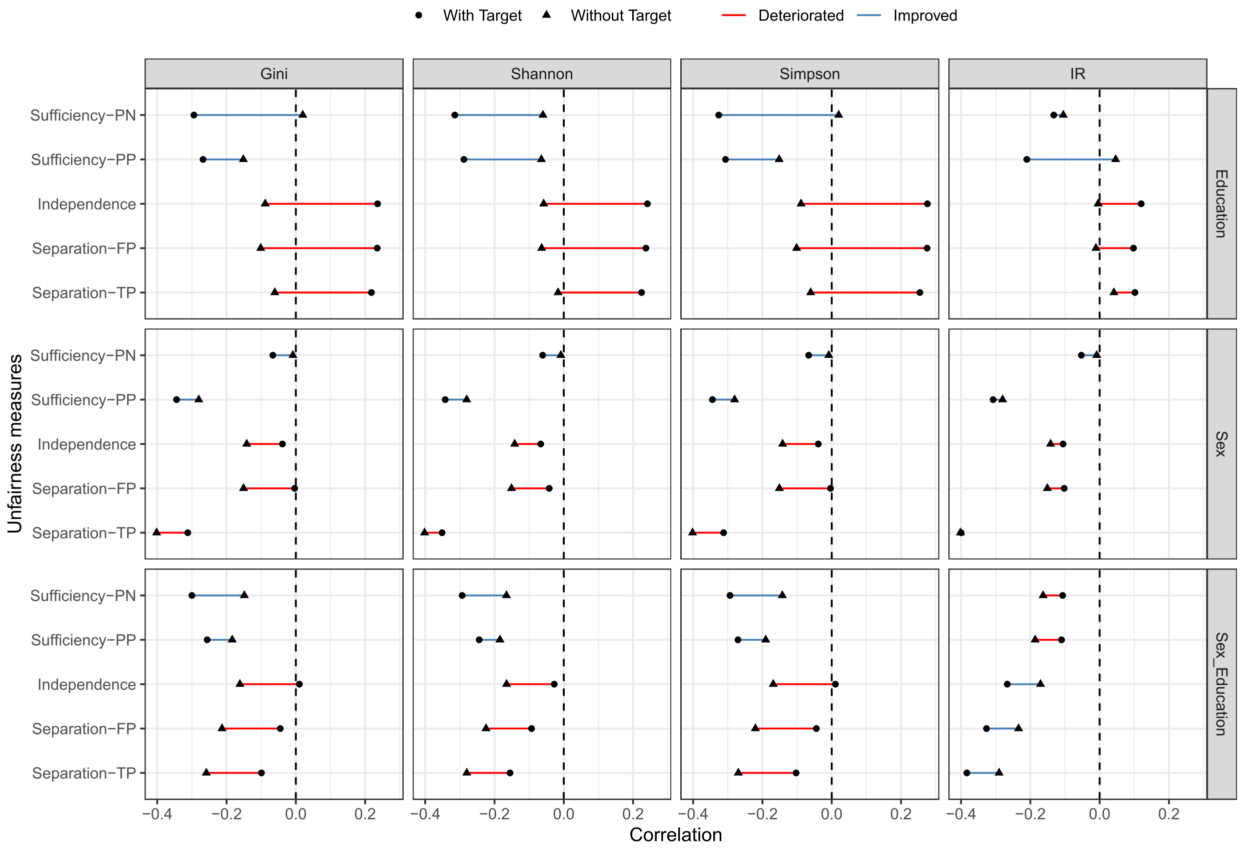


FIGURE 3. Correlation Balance-Unfairness for protected attributes *with* target and *without* target, for different fairness criteria and balance measures. The farther left to the zero the points, the better they are; if the circle marker is left of the triangle, then the combination *with* the target variable improves the unfairness detection.

pattern, but with the notable exception of the Imbalance Ratio index for which we find exactly the opposite pattern for all the fairness criteria. Hence, combining the target variable with the intersectional protected attribute improves the identification of the unfairness assessed through the sufficiency criterion with respect to the Gini, Shannon and Simpson indexes, but not to the IR index –which by contrast worsens in correspondence of the sufficiency criterion, and improves according to the independence and separation criteria.

