Identifying Imbalance Thresholds in Input Data to Achieve Desired Levels of Algorithmic Fairness

Mariachiara Mecati
Politecnico di Torino
Torino, Italy
mariachiara.mecati@polito.it

Andrea Adrignola
Politecnico di Torino
Torino, Italy
andrea.adrignola@studenti.polito.it

Antonio Vetrò
Politecnico di Torino
Torino, Italy
antonio.vetro@polito.it

Marco Torchiano
Politecnico di Torino
Torino, Italy
marco.torchiano@polito.it

***Abstract*—Software bias has emerged as a relevant issue in the latest years, in conjunction with the increasing adoption of software automation in a variety of organizational and production processes of our society, and especially in decision-making.**
Among the causes of software bias, data imbalance is one of the most significant issues. In this paper, we treat imbalance in datasets as a risk factor for software bias. Specifically, we define a methodology to identify thresholds for balance measures as meaningful risk indicators of unfair classification output. We apply the methodology to a large number of data mutations with different classification tasks and tested all possible combinations of balance-unfairness-algorithm.
The results show that on average the thresholds can accurately identify the risk of unfair output. In certain cases they even tend to overestimate the risk: although such behavior could be instrumental to a prudential approach towards software discrimination, further work will be devoted to better assess the reliability of the thresholds.
The proposed methodology is generic and it can be applied to different datasets, algorithms, and context-specific thresholds.

***Index Terms*—Data bias, Data imbalance, Algorithmic fairness, Risk analysis, Automated decision-making**

I. Introduction

Automated decision-making (ADM) systems are deployed in a large (and continuously increasing) number of sectors of our society, both in private and public organizations: automated decisions have an impact on important aspects of people's lives, such as in recruiting, education, finance, justice, just to mention a few cases [1]. In this context, a relevant socio-technical issue that emerged in recent years is the problem of biased software [2], i.e. software that *"systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate"* [3]. Often this problem is caused by imbalanced data, i.e. the condition of uneven distribution of data among the classes of a given attribute, which causes highly heterogeneous accuracy across different classes [4]. The problem of bias is a socio-technical issue because it often occurs when the target of predictions/classifications are people, and consequently, there is a disparate impact on specific social groups. We define a social group as a group of individuals who share the same physical, cultural or identitarian characteristics. When such characteristics are recorded in datasets, those groups correspond to persons sharing the same value of a given protected attribute. Being the problem of software bias a problem of automated discrimination, we identify as protected attributes those listed in "Article 21 - Non- discrimination" of the EU Charter of Fundamental Rights [5]: "*Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.*"

In previous works we successfully tested the reliability of balance measures as risk indicators [6] [7] [8]. In this paper, we move the investigation forward and we define a methodology to identify which thresholds of balance (measured on protected attributes in the training set) should be used for detecting a defined level of algorithmic fairness.

We show the methodology in section II and we describe the main elements of the experiment design in section III. Results are discussed in section IV, while we position our work in the literature and relate it with our previous studies in section V. Finally, we examine the limitations of our work in section VI, and we summarize the main aspect of the study in section VII, also highlighting potential future work.

II. Methodology for identifying Risk Thresholds

With a view to building risk thresholds for balance and unfairness measures, the methodology is aimed at answering the following *Research Question*:

> Is it possible to identify a threshold s (for balance measures) such that if the *balance* of the training set is greater than s, then the *unfairness* of the classification on the test set is expected to be less than a threshold f?

Balance is measured on protected attributes of the training set, while unfairness is computed on the test set.

We defined and applied the following procedure –separately for the binary and multiclass cases:

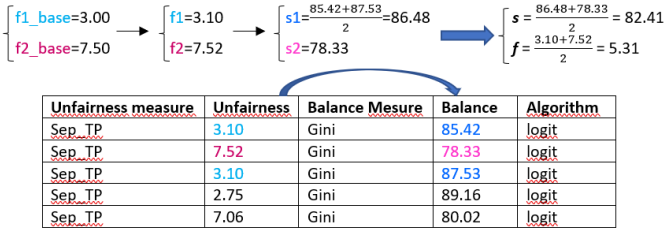
978-1-6654-8045-1/22/$31.00 ©2022 IEEE

1. we collected seven different datasets and, for each dataset, we selected both a binary and a multiclass protected attribute;
2. using two specific mutation techniques (one for the binary case and one for the multiclass case) we generated a large number of synthetic datasets with different levels of balance;
3. we assumed four classification algorithms, then, for each algorithm and for each synthetic dataset, we performed a classification with training-test sets randomly split of 70%-30%;
4. we computed the balance of the protected attributes in the training set through four different widely used *balance measures*;
5. we applied three different *unfairness measures* to the protected attributes in the test set –i.e. to the classifications obtained from the model– for a total of five unfairness measures on each protected attribute;
6. we built the thresholds s (for balance measures) and f (for unfairness measures) using the first collection of data by following the procedure specified below;
7. we generated a new collection of data by repeating steps 2 and 3;
8. using the second collection of data, we assessed and analyzed the performances of the thresholds previously defined through different evaluation metrics.

The method for identifying risk thresholds (step 5) relies on the first collection of data to identify the thresholds f and s: it is necessary to empirically observe the distribution of the unfairness to understand where the unfairness thresholds could be reasonably placed. We built five different configurations[1] in which f is placed differently relative to the distribution of the unfairness and, associating to each f the corresponding thresholds of balance s, we got five potential sets of thresholds; among them, we select the one that presented the highest accuracy.

We followed this procedure for each combination of balance measures, unfairness measures, and algorithms, basically filtering the collection of data with respect to those factors. In figure 1 we report a numerical example of the overall procedure, which is described as follows:

1) we define 2 theoretical values of unfairness thresholds, f1_base and f2_base, which identify the following brackets (where u = unfairness):
$$u \quad f1\_base \quad ! \text{ low unfairness}$$
$$f1\_base < u \quad f2\_base \quad ! \text{ medium unfairness}$$
$$u > f2\_base \quad ! \text{ high unfairness}$$
2) in the first collection of data, we select the values of unfairness that are nearest to f1_base and f2_base, and define them as f1 and f2;
3) as for each unfairness value there exists a corresponding value of balance –and vice versa–, we identify the two values of balance corresponding to f1 and f2, i.e. the values in correspondence of f1 and f2 in the data, and define them as s1 and s2. If more than one balance value is found corresponding to f1 or f2, we take their mean (e.g., if we find 2 values equal to f1, we can find two different values for the corresponding s1, thus we assume as s1 the mean of the two values);
4) we define the threshold of unfairness f as the mean between f1 and f2, and the threshold of balance s as the mean between s1 and s2.



