

Investigating Fairness with FanFAIR: is Pre-processing Useful Only for Performances?

Michele Rispoli*, Marco S. Nobile^{†||}, Luca Manzoni*, Alberto D’Onofrio*, Marco Confalonieri^{‡§},
Francesco Salton^{‡§}, Paola Confalonieri^{‡§}, Barbara Ruaro^{‡§}, and Chiara Gallese^{¶**}

*Department of Mathematics, Informatics and Geosciences, University of Trieste, Italy

[†]Department of Environmental Sciences, Informatics and Statistics, Ca’ Foscari University of Venice, Italy

[‡]Pulmonology Unit, University Hospital of Cattinara, Trieste, Italy

[§]Department of Medical Surgical and Health Sciences, University of Trieste, Italy

[¶]Department of Law, University of Turin, Italy

^{||}Bicocca Bioinformatics, Biostatistics and Bioimaging research center, Monza, Italy

Email: **chiara.gallese@unito.it

Abstract—Artificial Intelligence, and Machine Learning systems in general, are becoming pervasive in our society, from the industry to the public administration. AI can often provide a very efficient means to support decision-making, but it can represent a danger for high-risk applications such as bio-medicine and healthcare. In particular, biased datasets might lead to inaccurate or discriminatory ML systems, undermining the accuracy of their predictions and putting patients’ health at risk. FanFAIR is a python tool that provides the community with a semi-automatic tool for datasets’ fairness assessment. FanFAIR is designed to integrate qualitative considerations – such as ethics, human rights assessment, and data protection – with quantitative indicators of dataset’s fairness, such as balance, the presence of invalid entries, or outliers. In this work, we extend FanFAIR to deal with categorical data, and introduce a new algorithm for outlier detection in the presence of missing values. We then provide a case study on the data collected from COVID patients admitted to pneumology departments in Italy. We show how the successive steps of data cleaning and variable selection improve the indicators provided by FanFAIR. This shows that data cleaning procedures are not only necessary to improve the performance of the machine learning algorithm using the data for learning, but are also a way to improve (a measure of) fairness. Hence, the proposed case study provides an example in which performance and fairness are not in contrast, like it is commonly believed to be, but they improve together.

Index Terms—fairness, debiasing, preprocessing, data cleaning, sensitive attributes, dataset assessment

I. INTRODUCTION

As Artificial Intelligence (AI) continues to be increasingly employed across different domains, especially in high-stakes fields such as healthcare, discussion over its societal impact has gradually become more present in the literature. One of the main concerns in AI and Machine Learning (ML) ethics is the notion of fairness concerning datasets used in

This work was partially funded by the European Union’s Horizon Europe program with the DataCom Project (grant agreement no. 101108151). Views and opinions expressed are those of the author(s) only and do not reflect those of the EU or the European Commission. This work was also partially supported by DAIS – Ca’ Foscari University of Venice, within the ADIR program.

training AI models [1]. Fairness, in this context, relates to mitigating biases embedded in datasets to prevent AI systems from perpetuating or exacerbating inequities. While technical methods like debiasing have been advanced to address these challenges, a broader debate questions the effectiveness of such approaches, raising fundamental concerns about their ability to address systemic inequalities.

At the core of dataset fairness is the recognition that datasets can encode historical and structural biases that reflect an unfair world [2]. Such biases are shaped by gendered, racial, colonial, and other discriminatory practices that are embedded in healthcare, employment, and other societal structures [3].

The European Union has taken steps to address the issue of fairness in AI systems, such as enacting provisions that mandate fairness, transparency, and accountability in AI design [4]. However, as the European Digital Rights (EDRi) report highlights, these policies often take a limited view of fairness, focusing on debiasing data without addressing the larger societal context in which AI systems are deployed. Furthermore, the report suggest that even efforts to conform to the AI Act ensuring that the datasets are “representative, error-free, and complete” may still result in AI systems that reflect an inequitable world [2].