As a further observation, we note that given the nature of the fairness criteria –whose definitions are based on the target variable, score and protected attributes– and the two mutation techniques that we applied –which leave the distribution of the other variables unchanged, and thus also the distribution of the target variable remains unchanged– the level of balance of the target variable in the original dataset certainly plays a role in the final interpretation of our results. Indeed, in our datasets where the frequency of the positive target is much lower than the negative target (in the original dataset, only 6636 out of 30.000 (22%) belong to the positive class), the combination of protected attributes with the target variable improves identifying a discrimination risk when we apply the sufficiency criterion (except than in correspondence to the Imbalance Ratio index applied to intersectional protected attributes), which in fact implies the calibration of the model for the different groups as it requires the conditional probability of the target variable to be equal to 1. In conclusion, our answer to RQ3 is the following:

- The combination of primary protected attributes (binary or multiclass) with the target variable improves the detection of the unfairness measured through the sufficiency criterion (both Parity of Positives and Parity of Negatives), but worsens the detection of the unfairness measured through the independence or the separation criteria.
- The combination of intersectional protected attributes with the target variable improves the identification of the unfairness measured through the sufficiency criterion in the cases of the Gini, Shannon and Simpson indexes, but not in the case of the Imbalance Ratio index, for which the detection of the unfairness is improved when measured through the independence or the separation criteria.

D. THREATS TO VALIDITY AND LIMITATIONS

As concerns the limitations of our approach, we highlight that more indexes of balance are necessary to generalize the findings of this study, also by including measures for non-categorical data. A further extension regards the analysis of more datasets and several protected attributes, which would extend the study to a wider range of intersectional classes and combinations with target variables belonging to different domains of interest in the large landscape of automated decision-making systems. Moreover, we remark that in our dataset the predicted class was not present, therefore we ran a binomial logistic regression in order to build a classification label, but all the limitations of the algorithm hold: most notably, the assumption of limited or no multi-collinearity between independent variables, as well as the assumption of linearity between the dependent variable and the independent variables. Applying more classification algorithms (each with different parameters) would increase the generalizability of the results, by helping to identify how the different types of classification algorithms propagate the imbalance from the training set to the output. In addition, other kinds of mutation techniques could be considered by adopting different pre-processing methods in order to extend the variability and reliability of our results. Finally, a more in-depth analysis is also necessary to better understand the relation between the level of imbalance of the target variable in the dataset and the application of a determined fairness criterion, for instance by applying a mutation technique to the target variable and examining the behavior of the fairness criteria as a response to the balance measures, in order to better understand which fairness criteria to choose to evaluate the risk of discrimination –whose choice should depend on the domain and context of use as well.

V. CONCLUSION AND PRACTICAL IMPLICATIONS

In this paper, we studied to which extent it is possible to rely on balance measures as risk indicators of systematic discrimination when dealing with the intersection of protected attributes or with combinations of the target variable with protected attributes. We conducted an empirical study to test whether: i) it is possible to infer the balance of intersectional attributes from the balance of the primary attributes, ii) measures of balance on intersectional attributes are helpful to detect unfairness in the classification outcome, iii) the computation of balance on the combination of the target variable with protected attributes improves the detection of unfairness. To answer our research questions, we selected four indexes of balance (Gini, Simpson, Shannon, Imbalance Ratio), we generated a large number of synthetic datasets and measured different levels of imbalance in the training sets, whereas we evaluated the discrimination occurring in the classification outcome on the test sets.

Overall, the results on intersectional attributes show that balance measures are suitable for identifying unfairness risks in a classification output. Due to some variance in the observed trends, we strongly recommend first selecting a single fairness criterion of interest, and then choosing the balance measure that is more appropriate to the application case: Simpson or IR for rapid reaction to unfairness in all cases where small deviations of fairness correspond to severe damages to people's lives, and Gini/Shannon in all other cases because they have a more smooth response to unfairness risks and the cost of false detections of unfairness can be minimized. As far as the combination of protected attributes with a target variable is concerned, we recommend using the Gini, Shannon or Simpson indexes when the sufficiency criterion of fairness is preferable, otherwise the IR index.