Unfairness measure	Unfairness	Balance Mesure	Balance	Algorithm
Sep_TP	3.10	Gini	85.42	logit
Sep_TP	7.52	Gini	78.33	logit
Sep_TP	3.10	Gini	87.53	logit
Sep_TP	2.75	Gini	89.16	logit
Sep_TP	7.06	Gini	80.02	logit

Fig. 1. Numerical example of the procedure for the identification of the thresholds s and f, for the combination Gini-Sep_TP-logit.

A possible variation of this procedure consists in defining only one value of the unfairness f_base at step 1, instead of two different values. In this case, there is only one value for f and s at step 2-3, and it is possible to define the two thresholds at step 4 without computing a mean. The reason for these choices is to distribute the values of f evenly in the desired range, which is –based on the initial observation of the distribution of the unfairness– where we observed the highest concentration of unfairness values, approximately between the minimum and the mean of the distribution.

Note that for generating each of the two collections of data we varied a `seed` by setting 50 randomly sampled values between 1 and 1000, in order to keep track of both the samples and the mutations for reproducibility purposes, and with a view to increasing the variability –and thus the reliability– of our method.

III. Experimental Design

Hereinafter we describe all the elements of the experiment design.

A. *Data*

We selected datasets belonging to three different domains –financial, social and health– often involved in the discussion about discrimination risk derived from the application of ADM systems because of the potential impact of unfair decisions in such fields that could significantly affect people's lives. We retrieved a total of 7 datasets (one of them has been used in two separate classification tasks) from the UCI machine learning repository, and their relevant features are summarized in Table 1, while the specific predictors and target variables employed in the different classification tasks are provided in Appendix A[2]. These datasets contain

[1]The five configurations with all the specifications on their construction can be found in the Appendix B available at the following link: https://doi.org/10.5281/zenodo.7350599

[2]Appendix A is available at https://doi.org/10.5281/zenodo.7350599

data about individuals: some variables depend on the given context and are used as predictors, such as financial data, others are considered sensitive information, such as gender or age, among which we chose one binary and one multiclass attribute.

TABLE I
SUMMARY OF THE DATASET'S PROMINENT PROPERTIES.

Dataset	Domain	Target variable	Binary Protected attribute	Multiclass Protected attribute
Default of credit cards clients (Dccc)	Financial	default payment next month	sex	education
Statlog	financial	creditworthness	sex	education
Student performance (Math)	social	final grade	sex	father education
Student performance (Portuguese)	social	final grade	sex	father job
Census income	financial	income bracket	sex	race
Drug consumption (cannabis)	social	cannabis consumption	sex	ethnicity
Drug consumption (impulsive)	social	impulsiveness	sex	ethnicity
Heart disease	health	diagnosis	sex	education

***Credit card default**:* this dataset contains informations about default payments of credit card clients in Taiwan from April 2005 to September 2005 [9]. It includes credit data, history of payment, bill statements, together with demographic information. The dataset is composed of 30000 instances with 25 variables, mostly categorical. For the purpose of having results within a reasonable time with limited computing resources, we sampled 30% of the original dataset; even so, it has still a considerable amount of instances (9000), more than most of the datasets considered here.

***Statlog**:* this German credit dataset has been provided by the German professor Hans Hofmann as part of a collection of datasets from a European project called "Statlog" [10]. The data are a stratified sample of 1000 credits (700 good ones and 300 bad ones) and have been collected between 1973 and 1975 from a large regional bank in southern Germany, which had about 500 branches, both urban and rural ones. Bad credits have been heavily over-sampled, in order to acquire sufficient data for discriminating them from good ones. Specifically, the dataset contains 20 categorical attributes: each entry represents a person who takes credit from a bank and is classified as a good or bad credit risk.

***Student performance.**:* this is a set of two datasets containing information on student achievement in secondary education of two Portuguese schools; they have been built by using school reports and questionnaires in 2014 [11]. There are a total of 624 instances, and the attributes include student grades, as well as demographic, social and school-related features. The set of features are the same for the two datasets, including the target variable, which represents the final grade for Math or Portuguese (each dataset is about one of the two subjects); the final grade was divided into two classes by taking 9/20 as a threshold ($<= 9$, > 9).

***Census income.**:* these data were extracted by Barry Becker from the 1994 Census database and is also known as "Census Income" dataset [12]; the associated prediction task is to determine whether a person makes over \$50, 000 a year based on a set of reasonably clean records (the two classes to predict are then high or low income). It counts over 48000 instances and 15 variables: again, to avoid unnecessarily long training time, we took a sample of 30% of the original dataset.

***Drug consumption.**:* it contains records for 1885 respondents; for each of them, personality measurements are known, together with some demographic data [13]. In addition, participants were questioned concerning their use of 18 legal and illegal drugs: for each drug they have to select one of the answers: never used, used over a decade ago, or used in the last decade, year, month, week, or day. There is also one fictitious drug (Semeron) that was introduced to identify over-claimers. All variables are quantified, with fixed values representing specific categories. This dataset has been used for two different classification tasks: first, to predict the consumption of a drug given the personality data, and second, to predict a personality trait given the consumption of drugs. These two versions are considered as two different datasets in the following.

***Heart disease.**:* this dataset describes a range of conditions that could affect the heart [14]. These include blood vessel diseases, such as coronary artery disease, heart rhythm problems and congenital heart defects, as well as others. It consists of 303 instances and 14 variables, mostly clinical data. The original dataset contains 76 variables, but all published experiments refer to the subset considered in our study.