This highlights a significant tension: while datasets can be technically debiased – according to some predefined fairness metrics – they may still perpetuate and amplify the injustices inherent in the systems from which they were derived. In line with this critique, Hanna et al. emphasize the importance of reframing discussions about fairness in AI away from the algorithmic level and toward the social and institutional contexts in which these systems are implemented [5]. Nevertheless, both the EDRi report and the literature recognize the importance of data quality and statistical considerations pertaining to data, as those significantly influence the quality of the ML model [6]–[8]. This tension between the goals of dataset fairness and the realities of an unjust world points to a need for broader, more comprehensive approaches to

fairness in AI. It is not enough to focus solely on the technical aspects of fairness – such as ensuring that datasets are balanced and free from errors – without also considering the social, political, and economic systems that shape AI’s development and deployment. Addressing fairness in AI requires engaging with these larger systems of power and inequality, rather than relying solely on technical solutions to solve deeply entrenched social problems.

More concretely, we believe that unfairness is not solely depending on the data, but it extends to several human activities that impact how data is collected and processed. Some examples are ensuring the consent to collect and reuse patients’ personal data [9], enacting the transparency principle over the reuse of data [10], and performing an extensive ethics assessment, which is much broader than just obtaining the permission from an Ethical Review Board [11]. The ethics assessment includes abiding to principles such as Accountability, Dignity and Self-Determination, Traceability, Involvement of Stakeholders, Risk Assessment, and Impact on Society, which are important evaluations for the whole AI life cycle and can help in mitigating and preventing several AI harms.

FanFAIR was published in this context: a software solution for the evaluation of (medical) datasets, that exploits a rule-based fuzzy inference system to provide a quick, semi-automatic assessment of the fairness of data [12]. In FanFAIR, all the aforementioned literature was considered by integrating statistical properties of the dataset, which are computed autonomously by our software, with qualitative considerations entered by the user. FanFAIR can be useful to decide how to pre-process the dataset, and even to decide whether discarding the dataset altogether would be the most appropriate choice, in order to avoid the risk of increasing AI harm. We thus believe that the assessment provided by FanFAIR may serve as the basis for the dataset evaluation, offering helpful insights since the earliest stages of development of AI systems.

In this work, we extend FanFAIR with additional functionalities to simplify data import (Section II-B3) and to perform an improved outliers detection also in presence of missing data (Section II-B2). We also introduce a novel facility for the analysis of sensitive variables, which was integrated with the fuzzy inference system (Section II-B1). Additionally, we test our improved method on a real world dataset about COVID patients, presenting a concrete example of analysis performed with FanFAIR, and discussing the impact of pre-processing on fairness and predicting performance of ML models trained on our data (Section II-C). After presenting our results in Section III, we conclude the manuscript with some comments about limitations of FanFAIR, and possible future developments.

Both the source code and the documentation for FanFAIR are available on GITHUB: <https://github.com/aresio/FanFAIR>. FanFAIR can also be installed using pip.

II. METHODS

A. FanFAIR

FanFAIR is a Python library designed for the semi-automatic assessment of dataset fairness using a rule-based

approach that leverages fuzzy logic [12]. The tool calculates multiple fairness metrics over a dataset, and combines them into a single score, which enables researchers to evaluate a dataset’s fairness more efficiently. The metrics considered by FanFAIR for the assessment of a dataset are the following [12]:

- *balance*: how balanced the dataset is, with respect to the output labels;
- *numerosity*: whether the number of samples is reasonable with respect to the number of variables;
- *unevenness*: how frequent outliers are;
- *incompleteness*: how frequent missing values are within the dataset;
- *quality*: a manually set feature evaluating the quality of the dataset, with respect to noise or similar characteristics;
- (*legal*) *compliance*: whether the creator(s) of the dataset respected all laws and legal duties.

FanFAIR aggregates these features into an overall fairness score by using a Fuzzy Inference System (FIS) based on a 0-order Sugeno reasoner. The system relies on fuzzy logic to handle the inherent fuzziness in fairness assessments, which is generally not black-or-white and cannot rely on crisp arbitrary thresholds. FanFAIR is semi-automatic because most of the process is automated, although two metrics (namely, quality and legal compliance) inherently require a human intervention.

B. New functionalities in FanFAIR

In this work, we updated FanFAIR with three new functionalities: (i) the analysis for sensitive variables; (ii) the identification of outliers with an improved isolation forest able to deal with missing values in data; (iii) extended support for Pandas dataframes and non real-valued variables in the dataset. These functionalities are described in the following subsections.

1) *Sensitive variables analysis*: FanFAIR exposes a new method to specify which variables of the dataset should be considered sensitive. In this work, we define as “sensitive” all the variables that should not be directly responsible for a given prediction of the system.