However, further work is needed in order to improve the generalizability of our findings, for instance by exploring more metrics of balance, also by including measures for non-categorical data, and extending the study to more datasets and protected attributes, which would increase the number of possible intersectional attributes and combination with target variables. In addition, more classification algorithms and mutation techniques would extend the variability and reliability of our results. A more thorough analysis of the relation between the level of imbalance of the target variable and the application of a specific fairness criterion is necessary to better understand which fairness criteria to choose in order to evaluate the risk of discrimination.

We conclude our paper by remarking that we look at data imbalance as a risk factor and not as a technical fix, to create space for active human considerations and interventions, thus entrusting the ultimate responsibility to human decisions: we strongly recommend keeping in mind this important premise when applying our approach to real cases or further scientific experimentation. We also suggest getting the assistance of domain experts, professionals from the human and social science, and impacted stakeholders when selecting which combination of balance and fairness measures is more appropriate to the case at hand, to fully reflect the socio-technical nature of the bias problem in information systems.

APPENDIX
CORRELATION TABLES

As a complement to the discussion of the third research question in Subsection IV-C, in this Appendix we report correlations between balance and unfairness measures for both protected attributes combined with the target variable and protected attributes without target (in Tables 6,7,8,10,11,13). We also report the differences between the aforementioned correlations, for the protected attributes sex, education and sex_education (in Tables 9,12,14). For the sake of better interpretability of the numerical values in the tables, we make the following specification: as we expect the correlation between balance measures and fairness criteria to be negative for a given protected attribute, we assess the difference between the correlation of a protected attribute (primary or intersectional) and the correlation of the same attribute combined with the target. If this difference is positive,

TABLE 7. Correlation between balance and unfairness measures for the primary attribute *sex*: $\mathfrak{B}$(sex) $\sim$ $\mathfrak{U}$(sex).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.1417	-0.1417	-0.1417	-0.1416
Separation – TP	-0.4017	-0.4018	-0.4017	-0.4017
Separation – FP	-0.1509	-0.1510	-0.1509	-0.1509
Sufficiency – PP	-0.2801	-0.2802	-0.2802	-0.2801
Sufficiency – PN	-0.0085	-0.0085	-0.0084	-0.0084

TABLE 8. Correlation between balance and unfairness measures for the attribute *sex_target*: $\mathfrak{B}$(sex_target) $\sim$ $\mathfrak{U}$(sex).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.0388	-0.0665	-0.0388	-0.1056
Separation – TP	-0.3119	-0.3513	-0.3119	-0.3999
Separation – FP	-0.0035	-0.0420	-0.0035	-0.1024
Sufficiency – PP	-0.3443	-0.3422	-0.3443	-0.3075
Sufficiency – PN	-0.0664	-0.0609	-0.0664	-0.0527

TABLE 9. Difference between the correlation tables 7 and 8: diff_{sex} = cor(*sex*) – cor(*sex_target*).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.1029	-0.0752	-0.1029	-0.0360
Separation – TP	-0.0898	-0.0504	-0.0898	-0.0018
Separation – FP	-0.1474	-0.1089	-0.1474	-0.0485
Sufficiency – PP	0.0641	0.0620	0.0641	0.0274
Sufficiency – PN	0.0579	0.0525	0.0580	0.0444

TABLE 10. Correlation between balance and unfairness measures for the primary attribute *education*: $\mathfrak{B}$(education) $\sim$ $\mathfrak{U}$(education).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.0885	-0.0581	-0.0885	-0.0044
Separation – TP	-0.0608	-0.0165	-0.0608	0.0411
Separation – FP	-0.1013	-0.0637	-0.1014	-0.0106
Sufficiency – PP	-0.1515	-0.0647	-0.1516	0.0462
Sufficiency – PN	0.0200	-0.0601	0.0201	-0.1049

TABLE 11. Correlation between balance and unfairness measures for the attribute *education_target*: $\mathfrak{B}$(education_target) $\sim$ $\mathfrak{U}$(education).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	0.2358	0.2413	0.2762	0.1199
Separation – TP	0.2181	0.2244	0.2544	0.1018
Separation – FP	0.2348	0.2370	0.2753	0.0976
Sufficiency – PP	-0.2679	-0.2882	-0.3065	-0.2103
Sufficiency – PN	-0.2941	-0.3145	-0.3260	-0.1325

TABLE 12. Difference between the correlation tables 10 and 11: $\text{diff}_{education}$ = cor(*education*) – cor(*education_target*).