B. Mutation techniques

We adopted two specific *pre-processing* methods as mutation techniques, one for the binary attribute and one for the multiclass attribute, in order to generate synthetic datasets with different levels of balance; note that in both cases 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.

For *binary attributes*, we applied the function `ovun.sample`[3] from the `ROSE-package`. The relevant parameter of this mutations is p, which determines the probability of resampling from the minority class. Thus, we set 9 values for p, ranging from 0.01 (high imbalance) to 0.5 (perfect balance):

p = f0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5g

If we factor in 8 datasets and 50 seeds, we obtain $8 \quad 50 \quad 9 = 3600$ synthetic datasets. Because each dataset is processed by 4 different algorithms, we have a total of $3600 \quad 4 = 14400$ classifications.

[3]https://www.rdocumentation.org/packages/ROSE/versions/0.0-4/topics/ovun.sample, last visited on October 10, 2022

For *multiclass attributes*, we applied the function `SmoteClassif`[4] from the `UBL-package`. In this case, the relevant parameter is *c.perc*, which is a named list containing the percentages of under-sampling or/and over-sampling to apply to each class of the sensitive attributes. We examined five different configurations for the parameter C.perc: first, the default configuration "balance" (namely, the perfect uniform distribution, with all the occurrences equally distributed among the different classes) together with four additional configurations, corresponding to 4 exemplar distributions "Power2", "HalfHigh", "OneOff" and "QuasiBalance".

Power 2: occurrences are distributed according to a power-law with base 2, i.e., distributions among the classes increase like the powers of 2;
Half High: occurrences are distributed mostly among half of the classes while the remaining ones have a very low frequency (in particular, a ratio of 1:9 has been chosen for the frequencies of the two halves);
One Off: occurrences are distributed among all classes but one (which has 0 occurrences);
Quasi Balance: half of the classes are 10% higher w.r.t. max balance and the other half is 10% lower.

In addition, for each exemplar distribution we considered 4 permutations of the percentages assigned to the different classes. For instance, in the *One Off* configurations the four different permutations have each a different class with zero occurrences. If we factor in 8 datasets , and 50 seeds, we obtain 8 50 (4 4 + 1) = 6800 synthetic datasets. Considering that each dataset is processed by 4 different algorithms, we have a total of 6800 4 = 27200 classifications.

C. Algorithms

In our analysis, we used four different algorithms in order to simulate different classification tasks: specifically, to better generalize our study, we aimed at analyzing possible significant differences when establishing the thresholds with respect to the different algorithms, but we were not interested in the specific algorithm performance; for this reason, we did not perform hyper-parameters tuning and we kept the default parameters. The four algorithms are summarized as follows:

Logistic regression (logit): function *glm*, with argument family=binomial(link="logit"), from the package `stat` [5];
Support vector machine (svm): function *svm* from the package `e1071` [6];
Random forest: function *randomForest* from the package `randomForest` [7];
K-nearest neighbors (K-nn): function *knn* from the package `class` [8].

D. Balance measures

In this study, we limited our attention to *categorical* attributes and we selected the same four indexes already analyzed in our previous studies (see table II). The measures have been normalized in order to meet two conditions:

range in the interval $[0, 1]$;
share the same interpretation: the closer the measure to 1 and the higher the balance (i.e. categories have similar frequencies); vice-versa, values closer to 0 means more concentration of frequencies in few categories, thus an imbalanced distribution.

Note that we dealt with empty classes, i.e. classes that exist (potentially there could be occurrences) but are not represented; thus, we took into account *all* the classes of each selected sensitive attribute, including also the classes with zero occurrences. The reason for this choice is that, in our view, a dataset that contains no instances of a given class is imbalanced.

Gini Index: it is a measure of heterogeneity, which reflects how many types of a particular group are represented. It is used in a number of fields, such as political polarization or market competition, and often with different designations. In statistics, the heterogeneity of a discrete random variable which assumes m categories with frequency f_i (with i = 1, ..., m) can vary between a degenerate case (minimum value of heterogeneity) and an equiprobable case (maximum value of heterogeneity, since categories are all equally represented). Thus, for a given number of categories, the heterogeneity increases if the frequencies of the different classes are as equal as possible, i.e. the classes are similarly represented.

Shannon Index: it is a measure of species diversity in a community, which is a widely employed concept in biology, phylogenetics and ecology; it is a useful index to assess the imbalance in the composition of a community by taking into account the relative amounts of different species (classes).

Simpson Index: it is another index of diversity that measures the probability that two individuals randomly selected from a sample belong to the same species, i.e. the same class; it is employed in social and economic sciences for measuring wealth, equity and uniformity, as well as in ecology for measuring the diversity of living beings in a given location.

Imbalance Ratio: it is a widely used measure made of the ratio between the highest and the lowest frequency of the classes, but we take the inverse in order to normalize it in the range $[0, 1]$, analogously to the previous balance measures; it is particularly sensitive to class imbalance, as it always considers only the most represented class and the least represented class, regardless of the number of classes of a given attribute.

[4]https://www.rdocumentation.org/packages/UBL/versions/0.0.6/topics/SmoteClassif, last visited on October 10, 2022

[5]https://www.rdocumentation.org/packages/stats/versions/3.6.2, last visited on October 10, 2022

[6]https://www.rdocumentation.org/packages/e1071/versions/1.7-11, last visited on October 10, 2022

[7]https://www.rdocumentation.org/packages/randomForest/versions/4.7-1.1/topics/randomForest, last visited on October 10, 2022

[8]https://www.rdocumentation.org/packages/class/versions/7.3-20, last visited on October 10, 2022

TABLE II
THE *balance measures* WITH THE RESPECTIVE FORMULA, WHERE WE CONSIDER A DISCRETE RANDOM VARIABLE WITH m CLASSES, EACH WITH FREQUENCY f_i (= PROPORTION OF THE CLASS i W.R.T. THE TOTAL) WHERE $i = 1, ..., m$.