The user can now specify a list of sensitive variables using a novel `set_sensitive_variables()` method of the FanFAIR object (we will denote by \mathcal{S} the list of sensitive variables). So doing, FanFAIR will automatically calculate the Pearson’s correlation coefficient between each input variable $x \in \mathcal{S}$, and the output variable y^1 , which is computed as follows:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (1)$$

where \bar{x} and \bar{y} denote the sample mean of the sensitive variable and the output variable, respectively. Pearson’s correlation is symmetric, so that $r_{xy} = r_{yx}$. The possible values of r_{xy}

¹Please note that FanFAIR is currently limited to datasets related to classification problems. Hence, it can be used for regression problems only if the output value is properly converted, e.g., by means of binning.

range between -1 and $+1$, representing negative and positive correlation, respectively. Since we are only interested in detecting correlation, we will discard the sign by considering the absolute value, i.e., $R_{xy} = |r_{xy}|$. Finally, we will assess the fairness with respect to sensitive input variables of the dataset as:

$$\rho = \max_{x \in \mathcal{S}} (R_{xy}). \quad (2)$$

The rationale of Equation 2 is that a high correlation of the output to even a single variable marked as sensitive (e.g., gender, ethnicity, age, political orientation) is enough to classify the whole dataset as unfair. In such a case, the removal of the variable from the dataset before the training of the model is strongly advised.

2) *Improved Isolation Forest*: We integrated an improved implementation of Isolation Forest [13], [14], which enables FanFAIR to perform multivariate outlier detection leveraging any combination of numeric, boolean, or categorical variables, and also providing support for data with missing values. The underlying implementation is provided by the Python module “isotree”², which was added to FanFAIR’s dependencies. This new method may be used by setting `outliers_detection_method="isotree"` during the creation of a FanFAIR object.

Internally, a score between 0 and 1 is computed for each row in the data, with higher scores indicating that the corresponding sample presents a combination of values that is more unusual than those seen in the rest of the data. We then determine the outlier status O_i of each row by performing the following computation:

$$O_i = \begin{cases} \text{TRUE} & \text{if } o_i > \min(0.7, \mu_o + 3\sigma_o) \\ \text{FALSE} & \text{otherwise} \end{cases} \quad (3)$$

where o_i is the score associated to the i -th row, and μ_o and σ_o are the mean and standard deviation of the score across the dataset, respectively. We opted to use the dynamic thresholding detailed in equation (3), rather than applying a fixed threshold, to adapt the sensitivity of the method to the sparsity of the available data. Furthermore, we employ the default parameters for the detector (500 trees; 1 randomly selected variable per split; maximum tree depth = height of balanced binary tree with a number of leaves equal to the samples), which we reckon should perform well in most cases.

3) *Extended support for datatypes and dataframes*: It is now possible to feed a Pandas [15] dataframe directly to FanFAIR through the `dataframe` parameter when creating a FanFAIR object. This enables an easier integration of FanFAIR into pipelines that adopt this very popular Python module, and, more crucially, grants the user control over the determination of variables’ datatypes, providing an alternative to automatic type inference carried on by the csv parser.

Concurrently, we provided FanFAIR with the ability to automatically select the appropriate variables during each

step of the computation of the fairness score, accordingly to the chosen parameters (e.g., during the determination of outliers, binary and categorical variables are only preserved when employing “isotree”, while dates are excluded in all cases), informing the user of the actions taken. These additions improve FanFAIR’s usability, and extend its applicability to a wider variety of datasets.

C. Case study: COVID dataset

The goal of this study is to investigate how data processing applied during the development of an AI system can affect the fairness of a dataset. We relied upon a recently published study [16], employing the dataset as a test case for the present study.