Balance Measures Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.3243	-0.2994	-0.3647	-0.1243
Separation – TP	-0.2789	-0.2409	-0.3152	-0.0607
Separation – FP	-0.3361	-0.3008	-0.3767	-0.1082
Sufficiency – PP	0.1164	0.2235	0.1549	0.2565
Sufficiency – PN	0.3141	0.2544	0.3461	0.0276

it means that the correlation between balance measures and fairness criteria for that protected attribute combined *with the target variable* is stronger than the correlation obtained *without* combining the protected attribute with the target variable, thus adding the target will improve the unfairness detection. On the contrary, if the difference is negative, the combination of the protected attribute with the target variable does *not* improve identifying the unfairness.

TABLE 13. Correlation between balance and unfairness measures for the attribute *sex_education_target*: $\mathfrak{B}$(sex_education_target) $\sim$ $\mathfrak{U}$(sex_education).

Balance Measures / Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	0.0103	-0.0275	0.0107	-0.2666
Separation – TP	-0.0993	-0.1552	-0.1028	-0.3832
Separation – FP	-0.0450	-0.0930	-0.0443	-0.3266
Sufficiency – PP	-0.2559	-0.2441	-0.2706	-0.1100
Sufficiency – PN	-0.3000	-0.2934	-0.2935	-0.1069

TABLE 14. Difference between the correlation tables 6 and 13: diff$_{sex\_education}$ = cor(*sex_education*) – cor(*sex_education_target*).

Balance Measures / Fairness criteria	Gini	Shannon	Simpson	Imbalance Ratio
Independence	-0.1717	-0.1376	-0.1794	0.0960
Separation – TP	-0.1590	-0.1247	-0.1667	0.0930
Separation – FP	-0.1680	-0.1314	-0.1760	0.0926
Sufficiency – PP	0.0723	0.0600	0.0800	-0.0762
Sufficiency – PN	0.1513	0.1279	0.1510	-0.0562

REFERENCES

[1] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Trans. Inf. Syst.*, vol. 14, no. 3, pp. 330–347, Jul. 1996.

[2] S. Barocas and A. D. Selbst, "Big data's disparate impact," Rochester, NY, USA, Tech. Rep., 2016. [Online]. Available: https://papers.ssrn.com/abstract=2477899

[3] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Broadway Books, Sep. 2017.

[4] *Ethics Guidelines for Trustworthy AI*, Eur. Commission Directorate-Gen. Commun. Netw., Content Technol., Publications Office, Brussels, Belgium, 2019.

[5] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall. (Mar. 15, 2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464

[6] *Information Technology—Artificial Intelligence (AI)—Bias in AI Systems and AI Aided Decision Making*, Standard ISO/IEC TR 24027:2021, 2021. [Online]. Available: https://www.iso.org/standard/77607.html

[7] A. Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (Information Technology and Law Series), vol. 36. Berlin, Germany: Asser Press, 2022.

[8] F. A. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim, "Artificial intelligence & human rights: Opportunities & risks," Berkman Klein Center Res. Publication, Cambridge, MA, USA, Tech. Rep., 2018.

[9] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019. [Online]. Available: https://www.nature.com/articles/s42256-019-0088-2

[10] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, "Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, pp. 1–30, Nov. 2019.

[11] Z. Obermeyer and S. Mullainathan, "Dissecting racial bias in an algorithm that guides health decisions for 70 million people," in *Proc. Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jan. 2019, p. 89.

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

[13] A. Fabris, A. Mishler, S. Gottardi, M. Carletti, M. Daicampi, G. A. Susto, and G. Silvello, "Algorithmic audit of Italian car insurance: Evidence of unfairness in access and pricing," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 458–468, doi: 10.1145/3461702.3462569.

[14] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? How? What to do?" in *Proc. 29th ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng.*, Aug. 2021, pp. 429–440, doi: 10.1145/3468264.3468537.

[15] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[16] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[17] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[18] E. Strickland, C. Q. Choi, S. K. Moore, and P. Patel, "DALL-E2's failures reveal the limits of AI—OpenAI's text-to-image generator struggles with text, science, and bias," *IEEE Spectr.*, vol. 59, no. 8, pp. 5–12, Aug. 2022.

