Integrating SQuaRE data quality model with ISO 31000 risk management to measure and mitigate software bias

Alessandro Simonetta
Department of Enterprise Engineering
University of Rome Tor Vergata
Rome, Italy
alessandro.simonetta@gmail.com
ORCID: 0000-0003-2002-9815

Antonio Vetrò
Dept. of Control and Computer Eng.
Politecnico di Torino
Turin, Italy
antonio.vetro@polito.it
ORCID: 0000-0003-2027-3308

Maria Cristina Paoletti
Rome, Italy
mariacristina.paoletti@gmail.com
ORCID: 0000-0001-6850-1184

Marco Torchiano
Dept. of Control and Computer Eng.
Politecnico di Torino
Turin, Italy
marco.torchiano@polito.it
ORCID: 0000-0001-5328-368X

***Abstract* — In the last decades the exponential growth of available information, together with the availability of systems able to learn the knowledge that is present in the data, has pushed towards the complete automation of many decision-making processes in public and private organizations. This circumstance is posing impelling ethical and legal issues since a large number of studies and journalistic investigations showed that software-based decisions, when based on historical data, perpetuate the same prejudices and bias existing in society, resulting in a systematic and inescapable negative impact for individuals from minorities and disadvantaged groups. The problem is so relevant that the terms data bias and algorithm ethics have become familiar not only to researchers, but also to industry leaders and policy makers. In this context, we believe that the ISO SQuaRE standard, if appropriately integrated with risk management concepts and procedures from ISO 31000, can play an important role in democratizing the innovation of software-generated decisions, by making the development of this type of software systems more socially sustainable and in line with the shared values of our societies. More in details, we identified two additional measure for a quality characteristic already present in the standard (completeness) and another that extends it (balance) with the aim of highlighting information gaps or presence of bias in the training data. Those measures serve as risk level indicators to be checked with common fairness measures that indicate the level of polarization of the software classifications/predictions. The adoption of additional features with respect to the standard broadens its scope of application, while maintaining consistency and conformity. The proposed methodology aims to find correlations between quality deficiencies and algorithm decisions, thus allowing to verify and mitigate their impact.**

Keywords— ISO Square, ISO31000, data ethics, data quality, data bias, algorithm fairness, discrimination risk

I. INTRODUCTION

Software nowadays replace most human decisions in many contexts [1] ; the rapid pace of innovation suggests that this phenomenon will further increase in the future [2]. This trend has been enabled by the large availability of data and of the technical means to analyze them for building the predictive, classification, and ranking models that are at the core of automated decision making (ADM) systems. Advantages for using ADM systems are evident and they concern mainly scalability, efficiency, and removal of decision makers' subjectivity. However, several critical aspects have emerged: lack of accountability and transparency [3], massive use of natural resources and low-unpaid labor to building extensive training sets [4] , the distortion of the public sphere of political discussion [5], and the amplification of existing inequalities in society [6]. This paper focuses on the latter problem, which occurs when automated software decisions "*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*" [7] . In practice, software systems may perpetuate the same bias of our societies, systematically discriminating the weakest people and exacerbating existing inequalities [8]. A recurring cause for this phenomenon is the use of incomplete and biased data, because of errors or limitations in the data collection (e.g., under-sampling of a specific population group) or simply because the distributions of the original population are skewed. From a data engineering perspective, this translates into imbalanced data, i.e. a condition with an unequal distribution of data between the classes of a given attribute, which causes highly heterogeneous accuracy across the classifications [9] [10]. Imbalanced data is known to be problematic in the machine learning domain since long time [11]. In fact, imbalanced datasets may lead to imbalanced results, which in the context of ADM systems means differentiation of products, information and services based on personal characteristics. In applications such as allocation of social benefits, insurance tariffs, job profiles matching, etc., such differentiations can lead to unjustified unequal treatment or discrimination.

For this reason, we maintain that imbalanced and incomplete data shall be considered as a risk factor in all the ADM systems that rely on historical data and operate in relevant aspects of the lives of individuals. Our proposal relies on the integration of the measurement principles of the ISO SQuaRE [12] with the risk management process defined in ISO 31000 [13] to assess the potential risk of discriminating software output and take action for remediations. In the paper, we describe the theoretical foundations, and we provide a workflow of activities. We believe that the approach can be

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3114/paper-04.pdf

useful to a variety of stakeholders for assessing the risk of discriminations, including the creators or commissioners of software systems, researchers, policymakers, regulators, certification or audit authorities. Assessments should prompt taking appropriate action to prevent adverse effects.

II. METHODOLOGY

Figure 1 gives an overview of the proposed methodology. The process begins with the common subdivision of the original data into training and test data. At this point, it is possible to measure the quality in the training data (balance and completeness) and the fairness in the results obtained on the test data. Data balance measures extend the characteristics of the data quality model (ISO/IEC 25012), while completeness measures complement it. Data quality measures give rise to an indicator for unbalanced or incomplete data for the sensitive characteristics, which implicates a risk of biased classifications by the algorithm. In this circumstance, it is necessary to also assess the fairness of the algorithms used through the measures outlined in this paper. The presence of unfair results from the point of view of sensitive features in correspondence with poor quality data leads to the necessary data enrichment step to try to mitigate the problem. Thus, our proposed methodology is composed by two main blocks:

A. Risk analysis: measuring the risk that a training set could contain unbalanced data, integrating the SQuaRE approach with ISO 31000 risk management principles;

B. Risk evaluation: verify that a high level of risk corresponds to unfairness, and in positive case enrich original data with synthetic data to mitigate the problem.

A. *Risk analysis: where SQuaRE and ISO3100 meet*

We integrate the SQuaRE theoretical framework with the ISO 31000 risk management principles to measure the risk that an unbalanced or incomplete training set might cause discriminating software output. Since the primary recipients of this document are the participants of the "*3rd International Workshop on Experience with SQuaRE series and its Future Direction*"[1], we do not describe here the standard, however we summarize the most important aspects for the scope of the paper. Firstly, we remind that SQuaRE includes quality modeling and measurements of software products[2], data and software services. According to the philosophy and organization of this family of standards, quality is categorized into one or more quantifiable characteristics and sub-characteristics. For example, the standard ISO/IEC 25012:2011 formalizes the product quality model as composed of eight characteristics, which are further subdivided into sub-characteristics. Each (sub) characteristic relates to static properties of software and dynamic properties of the computer system[3]. The ISO/IEC 25012:2008 on data quality has 15 characteristics: 5 of them belongs to the "inherent" point of view (i.e., the quality relies only on the characteristics of the data per se), 3 of them are system-dependent (i.e., the quality depends on the characteristics of the system hosting the data and making it available), the remaining 7 belonging to both points of view. Data balance is not recognized as a characteristic of data quality in ISO/IEC 25012:2008: it is proposed here as an additional inherent characteristic. Because of its role in the generation of biased

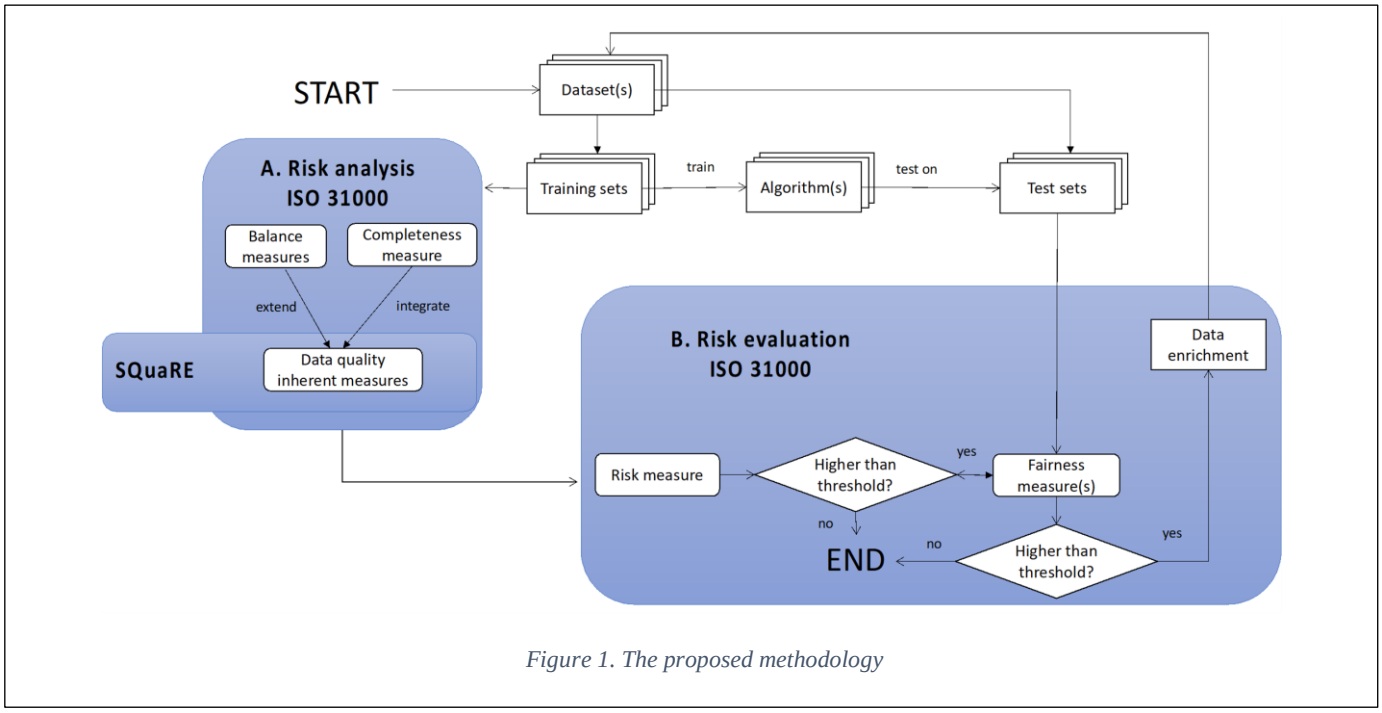

Figure 1. The proposed methodology

[1] See http://www.sic.shibaura-it.ac.jp/~tsnaka/iwesq.html

[2] A software product is a "set of computer programs, procedures, and possibly associated documentation and data" as defined in ISO/IEC 12207:1998. In SQuaRE standards, software quality stands for software product quality.

[3] A system is the "combination of interacting elements organized to achieve one or more stated purposes" (ISO/IEC 15288:2008), for example the aircraft system. It follows that a *computer system* is "a system containing one or more components and elements such as computers (hardware), associated software, and data", for example a conference registration system. An ADM system that determines eligibility for aid for drinking water is a *software system*.

software output, data balance reflects the propagation principle of SQuaRE: the quality of the software product, service and data affects the quality in use. Therefore, evaluating and improving product/service/data quality is one mean of improving the system quality in use. A simplification of this concept is the GIGO principle ("garbage in, garbage out"): data that is outdated, inaccurate and incomplete make the output of the software unreliable. Similarly, unbalanced data will probably cause unbalanced software output, especially in the context of machine learning and AI systems trained with that data. This principle applies also to completeness, which is already an inherent characteristic of data quality in SQuaRE: in this work we propose an additional metric to those proposed in ISO/IEC 25024:2015, that is more suitable for the problem of biased software.

To better address the problem of biased software output, we consider the measures of data balance and completeness not only as extension of SQuaRE data quality modelling but also as risk factors. Here comes the integration of SQuaRE theoretical and measurement framework with the ISO 31000:2018 standard for risk management. The standard defines guiding principles and a process of three phases: risk identification, risk analysis and risk evaluation. Here, we briefly describe them and specify the relation with our approach.

Risk identification refers to finding, recognizing and describing risks within a certain context and scope, and with respect to specific criteria defined prior to risk assessment. In this paper, this is implicitly contained in the motivations and in the problem formulation: it is the risk of discriminating individuals or groups of individuals by operating software systems that automate high-stake decisions for the lives of people.

Risk analysis is the understanding of the characteristics and levels of the risk. This is the phase where measures of data balance and completeness are used as indicators, due to the propagation effect previously described.

Risk evaluation, as the last step, is the process in which the results of the analysis are taken into consideration to decide whether additional action is required. If affirmative, this process would then outline available risk treatment options and the need for conducting additional analyses. In our case, specific thresholds for the measures should be decided for the specific prediction/classification algorithms used, the social context, the legal requirements of the domain, and other relevant factors for the case at hand. In addition to the technical actions, the process would define other types of required actions (e.g., reorganization of decision processes, communication to the public, etc.) and the actors who must undertake them.

1) Completeness measure

The completeness measure proposed is agnostic with respect to classical ML data classification because for our purposes we are interested in evaluating those columns that assume values in finite and discrete intervals, which we will call categorical with respect to the row data. This characteristic will allow us to consider the set of their values as the digits constituting a number in a variable base numbering system. The idea of the present study is based on the principle that a learning system provides predictions consistent with the data with which it has been trained. Therefore, if it is fed with non-homogeneous data it will provide unbalanced and discriminatory predictions with respect to reality. For this reason, the methodology we propose starts with the analysis phase of the reality of interest and of the dataset, an activity that must be carried out even before starting the pre-training phase in line with previous studies where some of the authors proposed the use of balance measures in automated decision

Index	Formula	Normalized formula	Notes
Gini	$\mathcal{G} = 1 - \sum_{i=1}^{m} f_i^2$	$\mathcal{G}_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right)$	m is the number of classes f is the relative frequency of each class $f_i = \frac{n_i}{\sum_{i=1}^{m} n_i}$ n_i= absolute frequency The higher G and G_n, the higher is the heterogeneity: it means that categories have similar frequencies The lower the index, the lower is the heterogeneity: a few classes account for majority of instances
Simpson	$D = \frac{1}{\sum_{i=1}^{m} f_i^2}$	$D_n = \frac{1}{m-1}\left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$	For m, f, f_i and n_i check Gini Higher values of D and D_n indicate higher diversity in terms of probability of belonging to different classes The lower the index, the lower is the diversity, because frequencies are concentrated in a few classes

Table 1. Example of measures of balance

making systems [14][15][16][17]. In particular, during this phase, it is necessary to identify all the independent columns that define whether the instance belongs to a class or category. Suppose we have a structured dataset as follows:

$$DS = \{ C_0, C_1, \dots , C_{n-1} \} \quad (1)$$

Indicating with the set S the positions of the columns categorising the instances, functionally independent of the other columns in the dataset:

$$S \subseteq \{ 0, 1, \dots , n-1 \}, dim(S) = m , m \leq n \quad (2)$$

we can analyze the new dataset consisting of the columns CS(j) with j ∈ [0, m − 1].

Having said that, we can decide to use two different notions of completeness: maximum or minimum. In the first case the presence in the dataset of a greater number of distinct instances that belong to the same categorising classes constitutes a constraint for all the other instances of the dataset. That is, one must ensure that one has the same number of replicas of distinct class combinations for distinct instances. Instead, in the second case it is sufficient to have at least one combination of distinct classes among those possible for each instance. For simplicity, but without loss of generality of the procedure, we will explore the minimum completeness of the dataset, then we will reduce the dataset to just the columns (*j*) by removing duplicate rows. We will use the Python language to explicate the calculation formulas and make the mathematical logic implied less abstract. The Python language has the pandas library, which makes it possible to carry out analysis and data manipulation in a fast, powerful, flexible and easy-to-use manner. Through the DataFrame class it is possible to load data frames from a simple csv file:

```
import pandas as pd
df=pd.read_csv(<file_name>)
```

The ideal value of minimum completeness for the combinatorial metric is when in the dataset there is at least one instance that belongs to each distinct combination of categories. The absence of some combination could create the lack of information that we do not want to exist. To calculate the total number of distinct combinations we need to calculate the product of the distinct replicas per single category.

```
k=( df['CS0'].unique().size *
    df['CS1'].unique().size *...*
    df['CSm-1'].unique().size )
```

On the other hand, in the dataset we only have the characterising columns so we can derive the true number of distinct instances in order to determine how far the data deviates from the ideal case.

```
len (df.drop_duplicates())/ k
```

The value for maximum completeness is calculated from the maximum number of duplicates of the same combinations of characterising columns. For this reason it is necessary to maintain in the dataset in addition to the columns (*j*) a discriminating identification field of the rows with the same values in these columns. To determine the potential total, once the maximum number of duplications (M) has been determined, it is necessary to extend this multiplication factor to all other classes.

```
M= df.groupby(['CS0',...,'CSm-1']).
   size().reset_index(name='counts').
   counts.max()

len(df)/(M*k)
```

2) Balance measures

Since imbalance is defined as an unequal distribution between classes [9], we focus on categorical data. In fact, most of the sensitive attributes are considered categorical data, such as gender, hometown, marital status, and job. Alternatively, if they are numeric, they are either discrete and within a short range, such as family size, or they are continuous but often re-conducted to distinct categories, such as information on "age" which is often discretized into ranges such as "< 18", "19-35", "36-50", "51-65", etc. We show two examples of measures in Table 1, retrieved from the literature of social and natural sciences, where imbalance is known in terms of (lack of) heterogeneity and diversity. They are normalized in the range 0-1, where 1 correspond to maximum balance and 0 to minimum balance, i.e. imbalance. Hence, lower level of balance measures mean a higher risk of bias in the software output.

B. Risk evaluation with fairness measures

The majority of fairness measures in machine learning literature rely on the comparison of accuracy, computed for each population group of interest [18]. For computing the accuracy, two different approaches can be adopted: the first attempts to measure the intensity of errors, i.e. the deviation between prediction and actual value (precision), while the other measures the general direction of the error. Indicating with e_i the i^{th} error, with f_i and d_i respectively the i^{th} forecast and demand, we have:

$$e_i = f_i - d_i \quad (3)$$

At this point we can add up all the errors with sign and find the average error:

$$average\ error = \frac{1}{n}\sum_i e_i \quad (4)$$

However, this measure is very crude because error compensation phenomena may be present so generically it is preferred to use the mean of the absolute error or the square root of the mean square error:

$$MAE = \frac{1}{n}\sum_i |e_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_i {e_i}^2} \quad (6)$$

RMSE is sensitive to important errors, while from this point of view MAE is fairer because it considers all errors at the same level. Moreover, if our prediction tends to the median it will get a good value of MAE, vice versa if it approaches the mean it will get a better result on RMSE.

Under conditions where the median is lower than the mean, for example in processes where there are peaks of demands compared to normal steady state operation, it will not be convenient to use MAE which will introduce a bias while it will be more convenient to use RMSE. Things are

reversed if outliers are present in the distribution as MAE is less sensitive than RMSE.

To measure model performance, you can choose to measure error with one or more KPI.

In the case of classification algorithms instead you can use the confusion matrix that allows you to compute the number of true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs):

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \cdots & \cdots & \cdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \quad (7)$$

You can use the following equations to calculate the following values:

$$TP(i) = p_{ii} \quad (8)$$

$$FP(i) = \sum_{k=1,k\neq i}^{n} p_{ki} \quad (9)$$

$$FN(i) = \sum_{k=1,k\neq i}^{n} p_{ik} \quad (10)$$

$$TN(i) = \sum_{k=1,k\neq i}^{n} p_{kk} \quad (11)$$

At this point, the single values could be computed for each population subgroup (e.g., "Asian" vs "Caucasian" vs "African-American" etc. , or "Male" vs "Female", etc.) and the same applies to the concepts of precision, recall, and accuracy, known from the literature and reported here:

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$Recall = \frac{TP}{TP + FN} \quad (13)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

Fairness measures should be then compared with appropriate thresholds selected with respect to social context in which the software application is used. If the unfairness is higher than the maximum allowed thresholds, then the original dataset should be integrated with synthetic data to mitigate the problem. One way to repopulate the dataset without causing distortion in the data is to add replicas of data selected from the same set at random (known as bootstrapping [19]); other rebalancing techniques have been proposed in the literature (e.g. SMOTE [20], ROSE [21]).

III. *Relation with Literature and Our Past Studies*

An approach similar to ours is the work of Takashi Matsumoto and Arisa Ema [22], who proposed a risk chain model (RCM) for risk reduction in Artificial Intelligence services: the authors consider both data quality and data imbalance as risk factors. Our work can be easily integrated into the RCM framework, because we offer a quantitative way to measure balance and completeness, and because it is natively related to the ISO/IEC standards on data quality requirements and risk management.

Other approaches which can be connected to ours are in the direction of labeling datasets: for example "The Dataset Nutrition Label Project" [4] aims to identify the "key ingredients" in a dataset such as provenance, populations, and missing data. The label takes the form of an interactive visualization that allows for exploring the previously mentioned aspects and spot flawed, incomplete, or problematic data. One of the author of this paper took inspiration from this study in previous works for "Ethically and socially-aware labeling" [16] and for a data annotation and visualization schema based on Bayesian statistical inference [17] always for the purpose of warning about the risk of discriminatory outcomes due to poor quality of datasets. We started from that experience to conduct preliminary case studies on the reliability of the balance measures [14] [15]: in this work we continue in that direction by adding a measure of completeness and proposing an explicit workflow of activities for the combination of SQuaRE with ISO 31000.

IV. *Conclusion and Future Work*

We propose a methodology that integrates SQuaRE measurement framework with the ISO 31000 process with the goal of evaluating balance and completeness in a dataset as risk factors of discriminatory outputs of software systems. We believe that the methodology can be a useful instrument for all the actors involved in the development and regulation of ADM systems, and, from a more general perspective, it can play an important role in the collective attempt of placing democratic control on the development of these systems, that should be more accountable and less harmful than how they are now. In fact, the adverse effects of ADM systems are posing a significant danger for human rights and freedoms as our societies increasingly rely on automated decision making. It must be stressed that this is still at a prototypical stage and further studies are necessary to improve the methodology and to assess the reliability of the proposed measures, for example to find meaningful risk thresholds in relation to the context of use and the severity of the impact on individuals. The current paper is also way to seek engagement from other researchers in a community effort to test the workflow in real settings, improve it and build an open registry of additional measures combined with evaluations benchmark. Finally, we are conscious that technical adjustments are not enough, and they should be integrated with other types of actions because of the socio-technical nature of the problem.

References

[1] F. Chiusi, S. Fischer, N. Kayser-Bril, and M. Spielkamp, "Automating Society Report 2020," Berlin, Oct. 2020. Accessed: Nov. 10, 2020. [Online]. Available: https://automatingsociety.algorithmwatch.org

[2] E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, Reprint edition. New York London: W. W. Norton & Company, 2016.

[3] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard Univ Pr, 2015.

[4] It is a joint initiative of MIT Media Lab and Berkman Klein Center at Harvard University: https://datanutrition.org/.

[4] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence: The Real Worlds of Artificial Intelligence*. New Haven: Yale Univ Pr, 2021.

[5] P. N. Howard, *Lie Machines: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives*. New Haven ; London: Yale Univ Pr, 2020.

[6] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press, 2018.

[7] B. Friedman and H. Nissenbaum, "Bias in Computer Systems," *ACM Trans Inf Syst*, vol. 14, no. 3, pp. 330–347, Jul. 1996, doi: 10.1145/230538.230561.

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

[9] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[10] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.

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

[12] International Organization for Standardization, "ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE," *ISO-International Organization for Standardization*, 2014. https://www.iso.org/standard/64764.html (accessed Nov. 10, 2020).

[13] International Organization for Standardization, "ISO 31000:2018 Risk management — Guidelines," *ISO - International Organization for Standardization*, 2018. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/56/65694.html (accessed Nov. 10, 2020).

[14] M. Mecati, F. E. Cannavò, A. Vetrò, and M. Torchiano, "Identifying Risks in Datasets for Automated Decision–Making," in *Electronic Government*, Cham, 2020, pp. 332–344. doi: 10.1007/978-3-030-57599-1_25.

[15] A. Vetrò, M. Torchiano, and M. Mecati, "A data quality approach to the identification of discrimination risk in automated decision making systems," *Gov. Inf. Q.*, p. 101619, Sep. 2021, doi: 10.1016/j.giq.2021.101619.

[16] E. Beretta, A. Vetrò, B. Lepri, and J. C. De Martin, "Ethical and Socially-Aware Data Labels," in *Information Management and Big Data*, Cham, 2019, pp. 320–327.

[17] E. Beretta, A. Vetrò, B. Lepri, and J. C. D. Martin, "Detecting discriminatory risk through data annotation based on Bayesian inferences," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Mar. 2021, pp. 794–804. doi: 10.1145/3442188.3445940.

[18] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019.

[19] "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques - ScienceDirect." https://doi.org/10.1016/j.patrec.2013.04.019 (accessed Sep. 18, 2021).

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.

[21] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 92–122, Jan. 2014, doi: 10.1007/s10618-012-0295-5.

[22] T. Matsumoto and A. Ema, "RCModel, a Risk Chain Model for Risk Reduction in AI Services," *ArXiv200703215 Cs*, Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.03215