Gini	$G = \frac{m}{m-1} \left(1 - \sum_{i=1}^{m} f_i^2\right)$
Simpson	$D = \frac{1}{m-1} \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$
Shannon	$S = \left(\frac{1}{\ln m}\right) \sum_{i=1}^{m} f_i \ln f_i$
Imbalance Ratio	$IR = \frac{\min(\{f_{1::m}\})}{\max(\{f_{1::m}\})}$

E. *Fairness assessment*

We assessed the *unfairness* of automated classifications relying on three criteria formalized in [15]. Hereinafter we say indistinctly "Unfairness measures" and "Fairness criteria", as we assume the fairness criteria as measures of unfairness in a classification output. Note that we deal with categorical attributes and the measures respect the following conditions:

values are in the range [0, 1];
the higher the fairness of the outcome, the lower the value of the measure (opposite behavior with respect to the balance measures);
if the conditions for the specific criterion are not satisfied, we get an "NA".

In general, to evaluate unfairness we consider a sensitive categorical attribute A that can assume different values $(a_1, a_2, ...)$, a target variable Y and a predicted class R where Y is binary (i.e., $Y = 0$ or $Y = 1$, and thus also the classifier is binary $R = 0$ or $R = 1$). In practice, we aim to check whether the ADM system, which assigned a predicted class, behaved fairly w.r.t. the different values of a sensitive attribute.

Independence*:* this criterion requires the acceptance rate to be the same in all groups, where acceptance corresponds to the event $R = 1$, and it has been explored through many equivalent terms or variants referred to as, for instance, demographic parity or statistical parity, since it enforces groups to have equal selection rates. If A is binary (that is, $A = a_1$ or a_2), then we can compute the Independence unfairness measure as:

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

Separation*:* in simple words, as in many scenarios the sensitive characteristic may be correlated with the target variable, the separation criterion allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable, reason why it is also said equalized odds, equality of opportunity, or even conditional procedure accuracy. Specifically, the separation criterion requires the equivalence of true positive rate and false positive rate for each level of the protected attributed under analysis. If A is binary we can compute two Separation unfairness measures:

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

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

Sufficiency*:* the third criterion implies calibration of the model for the different groups, that is, Parity of Positive predictive values and Parity of Negative predictive values across all groups. As before, if A is binary we can compute two Sufficiency unfairness measures as follows:

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

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

We observe that overall we have three criteria but five conditions in total, i.e. five unfairness measures, and all the definitions above can be extended to the case of non-binary attributes by taking the mean of indexes computed considering all the possible pairs of levels in A:

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

F. *Evaluation Metrics*

We assumed different evaluation metrics to assess the reliability of the thresholds. First, we remind that the first collection of data has been used to build the thresholds, whereas the second has been used to evaluate them: in simple words, we evaluate whether a classification (obtained with the second collection of data) respects or not the conditions on balance and unfairness measures defined through the first collection. Given the two thresholds s and f (for balance measures and unfairness measures respectively), when the balance of the training set is over s, we expect the unfairness of the classification to be under f; if this happens, we have a positive instance, otherwise we have a negative instance. Hence, we define the following instances related to the confusion matrix in figure 2:

if balance<s & unfairness>f ! True Positive (TP)
if balance<s & unfairness<f ! False Positive (FP)
if balance>s & unfairness<f ! True Negative (TN)
if balance>s & unfairness>f ! False Negative (FN)

		Predicted values: Positive	Predicted values: Negative
Actual values	Positive	TP = balance < s & unfairness > f	FN = balance > s & unfairness > f
	Negative	FP = balance < s & unfairness < f	TN = balance > s & unfairness < f

Fig. 2. Confusion matrix of the statistical classification based on the different levels of balance/unfairness.

In particular, we adopted the five sensitivity indexes reported in table III, whose values range in the interval [0, 1]: Accuracy evaluates the percentage of correctly classified values, while Precision (also called Positive Predictive Value) represents the

fraction of positive instances correctly identified with respect to all the positive predicted instances; Sensitivity, also called Recall, indicates how many positive instances are correctly detected (TP) among those that actually present the condition; instead, Specificity represents how many negative instances are correctly identified (TN) among all those that do not present the condition; finally, F1-score is the harmonic mean of precision and sensitivity.

TABLE III
THE *evaluation metrics* WITH THE RESPECTIVE FORMULA.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
F1-score	$\frac{2TP}{2TP+FP+FN}$

IV. RESULTS AND DISCUSSION

Before addressing the main research question, we report a preliminary analysis of the correlation between fairness criteria and balance measures; after that, we show the overall results for thresholds and evaluation metrics, and finally we assess the goodness of our results by aggregating with respect to balance measures, fairness criteria and algorithms, in order to better understand how the different factors affect the goodness of the outcomes.

A. Analysis of the correlation between balance measures and fairness criteria.

Before addressing the *Research Question*, we assessed the correlation between balance and unfairness measures in order to verify whether the negative correlation holds (the higher the balance, the lower the unfairness). Indeed, compared to our previous studies [6] [7] [8], in this analysis we introduced much more datasets and algorithms to classify data, as specified also in section V, so as to increasingly thoroughly assess the balance measures as a risk indicator.
To understand the values, we remind from the results of our previous studies that we expect the correlation between balance measures and fairness criteria to be negative, as we expect the balance measures to be high (indicating low imbalance) if the unfairness values are low (i.e., higher fairness). Thus, the stronger the negative correlation, the stronger the relationship between balance and unfairness measures.

As we can observe from table IV, most of the balance measures present a moderate or low negative correlation with the fairness criteria, above all in the case of binary attributes, meaning that the higher the indexes of balance, the lower the unfairness measures. Note that the computations reveal that such values are all significant, with a p-value$<$0.05.
More in detail, for binary attributes we observe correlation values between -0.063 and -0.407 for all the fairness criteria except for the Independence criterion, which presents no correlation (around 0.008) indicating that this criterion is the most difficult to detect; whereas the Sep_TP and the Suf_PP criteria present the stronger negative correlation values, between -0.341 and -0.407.
Specifically for multiclass attributes, instead, we note a weak negative correlation in correspondence of the Sep_TP, Suf_PP and Suf_PN criteria, between -0.008 and -0.045, and a weak positive correlation in the range 0.012–0.133 for the Independence and the Sep_FP criteria, meaning that overall the level of unfairness in the case of multiclass attributes is more difficult to detect.