[19] J. Dastin. (Oct. 2018). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. [Online]. Available: https://reut.rs/2Od9fPr

[20] E. Gómez, L. Boratto, and M. Salamó, "Provider fairness across continents in collaborative recommender systems," *Inf. Process. Manage.*, vol. 59, no. 1, Jan. 2022, Art. no. 102719. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645732100203X

[21] A. Vetrò, M. Torchiano, and M. Mecati, "A data quality approach to the identification of discrimination risk in automated decision making systems," *Government Inf. Quart.*, vol. 38, no. 4, Oct. 2021, Art. no. 101619.

[22] M. Mecati, A. Vetrò, and M. Torchiano, "Detecting risk of biased output with balance measures," *J. Data Inf. Qual.*, vol. 14, no. 4, pp. 1–7, Nov. 2022, doi: 10.1145/3530787.

[23] L. Bowleg, "When black + lesbian + woman ≠ black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research," *Sex Roles*, vol. 59, nos. 5–6, pp. 312–325, Sep. 2008.

[24] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, Nov. 2016.

[25] D. Firmani, L. Tanca, and R. Torlone, "Ethical dimensions for data quality," *J. Data Inf. Qual.*, vol. 12, no. 1, pp. 1–5, Mar. 2020.

[26] B. Hutchinson and M. Mitchell, "50 years of test (un)fairness: Lessons for machine learning," in *Proc. Conf. Fairness, Accountability, Transparency (FAT)*, Jan. 2019, pp. 49–58.

[27] E. Pitoura, "Social-minded measures of data quality: Fairness, diversity, and lack of bias," *J. Data Inf. Qual.*, vol. 12, no. 3, pp. 12:1–12:8, Jul. 2020.

[28] K. S. Chmielinski, S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, and Y. C. Qiu, "The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence," 2022, *arXiv:2201.03954*.

[29] M. S. A. Lee and J. Singh, "The landscape and gaps in open source fairness toolkits," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–13, doi: 10.1145/3411764.3445261.

[30] K. Peng, J. Chakraborty, and T. Menzies, "FairMask: Better fairness via model-based rebalancing of protected attributes," *IEEE Trans. Softw. Eng.*, early access, Nov. 15, 2022, doi: 10.1109/TSE.2022.3220713.

[31] A. Fabris, S. Messina, G. Silvello, and G. A. Susto, "Algorithmic fairness datasets: The story so far," *Data Mining Knowl. Discovery*, vol. 36, no. 6, pp. 2074–2152, Sep. 2022. [Online]. Available: https://link.springer.com/10.1007/s10618-022-00854-z

[32] F. Königstorfer and S. Thalmann, "Software documentation is not enough! Requirements for the documentation of AI," *Digital Policy, Regulation Governance*, vol. 23, no. 5, pp. 475–488, 2021.

[33] A. Vetrò, "Imbalanced data as risk factor of discriminating automated decisions: A measurement-based approach," *J. Intellectual Property, Inf. Technol. E-Commerce Law*, vol. 12, no. 4, pp. 272–288, 2021. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0009-29-54528

[34] M. Mecati, A. Vetrò, and M. Torchiano, "Detecting discrimination risk in automated decision-making systems with balance measures on input data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 4287–4296.

[35] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]," in *Feminist Legal Theory*. Evanston, IL, USA: Routledge, 2018, pp. 57–80.

[36] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Accountability Transparency*, 2018, pp. 77–91.

[37] D. Holman, S. Salway, and A. Bell, "Mapping intersectional inequalities in biomarkers of healthy ageing and chronic disease in older English adults," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Aug. 2020.

[38] S. Subramanian, X. Han, T. Baldwin, T. Cohn, and L. Frermann, "Evaluating debiasing techniques for intersectional biases," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2492–2498. [Online]. Available: https://aclanthology.org/2021.emnlp-main.193

[39] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *Proc. IEEE 36th Int. Conf. Data Eng. (ICDE)*, Apr. 2020, pp. 1918–1921.

[40] K. Yang, J. R. Loftus, and J. Stoyanovich, "Causal intersectionality and fair ranking," in *Proc. 2nd Symp. Found. Responsible Comput. (FORC)*, in Leibniz International Proceedings in Informatics, vol. 192, K. Ligett and S. Gupta, Eds. Dagstuhl, Germany: Schloss Dagstuhl—Leibniz-Zentrum für Informatik, 2021, pp. 7:1–7:20. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2021/13875

[41] P. Branco, L. Torgo, and R. P. Ribeiro, "Pre-processing approaches for imbalanced distributions in regression," *Neurocomputing*, vol. 343, pp. 76–99, May 2019.