1) *Study and dataset overview*: The referenced work consists in a retrospective analysis, aimed at developing a ML algorithm for mortality prediction of hospitalized COVID-19 patients undergoing treatment with glucocorticoids (GCs), which could serve as a decision support for medical doctors. The initial dataset included 951 patients, and comprised the following 129 fields:

- **health history** (24) - collected at hospital admission, these include age, sex, body-mass index, and 21 risk factors (e.g., smoking habits, history of coronaropathy);
- **therapy** (20) - detailing which therapies were administered to the patient, including drugs (e.g., warfarin), and different types of ventilatory support (e.g., invasive mechanical ventilation (IMV));
- **complications** (21) - detailing complications that occurred during the treatment of the patients (e.g., acute renal failure);
- **serial measurements** (45) - these comprise (up to) five measurements for each of nine variables, including physiological quantities (e.g., C-Reactive Protein (CRP), Lymphocytes), and indicators of the status of pulmonary functions (e.g. arterial partial pressure of oxygen to fraction of inspired oxygen ratio (PaO₂/FiO₂));
- **dates** (18) - these include dates of birth, hospitalization, decease/discharge, sampling of serial values, and administration of some therapies;
- **outcome** (1) - binary indicator of patient’s death.

The ML task tackled in the study is an instance of binary classification problem, that is, the resulting algorithm predicts mortality of individual patients, according to the features collected during their stay at the hospital.

2) *Relevant stages of the data processing pipeline*: We used FanFAIR to evaluate four versions of our dataset, corresponding to different stages of the data processing pipeline adopted in the original study:

- 1) **raw** - (947 rows, 129 columns) imported from csv, it comprises all the available variables; minimal processing was applied, just to allow FanFAIR to evaluate the data, namely: datatype enforcement and automated removal of invalid entries of numeric variables;
- 2) **clean** - (825 rows, 79 columns) uninformative and highly sparse variables were removed, as well as rows

²The documentation, complete with the list of literature referenced by the developer, is available online on the original author’s GitHub repository.

pertaining patients not belonging to the population of interest (i.e., those that were not treated with GCs), or presenting clearly anomalous values. Only two out of the nine serially measured variables available were preserved (i.e. PaO₂/FiO₂ ratio and CRP level);

- 3) **reseried** - (825 rows, 57 columns) dates are removed, raw serial measurements are condensed into three fields per series, namely, first and last sample, and a binary "improving" field computed from the raw readings according to heuristics designed by medical experts;
- 4) **selected** - (825 rows, 10 columns) it only includes the 9 predictors determined during the variable selection phase of the processing pipeline adopted in the original study, and the outcome.

3) *Quality, compliance and sensitive variables*: In order to evaluate our dataset(s) with FanFAIR, we need to assess the appropriate values for the `quality` and `compliance` parameters, as well as identify sensible variables in our data (as per the definition given in section II-B1).

Concerning the quality, a fair assessment should be based on the reliability of the data collection and on the professional opinion of domain experts. In our case, the data collection consisted in the manual annotation of the values by clinical doctors over multiple sessions for each patient, and the final manual assembly of the collected data into a single Excel dataset. Human errors might occur at different steps of this articulated process, resulting in noisy data values. As a matter of fact, we discovered evidence of such errors during an extensive quality checking of the date variables in our dataset. The domain experts deliberated that a penalty of 0.1 would be sufficient to account for this noise. Additional source of noise, due to unforeseeable factors, should be accounted too. For these reasons, we decided to use an overall penalty value of 0.2. We consider the final value of 0.8 a very conservative estimate of the quality of the least processed version of our dataset (i.e., raw). In addition, it would be reasonable to assume that subsequent processing steps improved the quality of the data; nonetheless, we deemed it safer to adopt the same quality value across all the dataset, to prevent any artificial inflation of the final fairness score.

Compliance assessment was rather straightforward, since some of the authors of the present work were directly involved in the original studies that produced the dataset [17], [18]: the clinical and medical data were collected in compliance with legal and ethical standards. Specifically, the data were pseudonymized, handled in accordance with transparency obligations and the rights of the individuals, and the appropriate legal compliance was performed. Formal authorization from the Central Ethical Committee and informed patient consent were obtained prior to data collection. All relevant clinical and diagnostic regulations were followed. Lastly, principles of non-discrimination, fairness, and other standard ethical guidelines were adhered to in the collection, storage, and use of the patient data. On the basis of these considerations, we determined that our dataset meets all five compliance criteria

considered by FanFAIR, namely: *data protection, copyright, medical, non discrimination and ethics*.

Finally, we identified age and sex (referring to the assigned sex at birth, not the gender identity of the patient) as the only sensible variables in our dataset; the latter in particular was excluded during the variable selection phase of the original study, therefore it does not appear in the final instance of the dataset. It is worth noting that we do expect to detect some degree of correlation between age and mortality in our data, given that the results of the previous study indeed point to age as the second most influential variable, in terms of mean absolute SHAP value associated to the final model.

4) *Training and evaluating ML models*: To show how model performance and fairness evolve in parallel, we train a Random Forest (RF) classifier on each dataset, and estimate its generalization performances in terms of classic ML metrics. We adopt a very basic pipeline to train and evaluate the models, designed to ensure that both technical and problem-specific prerequisites are met in all four instances of the classification task (briefly summarized in Section II-C1). Specifically, the pipeline articulates as follows:

- 1) task-specific selection - patients that did not undergo GCs treatment are excluded (122, only in raw); variables that would make the prediction task trivial, according to our medical experts, are removed (2, in all datasets except selected);
- 2) technical preprocessing - date columns are dropped; categorical values are replaced with the respective numeric codes; missing values are filled with the median (or modal, in case of categorical) values of the respective variables;
- 3) model training and evaluation - a vanilla `RandomForestClassifier` (scikit-learn v1.5.1 [19]) is trained and evaluated adopting a 5-fold cross-validation scheme with stratification, to preserve outcome proportions.

Hyper-parameter tuning was skipped, as the achievement of optimal performances was not within the scope of the present work. It's also worth noting that, since the variables present in the last dataset (i.e., "selected") were chosen accordingly to a different pipeline, we do not expect that models trained on it according to the present pipeline will necessarily achieve the best overall performance scores.

III. RESULTS

We applied FanFAIR to the four versions of the COVID datasets described in Section II-C. We set "age" and "sex" as sensitive variables. The fairness scores calculated by FanFAIR are the following:

- raw : 75.9%
- clean : 82.3%
- reseried : 83.8%
- selected : 84.3%

These values show a clear improving trend in accordance with the progression of the data processing pipeline.

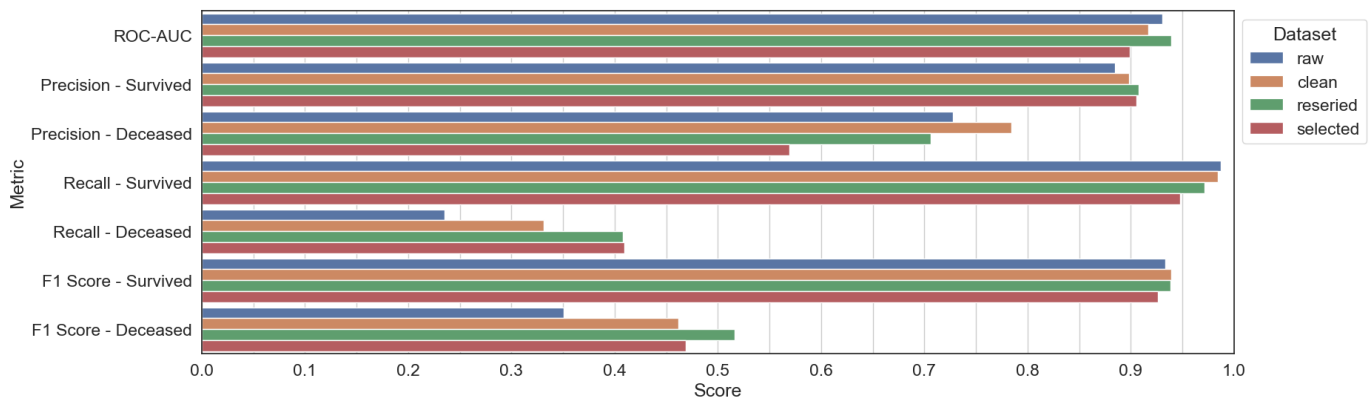


Fig. 1. Overview of performance metrics of the Random Forest models fitted on each dataset.

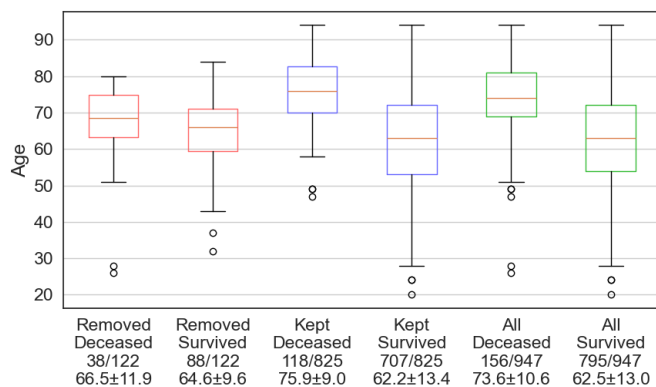


Fig. 2. Box-plots of age, stratified by outcome, in the group of patients excluded during the first data processing step (left) against the rest (center), and in the entire raw dataset (right); labels indicate numerosity, and age mean ± standard deviation for each group.

Concurrently, performance metrics (see Figure 1) indicate that models trained on more processed versions of the dataset are better at identifying patients who did not survive, which arguably constitutes the most crucial aspect of the task. Specifically, recall on deceased patients using the “selected” dataset almost doubles with respect to the “raw” dataset (from 23.6% to 40.9%), and likewise the F1 score on deceased patients improved by more than 10 percentile points (from 35.1% in “raw” to 46.9% in “selected”, and up to 51.7% in “reseried”); the other metrics are only slightly affected, with the sole exception of precision for deceased, which dropped from 72.8% in “raw” to 56.9% in “selected”, while still resulting improved with respect to the baseline in “clean” (78.4%).

Given the nature of the task and the strong unbalance in the data, we believe that these results support the hypothesis that adequate pre-processing of the data is not only necessary to improve the performance of ML solutions, but also constitutes a valid means to reduce the unfairness embedded in the data. In this specific case, the most substantial contribution to the fairness score (+6.4%) was achieved in the first step of the

data processing pipeline, which consisted in the exclusion of variables presenting a high level of sparsity (i.e., with 50% or more undefined values), and the restriction of the cohort to patients within the population of interest.

Concerning the analysis of sensitive variables, FanFAIR reports a noteworthy level of correlation between age and mortality (32% in raw, increased to 35% in all other datasets), although domain experts are keen to consider this as a manifestation of the well known “age pattern of mortality”, rather than the evidence of the unfair administration of treatments at the expense of elderly patients, especially considering that half of the cohort belongs to this category (median age is 65).

The 3% increase in the correlation can be explained by the fact that patients that were dropped from the dataset in the first step lowered the mean age among deceased in the first dataset (Figure 2). Furthermore, FanFAIR reports negligible correlation between sex and outcome (2% in “raw”, 0% in “clean” and “reseried”).

A further inspection of the membership functions reveals that an important issue of the dataset is the balance of the labels, which was assessed around 50% (see Figure 3, top-left panel). All other variables do not seem to have a negative impact to the fairness, with the exception of quality which could be slightly improved, e.g., by automating (parts of) the data collection procedure. The numerosity, according to FanFAIR, was excellent to develop a fairer predictive model.

For this dataset, the overall improvement of fairness due to pre-processing was limited (< 10%). Although this may not necessarily be the case for every dataset, this result is in agreement with the idea, discussed in this work, that the fairness of data (and hence, of the AI systems trained on them) is dependent also on additional factors that must be taken into account since the beginning of the study. These factors include (but are not limited to) a proper design of the data collection phase, and a careful review and fulfillment of all legal requirements that apply to research, at both general and case-specific levels.

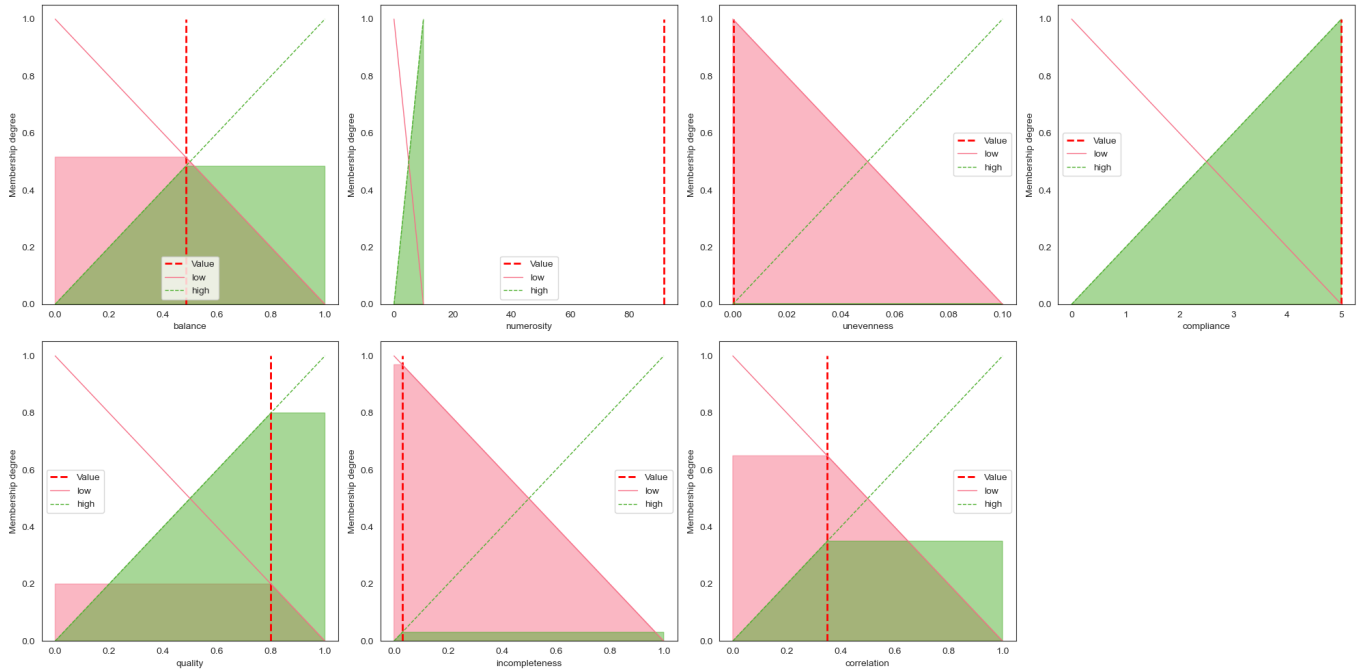


Fig. 3. Membership values calculated for the fully processed version (i.e., “selected”) of the COVID dataset.

IV. CONCLUSION

In this work we presented an improvement to FanFAIR, a Python library that leverages fuzzy reasoning for the semi-automatic assessment of datasets fairness. Specifically, we extended FanFAIR to support the specification of one or more sensitive variables. A correlation check is automatically performed on all sensitive variables and that information is later fed to a fuzzy rule, which determines whether the sensitive variables have an excessive influence on the output labels. This preliminary analysis can help to identify potential bias in the data that can be traced back to even a single variable. We also implemented a few missing but useful functionalities, e.g., the import of Pandas’ dataframes.

FanFAIR was applied to the analysis of a COVID dataset. Our results showed that both data fairness and model performance can improve in parallel when appropriate data processing is set in place during the development of AI systems. In this regard, FanFAIR can help researchers in deciding whether a given dataset should be pre-processed, or discarded altogether, due to its limitations.

From a computational standpoint, we report that FanFAIR was able to process our tabular dataset in under a minute on a consumer laptop equipped with a CPU Intel i3 of 12th generation, which is arguably efficient for any similar use case. Early testing performed during our study showed that the newly introduced multi-variate out-lier detection methods³ is more computationally demanding than the available uni-variate options, which could be expected. We plan to further inquire

³Benchmarks for the isotree and PyOD Python modules are available at the respective homepages.

on the complexity and scalability of FanFAIR by systematic benchmarking and testing, as this would provide the insight necessary to extend our tool’s applicability and reliability.

It is important to highlight that FanFAIR is not a debiasing algorithm, but rather a means to help healthcare workers and data scientists to pre-evaluate the dataset they intend to use to train a ML model, as we demonstrated by applying it to a study case on COVID-19 data. In addition, FanFAIR does not claim to address all potential biases related to AI fairness, as many discriminatory factors are systemic and embedded in society, such as the unavailability of healthcare for certain marginalized groups and their consequent absence in the dataset. Our tool aims to provide a way to assess the dataset so that the risk of unfairness is decreased.

It is often the case that a combination of variables might contribute to a discriminatory prediction. We will extend this feature in future versions of FanFAIR with the possibility to perform multi-variate influence analysis of sensitive variables and also to perform *post-hoc* analysis by leveraging user-provided ML models trained on the data.

FanFAIR is designed to be as intuitive as possible, with a minimal interface, and the capability to autonomously determine the most appropriate values for the parameters that regulate its functioning. The rationale is that our tool is designed to be used by anyone, including professionals who are not experts in the fields of machine learning and data science. Nevertheless, we understand that many practitioners might want to tweak and configure some of its internal settings (e.g., the hyper-parameters of the outlier detection algorithms). As future developments, we will provide FanFAIR with additional (optional) arguments to properly choose such settings.

REFERENCES

- [1] C. Sessa, C. Gallese, F. Schettini, D. Bellavia, F. Asperti, and E. Falletti, "Identifying bias in data collection: A case study on drugs distribution," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–10.
- [2] A. Balayn and S. Gürses, "Beyond debiasing: Regulating AI and its inequalities," *EDRI report*, 2021.
- [3] S. Milan and E. Treré, "Big data from the south (s): Beyond data universalism," *Television & New Media*, vol. 20, no. 4, pp. 319–335, 2019.
- [4] T. Scantamburlo, P. Falcarin, A. Veneri, A. Fabris, C. Gallese, V. Billa, F. Rotolo, and F. Marcuzzi, "Software systems compliance with the AI Act: Lessons learned from an international challenge," in *Proceedings of the 2nd International Workshop on Responsible AI Engineering*, 2024, pp. 44–51.
- [5] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 501–512.
- [6] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1–20, 2017.
- [7] R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang, "Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?" in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 325–336.
- [8] D. Thakkar, A. Ismail, P. Kumar, A. Hanna, N. Sambasivan, and N. Kumar, "When is machine learning data good?: Valuing in public health datafication," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–16.
- [9] C. Gallese, C. Fuchs, S. G. Riva, E. Foglia, F. Schettini, L. Ferrario, E. Falletti, and M. S. Nobile, "Predicting and characterizing legal claims of hospitals with computational intelligence: the legal and ethical implications," in *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2022, pp. 1–9.
- [10] C. Gallese, "Legal aspects of AI in the biomedical field. the role of interpretable models," *Big Data Analysis and Artificial Intelligence for Medical Sciences*, 2024.
- [11] E. R. Goffi, L. Colin, and S. Belouali, "Ethical Assessment of AI Cannot Ignore Cultural Pluralism: A Call for Broader Perspective on AI Ethic," *Arribat-International Journal of Human Rights Published by CNDH Morocco*, vol. 1, no. 2, pp. 151–175, 2021.
- [12] C. Gallese, T. Scantamburlo, L. Manzoni, and M. S. Nobile, "Investigating semi-automatic assessment of data sets fairness by means of fuzzy logic," in *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2023, pp. 1–10.
- [13] W. S. Al Farizi, I. Hidayah, and M. N. Rizal, "Isolation forest based anomaly detection: A systematic literature review," in *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*. IEEE, 2021, pp. 118–122.
- [14] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended isolation forest," *IEEE transactions on knowledge and data engineering*, vol. 33, no. 4, pp. 1479–1489, 2019.
- [15] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>
- [16] F. Salton, M. Rispoli, P. Confalonieri, A. De Nes, E. Spagnol, A. Salotti, B. Ruaro, S. Harari, A. Rocca, A. d'Onofrio *et al.*, "A tailored machine learning approach for mortality prediction in severe COVID-19 treated with glucocorticoids," *The International Journal of Tuberculosis and Lung Disease*, vol. 28, no. 9, pp. 439–445, 2024.
- [17] F. Salton, P. Confalonieri, S. Centanni, M. Mondoni, N. Petrosillo, P. Bonfanti, G. Lapadula, D. Lacedonia, A. Voza, N. Carpenè *et al.*, "Prolonged higher dose methylprednisolone versus conventional dexamethasone in COVID-19 pneumonia: a randomised controlled trial (MEDEAS)," *European Respiratory Journal*, vol. 61, no. 4, 2023.
- [18] F. Salton, P. Confalonieri, G. U. Meduri, P. Santus, S. Harari, R. Scala, S. Lanini, V. Vertui, T. Oggionni, A. Caminati *et al.*, "Prolonged low-dose methylprednisolone in patients with severe COVID-19 pneumonia," in *Open forum infectious diseases*, vol. 7, no. 10. Oxford University Press US, 2020, p. ofaa421.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.