In general, we observe that the balance measures respond very similarly to the different fairness criteria, therefore we deduce that the negative correlation depends mostly on the unfairness measures, rather than on the specific balance measure.

B. Assessment of the thresholds through evaluation metrics.

To define the thresholds s (for balance measures) and f (for fairness criteria), for each combination of balance-unfairness-algorithm we selected the configuration that has the highest accuracy. We chose the accuracy as a discriminant for the identification of the thresholds s and f because it showed the smallest interquartile range (or IQR, which graphically corresponds to the height of the box) indicating that the accuracy index is the one with the lowest variability among the selected evaluation metrics (see figure 3 and figure 4). The complete results are reported in Appendix C[9] in separate tables for the binary and the multiclass cases, ordered by balance measure (Gini, Shannon, Simpson, and IR indexes), for each combination of balance-unfairness-algorithm. For sake of legibility, we report values for the thresholds of both fairness criteria and balance measures multiplied by 100, i.e. on a scale $[0, 100]$. Hereinafter, we show the aggregated and overall results for thresholds and evaluation metrics. We remind that the aim of this study is to define two thresholds s (for balance measures) and f (for unfairness measures) such that if the balance of the training set is greater than s, then the unfairness of the classification on the test set is expected to be less than f. As before, we examine results separately for binary and multiclass attributes.

Regarding the binary case, overall the thresholds assume values close to the extremes of the range, with the thresholds for the fairness criteria being between 0 and 10, and the thresholds for the balance measures being between 80 and 100 except for the IR index, which presents lower balance threshold values, around 60 (we can retrieve such data from Appendix C). Looking at figure 3, we observe that the Accuracy is on average 0.7, but among all the evaluation metrics the Precision index is the one that presents the highest values, around 0.85 on average, indicating a high fraction of positive instances correctly identified with respect to all the positive predicted instances. Instead, the Sensitivity –or Recall– is on average around 0.75, meaning that the number

[9]All the tables are reported in the Appendix C available at https://doi.org/10.5281/zenodo.7350599

TABLE IV
CORRELATION BETWEEN BALANCE MEASURES AND UNFAIRNESS MEASURES, SEPARATELY FOR THE BINARY AND THE MULTICLASS CASES.

Fairness criteria \ Balance Measures	Gini (binary)	Shannon (binary)	Simpson (binary)	IR (binary)	Gini (multi)	Shannon (multi)	Simpson (multi)	IR (multi)
Independence	0.008	0.008	0.008	0.006	0.133	0.111	0.100	0.081
Separation – TP	-0.393	-0.396	-0.379	-0.341	-0.028	-0.022	-0.032	-0.008
Separation – FP	-0.073	-0.074	-0.071	-0.063	0.032	0.014	0.012	-0.011
Sufficiency – PP	-0.400	-0.407	-0.382	-0.345	-0.039	-0.036	-0.045	-0.016
Sufficiency – PN	-0.115	-0.116	-0.110	-0.097	-0.019	-0.032	-0.034	-0.017

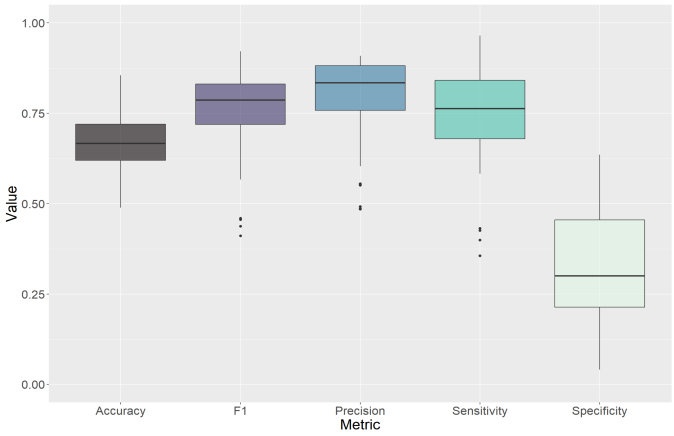


Fig. 3. Boxplot of the evaluation metrics in the binary case.

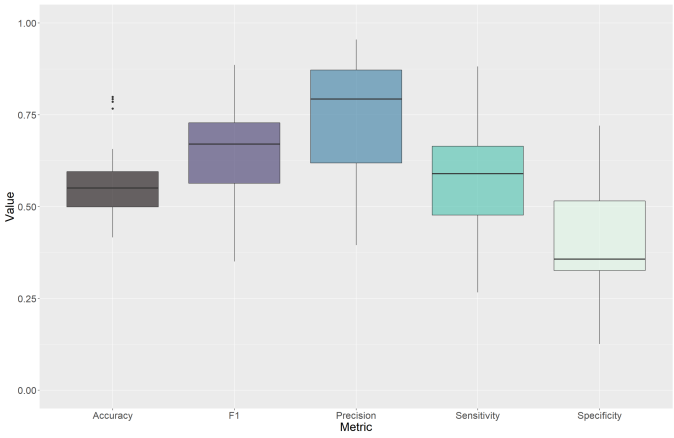


Fig. 4. Boxplot of the evaluation metrics in the multiclass case.

of instances misclassified as negatives (FN) is higher than the number of instances misclassified as positives (FP); in terms of thresholds, it means that the number of instances in which values of balance are over s (indicating high balance) and the unfairness is over f (indicating high unfairness) differently from the expectation to find low unfairness, is *higher* than the number of instances in which values of balance are under s (indicating low balance) and the unfairness is under f (indicating low unfairness) differently from the expectation to find high unfairness. Considering the F1-score, which represents the harmonic mean between Precision and Sensitivity, it assumes values around 0.8 on average. Finally, the worst performances are identified through the Specificity index, with significantly lower values (around 0.3 on average) with respect to the other indexes, indicating that the number of true negative instances is very small with respect to the number of false positives, i.e., when evaluating the thresholds we obtain a high number of instances in which values of balance are under s (indicating low balance) but the unfairness is under f (indicating low unfairness) differently from the expectation to find high unfairness.

As regards the multiclass case, in Appendix C we can observe that overall the thresholds for the balance measures are between 70 and 100 except for the IR index, which presents much lower balance threshold values of around 30, while the thresholds for the fairness criteria are in the range 0 and 15. Looking at figure 4 we note that overall the evaluation metrics assume lower values with respect to the binary case: Accuracy decrease to around 0.55, Precision is around 0.8 and Sensitivity is around 0.6 on average, with F1-score around 0.7; on the contrary, Specificity slightly increase to around 0.35 on average. Thus, according to the evaluation metrics taken into account, the identified thresholds perform better in the binary case than in the multiclass case.

Overall, we deduce that the thresholds are responsive to risk, i.e. values of balance under s indicate levels of unfairness over f, but they even tend to overestimate the risk (as we can infer from the low Specificity caused by the high number of false positives).

C. Assessments of the thresholds' goodness with respect to balance measures, fairness criteria and algorithms.

As we defined the thresholds for each combination of balance-unfairness-algorithm, hereinafter we assess the thresholds' accuracy with respect to the different balance measures, fairness criteria and algorithms involved in the study, in order to understand how and to which extent each factor affects the performances of the thresholds.

Concerning the binary case, looking at figure 5 we observe that the thresholds' accuracy with respect to the four balance measures is around 0.67 overall, with the IR index slightly higher than the other measures, but with no significant differences between the indexes.

Conversely, from figure 6 we note that the thresholds perform very differently with respect to the different fairness criteria: particularly, the Sep_TP criterion presents the highest values of

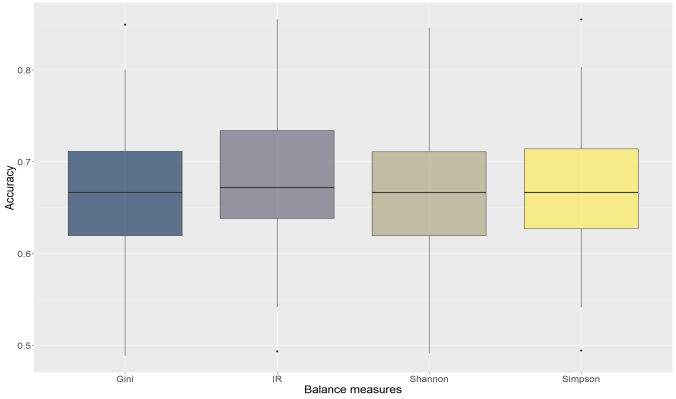


Fig. 5. Boxplot of the thresholds' Accuracy with respect to the *balance measures* in the binary case.

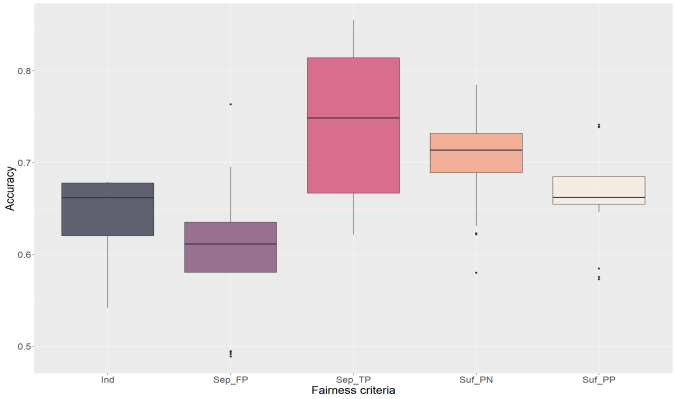


Fig. 6. Boxplot of the thresholds' Accuracy with respect to the *fairness criteria* in the binary case.

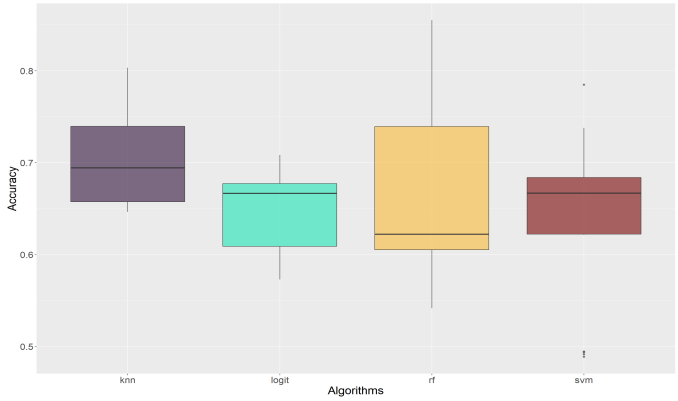


Fig. 7. Boxplot of the thresholds' Accuracy with respect to the *algorithms* in the binary case.

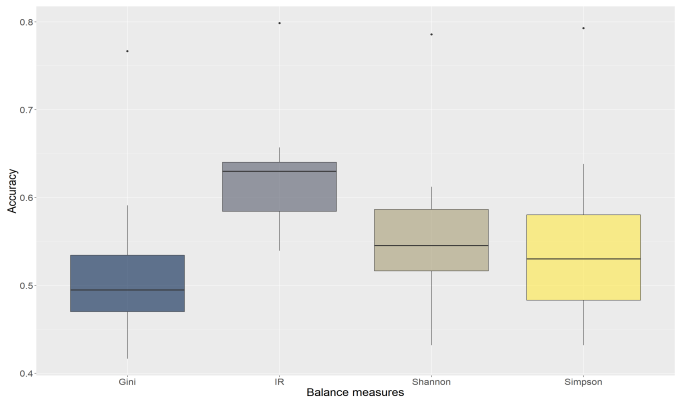


Fig. 8. Boxplot of the thresholds' Accuracy with respect to the *balance measures* in the multiclass case.

thresholds' accuracy, around 0.75 on average, with the largest interquartile range (or IQR, which graphically corresponds to the height of the box) indicating that the Sep_TP measure is the one with the largest variability of the accuracy values; then we find the two Sufficiency conditions, with thresholds' accuracy respectively around 0.72 for Suf_PN and 0.66 for Suf_PP, and the Independence criterion, always with an accuracy around 0.66 on average; the lowest values of the thresholds' accuracy are found in correspondence of Sep_FP, around 0.62.

Looking at figure 7 on the thresholds' accuracy with respect to the algorithms, we note that the best performances are reached with the K-nn classifier, with accuracy values around 0.70 on average; the logit and the svm algorithms present similar accuracy values around 0.67 on average, while the worst performances of the thresholds correspond to the random forest classifier with values around 0.67 on average; the random forest also presents the largest variability, with the highest values close to the ones of K-nn.

As regards the multiclass case, from figure 8 we note that the thresholds perform differently with respect to the different balance measures –contrary to the binary case–. Specifically, the IR index presents the highest accuracy values, around 0.63 on average, while the values decrease to around 0.55, 0.54 and 0.5 for the Shannon, Simpson and Gini indexes respectively. Then, looking at figure 9 on the thresholds' accuracy with respect to the fairness criteria, we observe the same pattern as in the binary case, but with lower values and greater variability for all the measures: the highest accuracy values are in correspondence of the Sep_TP condition and decrease to around 0.6 on average, followed by Suf_PN and Suf_PP (around 0.57), and by Independence and Sep_FP around 0.52 on average. Finally, about the thresholds' accuracy with respect to the algorithms represented in figure 10, we note a completely different pattern with respect to the binary case: the accuracy in correspondence of the random forest remains stable at around 0.6 with a wide variability and it presents the best accuracy value among the other algorithms (contrary to the binary case); indeed, the thresholds' accuracy decrease on average to around 0.58, 0.54 and 0.51 respectively for the svm, K-nn and the logit classifiers.

To conclude, the thresholds perform better in the binary case than in the multiclass case, with higher accuracy values overall; with respect to balance-unfairness-algorithm, the balance measures seem to have an impact only in the multiclass case, whereas the fairness criteria affect the thresholds' accuracy

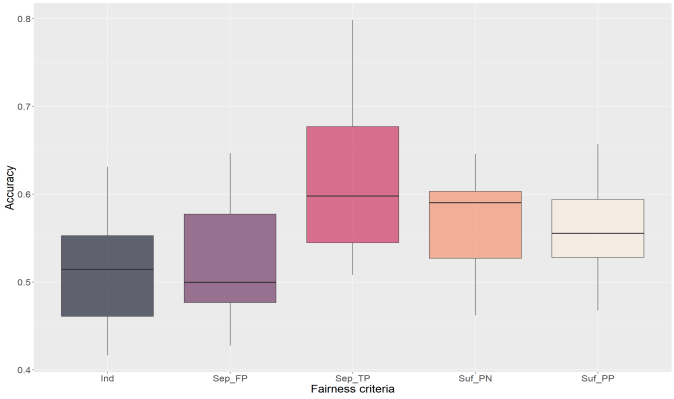


Fig. 9. Boxplot of the thresholds' Accuracy with respect to the *fairness criteria* in the multiclass case.

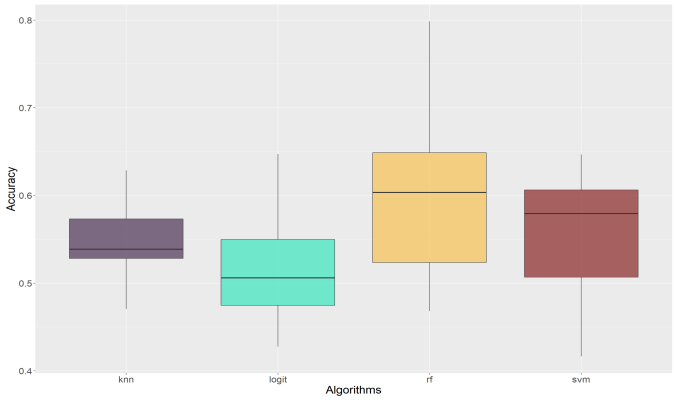


Fig. 10. Boxplot of the thresholds' Accuracy with respect to the *algorithms* in the multiclass case.

in both cases following the same pattern; also the algorithms have an impact –as they present different accuracy values– but without a precise trend.

V. Relation with previous studies and positioning in the fairness literature

The theoretical foundations of this study are detailed in [16]: the author describes the relations with the ISO/IEC standards on data quality measurement [17] and with risk management [18]. The proposed approach has been subsequently tested in [6] on a few hypothetical exemplar distributions. Then, in two subsequent studies we ran more exhaustive tests by applying two different mutation techniques to generate a number of derived synthetic datasets having different levels of balance, in one case to binary attributes [8] and in the other case to multiclass attributes [7]. This study moves forward this series of investigations: the fundamental novelty of this work is the construction of specific risk thresholds for balance measures and for fairness criteria, such that if the balance of the training set is greater than s, then the unfairness of the classification on the test set is expected to be less than f. Other novelties of this paper are given by the analysis of much more datasets and protected attributes (both binary and multiclass), other than the adoption of both the mutation techniques simultaneously on different protected attributes of a given dataset; moreover, we adopted four different algorithms to simulate different classification tasks in order to increase the variability of the output and the generalizability of the results. Note that, given the socio-technical nature of the issue, in our studies we look at data imbalance as a risk factor and not as a technical fix. Indeed, we believe that a risk assessment approach creates space for active human considerations and interventions, thus entrusting the ultimate responsibility to human decisions. In addition, fairness and bias are studied for a long time in the social and human sciences: an interdisciplinary approach would be more appropriate than a pure computational one.

Our contribution can be located in the main area of research on algorithmic bias and fairness, with a specific focus on inputs and processes, as suggested by several recent studies (e.g., [19], [20] and [21]). Similar to our approach but wider in scope, it is the work of Takashi Matsumoto and Arisa Ema [22]: they propose a risk chain model for risk reduction in Artificial Intelligence (AI) services, named RCM, where they consider both data quality and data imbalance as risk factors. Our work fits the RCM framework because we propose a quantitative way to measure balance. In addition, our work is complementary to the existing toolkits for bias detection and mitigation [23] since the balance measures proposed herein have not been taken into account yet.

VI. Limitations

As limitations of our approach, first of all we highlight that we did not perform the hyper-parameters tuning of the algorithms involved in our study as we were not interested in a specific algorithms performance analysis, but rather in varying the classifier in order to increase the variability of the output and the generalizability of the results; nevertheless, a better fitting of the data could reveal more meaningful differences among the different algorithms.
We also remark that we chose the accuracy as a discriminant for the identification of the thresholds, but an analogous study can be conducted by considering a different evaluation metric as a reference and identifying the thresholds according to the performances based on such metric.
Finally, it would be recommended to take into account other indexes of balance and unfairness, also by including measures for non-categorical data, in order to extend the findings of this study. Moreover, 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.

VII. Conclusions and future work

In this study we defined and tested a methodology to identify thresholds of balance such that the unfairness of the classification is expected to be less than the desired levels. To conduct this analysis, we adopted a previously defined metric-based approach to assess imbalance in a given dataset as a risk indicator of discriminatory classification outcomes of automated decision-making systems [6]. First of all, we selected

a set of balance measures (the Gini, Shannon, Simpson and Imbalance Ratio indexes), we generated a large number of synthetic datasets and measured the different levels of imbalance in the training sets, whereas through a set of fairness criteria we assessed on the test sets the discrimination occurring in the outcomes obtained from different classifiers. After that, we built the thresholds s and f by following a specific procedure, and we generated a new collection of data on which we evaluated the performances of the defined thresholds through different evaluation metrics (Accuracy, Precision, Sensitivity, Specificity and F1-score). Specifically, for each combination of balance-unfairness-algorithm we selected the configuration of thresholds that presented the highest accuracy. We conducted the experiment and analyzed the results separately for binary and multiclass attributes, and we assessed in detail the thresholds' goodness with respect to balance measures, fairness criteria and algorithms.

By assessing the thresholds through the evaluation metrics, we observed that the values of balance under s indicate levels of unfairness over f, but they even tend to overestimate the risk. In both the binary and the multiclass cases, the Precision index –which indicates the Positive Predictive value– is the one that presents the highest values among all the evaluation metrics, while the worst performances are identified through the Specificity index, suggesting that the thresholds tend to overestimate the risk. Overall, we also noted that the identified thresholds perform better in the binary case than in the multiclass case.

The evaluation of the thresholds' accuracy with respect to the balance measures revealed that there is no significant difference between the different indexes of balance, except for the IR index, which presents the highest accuracy values in both the binary and the multiclass cases. Conversely, the thresholds perform very differently with respect to the different fairness criteria, and overall, for the multiclass case we observe the same pattern as in the binary case, but with lower values and greater variability for all the fairness criteria. For the thresholds' accuracy with respect to the different algorithms, we found completely different values between the binary and the multiclass cases, thus the algorithms have an impact on the performances of the thresholds, but in this study we could not identify a specific pattern.

Hence, further work shall be devoted to a thorough and systematic investigation of the thresholds to be used relating to different classification algorithms, also by performing the hyper-parameters tunings for each classifier. Further work is also needed to better assess the reliability of the risk thresholds, for instance by conducting an analogous study by considering other evaluation metrics (different from the accuracy) as a discriminant to define the best thresholds for each combination of balance-unfairness-algorithm.

We hope that these findings on risk thresholds for detecting algorithmic fairness with balance measures will improve the identification and assessment of discrimination risks in ADM systems by measuring the imbalance of the protected attributes in the input data.

REFERENCES

[1] F. Chiusi, S. Fischer, N. Kayser-Bril, and M. Spielkamp, "Automating Society Report 2020," https://automatingsociety.algorithmwatch.org, Berlin, Oct. 2020.

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

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

[4] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[5] E. U. A. for Fundamental Rights, "EU Charter of Fundamental Rights - Article 21 - Non-discrimination," https://fra.europa.eu/en/eu-charter/article/21-non-discrimination, December 2007.

[6] A. Vetrò, M. Torchiano, and M. Mecati, "A data quality approach to the identification of discrimination risk in automated decision making systems," *Government Information Quarterly*, vol. 38, no. 4, 2021.

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

[8] M. Mecati, A. Vetrò, and M. Torchiano, "Detecting risk of biased output with balance measures," *J. Data and Information Quality*, vol. 14, no. 4, nov 2022. [Online]. Available: https://doi.org/10.1145/3530787

[9] U. M. Learning, "Default of Credit Card Clients Dataset," https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset, 2016.

[10] "UCI Machine Learning Repository: Statlog (German Credit Data) Data Set," https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data), 1994.

[11] "UCI Machine Learning Repository: Student Performance Data Set," https://archive.ics.uci.edu/ml/datasets/Student+Performance, 2014.

[12] "UCI Machine Learning Repository: Adult Data Set," https://archive.ics.uci.edu/ml/datasets/adult, 1996.

[13] "Drug consumption (quantified) data set," https://archive.ics.uci.edu/ml/datasets/Drug+consumption+\%28quantified\%29, 2016.

[14] "Heart disease data set," https://archive.ics.uci.edu/ml/datasets/heart+disease, 1988.

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

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

[17] International Organization for Standardization, "ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE," https://www.iso.org/standard/64764.html, 2014.

[18] ——, "ISO 31000:2018 Risk management — Guidelines," https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html, 2018.

[19] D. Firmani, L. Tanca, and R. Torlone, "Ethical dimensions for data quality," *Journal of Data and Information Quality (JDIQ)*, vol. 12, no. 1, pp. 1–5, 2019.

[20] E. Pitoura, "Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias," *Journal of Data and Information Quality*, vol. 12, no. 3, pp. 12:1–12:8, Jul. 2020.

[21] B. Hutchinson and M. Mitchell, "50 Years of Test (Un)fairness: Lessons for Machine Learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 49–58.

[22] T. Matsumoto and A. Ema, "RCModel, a Risk Chain Model for Risk Reduction in AI Services," http://arxiv.org/abs/2007.03215, Jul. 2020.

[23] M. S. A. Lee and J. Singh, "The landscape and gaps in open source fairness toolkits," 2020.