[42] N. Moniz, P. Branco, and L. Torgo, "Resampling strategies for imbalanced time series forecasting," *Int. J. Data Sci. Anal.*, vol. 3, no. 3, pp. 161–181, May 2017.

[43] X. Yu, J. Liu, Z. Yang, X. Jia, Q. Ling, and S. Ye, "Learning from imbalanced data for predicting the number of software defects," in *Proc. IEEE 28th Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2017, pp. 78–89.

[44] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020.

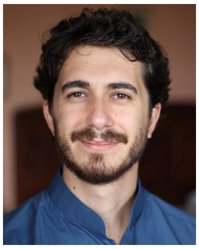
[45] EU Agency for Fundamental Rights. (Dec. 2007). *EU Charter of Fundamental Rights—Article 21—Non-Discrimination*. [Online]. Available: https://fra.europa.eu/en/eu-charter/article/21-non-discrimination

[46] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining Knowl. Discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[47] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. 2019. [Online]. Available: http://www.fairmlbook.org

MARIACHIARA MECATI (Member, IEEE) received the master's degree in mathematical engineering with a specialization in statistics and optimization on data and networks. She is currently pursuing the Ph.D. degree in computer and control engineering with Politecnico di Torino, Italy. She spent three months at the Department of Mathematics, University of Houston, TX, USA, to develop her master's thesis on neural networks applied to retina fundus images. Her research interests include data science with particular attention to the novel field of data ethics, aiming to investigate algorithmic fairness and, above all, the impact of poor data quality and biased data—specifically in terms of imbalanced data—when used by software applications to make automatic decisions affecting individuals.

MARCO TORCHIANO (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from Politecnico di Torino. He was a Postdoctoral Research Fellow with the Norwegian University of Science and Technology (NTNU), Norway. He was a Visiting Professor with Polytechnique Montréal, studying software energy consumption. He is currently an Associate Professor with the Control and Computer Engineering Department, Politecnico di Torino, Italy. He is the author or coauthor of over 150 research papers published in international journals and conferences and the book titled *Software Development: Case Studies in Java* (Addison-Wesley) and a Co-Editor of the book *Developing Services for the Wireless Internet* (Springer). His current research interests include green software, UI testing methods, open-data quality, and software modeling notations. The methodological approach he adopts is that of empirical software engineering. He is a member of the software engineering committee of UNINFO (part of ISO/IEC JTC 1) and a Faculty Fellow of the Nexa Center on Internet and Society.



ANTONIO VETRÒ (Member, IEEE) is currently an Assistant Professor with the Control and Computer Engineering Department, Politecnico di Torino. He is also a Senior Research Fellow with the Nexa Center for Internet and Society. In the past, he has been a Postdoctoral Research Fellow with the Software and System Engineering Department, Technische Universität München, Germany, and a Junior Scientist with the Fraunhofer Center for Experimental Software Engineering, MD, USA. He is currently studying how to detect and mitigate discrimination caused by automated decision systems. He has vast experience in empirically evaluating methods for software and data quality improvement. He is a member of the international committee ISO/IEC JTC1 SC7/WG6 software and system engineering—software product and system quality and the Italian Software Engineering Committee at UNINFO (delegated with ISO for standardization in information technologies). He serves on the scientific board of the Italian Institute for Advanced Studies in Torino "Umberto Eco" Scienza Nuova.

JUAN CARLOS DE MARTIN (Member, IEEE) was a Faculty Associate with the Berkman Klein Center for Internet and Society, Harvard University. He was a Co-Curator of Biennale Tecnologia. He is currently a Full Professor of computer engineering with the Polytechnic of Turin, the vice-rector for culture and communication, and a Co-Director of the NEXA Center for Internet and Society. His research interest includes the interplay between digital technologies and democracy, with a specific focus on digital power. Previously, he worked on the future of the university in the Internet age, a topic which he addressed in the book *Università Futura—Tra Democrazia e Bit* (in Italian, Codice Edizioni, 2017).

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement