Mitigating Unfairness in Machine Learning: A Taxonomy and an Evaluation Pipeline

Discussion Paper

Chiara Criscuolo[1,*], Tommaso Dolci[1] and Mattia Salnitri[1]

[1]*Politecnico di Milano – Department of Electronics, Information and Bioengineering*

Abstract

Big data poses challenges in maintaining ethical standards for reliable outcomes in machine learning. Data that inaccurately represent populations may result in biased algorithmic models, whose application leads to unfair decisions in delicate fields such as medicine and industry. To address this issue, many fairness mitigation techniques have been introduced, but the proliferation of overlapping methods complicates decision-making for data scientists. This paper proposes a taxonomy to organize these techniques and a pipeline for their evaluation, supporting practitioners in selecting the most suitable ones. The taxonomy classifies and describes techniques qualitatively, while the pipeline offers a quantitative framework for evaluation and comparison. The proposed approach supports data scientists in addressing biased models and data effectively.

Keywords

fairness, mitigation, machine learning, taxonomy, pipeline

1. Introduction

One of the challenges of big data is assuring high ethical standards to obtain *reliable and high-quality results* when machine learning algorithms are employed. Data that do not correctly represent the population sooner or later lead to biased machine learning models and wrong outcomes, with possible severe impacts on people and society. For example, a biased model for evidence-based medicine may lead to wrong diagnoses, or a biased model for the industry might lead to wrong business decisions. In both cases, consequences will be drastic, potentially leading to life-threatening situations in the former case and to business failures in the latter. To address this problem, *fairness mitigation techniques* can be used, to either modify the analyzed data or tune the model, with the goal of reducing or removing bias. Because of the importance of the issue, mitigation techniques are proliferating in the literature, frequently generating overlapping techniques that make the choice of the the data scientist extremely difficult.

Therefore, we propose a **taxonomy** to organize the fairness mitigation techniques that can be used to mitigate bias. Such a taxonomy will help data scientists navigate through the different mitigation techniques and easily identify the ones that can be applied to the specific context. To further help the selection of mitigation techniques, this paper proposes a **pipeline** for their

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.

chiara.criscuolo@polimi.it (C. Criscuolo); tommaso.dolci@polimi.it (T. Dolci); mattia.salnitri@polimi.it (M. Salnitri)

evaluation, whose objective is to be used for the creation of a shared repository of evaluations that can grow incrementally and become a reference knowledge-base.

The rest of the paper is structured as follows. Section 2 reports the theoretical foundations of the research work, and Section 3 describes the state of the art on taxonomies for unfairness mitigation techniques and pipelines for their evaluation. Section 4 introduces the proposed taxonomy, with an example of a structured description focused on one unfairness mitigation technique, while Section 5 illustrates the evaluation pipeline. Finally, Section 6 concludes the paper and discusses future works.

2. Preliminaries

Fairness is "the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" [1, p.100]. It is based on the idea of *protected* or *sensitive attribute*. A protected attribute is a characteristic for which non-discrimination should be established, such as religion, race, sex, and so on [2]. A protected group is a set of individuals identified by the same value of a protected attribute (e.g., females, young people, Hispanic people). Fairness can be measured with different metrics [2, 3, 4], each one expressing a specific aspect of fairness. Fairness can be considered satisfied based on the values of the aforementioned metrics.

The process to identify, measure and mitigate unfair behaviors of machine learning models is composed of three steps [5]: *(i)* train a classification algorithm to predict the binary value of the target class, that can be a positive or a negative outcome like not obtaining a loan or having a low income; *(ii)* use fairness metrics to understand whether the prediction of this model encompasses discrimination for the protected group; *(iii)* if the metrics results show unfair behavior, we conclude that the model learned the bias from the dataset, and a fairness mitigation technique is applied.

A fairness mitigation technique falls into one of the following three main categories [6, 7]:

- **Pre-processing**: tackle the fairness problem by removing the underlying discrimination from the training dataset, before the model is trained;
- **In-processing**: modify state-of-art algorithms to address fairness during the learning phase;
- **Post-processing**: act after the training phase has ended, by transforming the model output to satisfy fairness.

In the rest of the paper, we adopt the following terminology:

- D: the original dataset composed by N tuples;
- c: the trained algorithm classifier;
- g: the protected attribute;
- Y: the actual classification result, two possible values (or labels);
- $\hat{Y}$: the algorithm predicted decision;
- D', $\hat{Y}'$ and c' are the corresponding values after the application of a mitigation technique.

3. Related Work

In the recent years, numerous researchers have worked to provide a comprehensive overview of unfairness mitigation techniques in computer science and machine learning.

In [8] the authors present the various types of bias with examples of real-world applications. They propose a taxonomy for fairness definitions from the literature to address bias in machine-learning systems, categorizing them based on their downstream application (i.e., classification, regression, and other machine-learning methods). In [9] the authors give an overview of the main concepts of fairness, considering identification, measurement, and improvement of fairness, focusing solely on classification tasks. They briefly present various fairness-enhancing algorithms dividing them into pre-, in-, and post-processing techniques. A similar approach is adopted in [7], with the difference that pre-, in-, and post-processing algorithms are further sub-categorized into areas to provide a deeper insight. In [10], bias and unfairness are analyzed in various domains, primarily focusing on data management approaches to address unfairness. They cover the definitions of fairness metrics, how to identify unfairness, and mitigation techniques. The author suggest a shift towards a data-centered approach, by integrating bias constraints and curation methods into database management systems, to discover and mitigate bias as early as possible. In [11], the authors provide an overview of bias-mitigation research across various algorithmic systems, focusing on information retrieval, human-computer interaction, recommender systems, and machine learning. Additionally, they discuss the subjective nature of perceived fairness and the importance of explainability in managing bias. The survey identifies two types of attributes affected by bias: those describing the world and those describing information, suggesting different approaches for mitigation.

What emerges from this work is the need for an *up-to-date and comprehensive taxonomy* that presents and describes unfairness mitigation techniques, highlighting, through specific selection criteria, which ones are *available off-the-shelf and readily usable*. Our taxonomy addresses this gap, focusing on the qualitative aspects necessary for understanding and choosing the right technique to apply.

Finally, in [12] the authors present a data pre-processing pipeline to facilitate a direct comparison of mitigation techniques: they suggest a standard dataset pre-processing method and study the trade-off evaluation between fairness and accuracy. Our evaluation pipeline differs from this approach in comparing, through fairness metrics, the *impact of each mitigation strategy*.

4. Taxonomy of Mitigation Techniques

The first part of this section presents the taxonomy structure, the selection criteria, and its overall composition. The second part describes in detail the taxonomy elements. Finally, we present a structured description of a specific mitigation technique included in the taxonomy.

4.1. Taxonomy Definition

We propose a taxonomy that classifies unfairness mitigation techniques based on the targeted processing phase of machine learning models. Each technique is classified into pre-, in- and post-processing classes. The usefulness of the proposed taxonomy is based on the amount of

mitigation techniques that are included. The taxonomy can be seen as the result of systematic analyses of the state of the art, that can be cyclically performed by data scientists to include more and more techniques. We, therefore, envision the taxonomy to be constantly updated and extended. As a result, we designed it to be easily extensible, by following the classification proposed and including the required information for the structured description of the technique.

The choice of the mitigation techniques that are included in the taxonomy is based on the following selection criteria.

- **Applicability**: the mitigation technique should be usable off-the-shelf, with parametric customization only. Many prototypical mitigation techniques are available, yet their application is limited to specific cases, or they need heavy customization. Such an effort is not feasible for data scientists in most scenarios.
- **Tabular data**: tabular data is the most used type of data, especially in data science applications. The selected mitigation techniques are required to operate on tabular data. In some cases (e.g., waveforms) unstructured data can be transformed into tabular data.
- **Binary classification**: the selected mitigation techniques are required to mitigate bias for binary classification models. In particular, we restrict the scope to techniques that apply to prescriptive models, and guarantee fair, unbiased predictions.

Each element of the taxonomy is a mitigation technique, that is detailed using a predefined description structure that will guide the inclusion of new mitigation techniques, and ease the analysis of the taxonomy and of the included techniques.

General Description. This is a general description of the mitigation technique, that is used to provide the data scientists with the context of application of the technique.

Objectives. This field defines the objective and the scope of the mitigation technique. We describe and summarize the main steps of each technique, to make the technique more understandable.

Input Required. This specifies the type of input required by the mitigation technique: some techniques require specific information or pre-processing steps to make the dataset compliant with the requirements of the mitigation technique, e.g., if the dataset contains only categorical values. This field specifies the pre-processing actions possibly required by the technique.

Expected Output. The expected results are specified in this field. For instance, some techniques return a modified version D' of the input dataset, others an improved classifier c'.

Parameters. This field contains the number and types of parameters that need be specified for the mitigation techniques. A short description of each parameter is provided, including procedures for the identification of the best parameter values and instructions on fine-tuning. If the parameters are context-dependent, this is also reported.

Issues and Limitations. This field specifies the known limitations of the mitigation technique and also possible issues experienced in its application. Possible issues relate to: *(i)* machine learning support, i.e., any constraints with machine learning models, cross-validation or hold-out support; *(ii)* required post-processing of the mitigation output to compare it with the original dataset; *(iii)* the possibility of retrieving (or not) the modified dataset (if any).

Demonstration of Application. This provides a link to a working demonstration, if any. A working demonstration is central to understand how the technique is applied and can operate.

Table 1
This table reports the taxonomy composed by 12 mitigation techniques classified into the three categories: pre-processing, in-processing and post-processing. For each technique, we describe input, output, corresponding reference, i.e., the original paper in which the technique was presented, and the tool that implements the technique.

Category	Technique	Input	Output	Ref.	Tool
Pre-proc.	Reweighing	D, g, Y	D', W	[13]	AIF360
	Disparate Impact Remover	D, g, Y	D'	[14]	AIF360
	Learning Fair Representation	D, g, Y	D'	[15]	AIF360
	Optimized Pre-processing	D, g, Y	D'	[16]	AIF360
	Correlation Remover	D, g, Y	D'	[17]	FL
In-proc.	Adversarial Debiasing	D, g, Y	c'	[18]	AIF360, FL
	Prejudice Remover	D, g, Y	c'	[19]	AIF360
	Reductions Exponential Gradient	D, g, Y	c'	[20]	FL, AIF360
Post-proc.	Calibrated Equalized Post-Processing	$D, g, Y, \hat{Y}$	$\hat{Y}'$	[21]	AIF360
	Reject Option Classification	$D, g, Y, \hat{Y}$	$\hat{Y}'$	[22]	AIF360
	Equalized Odds Post-Processing	$D, g, Y, \hat{Y}$	$\hat{Y}'$	[21, 23]	FL, AIF360
	Threshold Optimizer	$D, g, Y, \hat{Y}$	$\hat{Y}'$	[23]	FL

Source Code. This specifies the link to the source code, if available.
Documentation. This specifies the link to the official documentation or any other useful documentation of the technique.

4.2. An Overview of the Proposed Taxonomy

Table 1 shows a general overview of the taxonomy, considering some mitigation techniques, presenting: the name of each technique, its category, the required input, and the expected output. The selected techniques satisfy all the criteria mentioned at the beginning of Section 4.1, i.e., the possibility of applying this mitigation technique to tabular data in a binary classification task. The two additional columns **Ref.** and **Tool** report the original paper in which the mitigation technique was first proposed, and which tool provides a library/source code available to use the specific mitigation technique. Two main tools are considered: *AIF360* [24], the acronym for *AI Fairness 360* tool implemented by IBM, and *FL* standing for *FairLearn* [17], implemented by Microsoft. Other state-of-the-art tools were discarded because they offered only methodologies for measuring fairness and not for its mitigation, such as *Aequitas* tool [3], *Google What-if* tool [25].

The following part reports, as an example, the structured description of *Reweighing* [13], a mitigation technique included in the taxonomy described in Table 1. This technique fits the selection criteria of the presented taxonomy: *(i)* it is applicable and off-the-shelf, *(ii)* it handles

tabular data, and *(iii)* it can be used to mitigate unfairness also for binary classification tasks.
General Description The intuition behind this strategy is that new *weights are assigned to data objects (tuples) to make the dataset discrimination-free.* [13, p.14].
Objectives The goal of *Reweighing* is to assign a weight to the tuple in the dataset, in a way that the new dataset $D^{'}$ is bias-free, i.e., g the protected attribute and Y the target variable, are statistically independent [13]. More specifically, the proposed equation is:

$$W(X) = \frac{P_{exp}(g = X(g) \wedge Y = X(Y))}{P_{obs}(g = X(g) \wedge Y = X(Y))} ,$$

where g is a protected attribute, Y is the target variable and X is a random unlabelled data object. If the expected probability is higher than the observed probability value, it shows the bias toward class, thus the technique assigns lower weights to individuals who have been deprived or favored. If we multiply the frequency of every tuple by its weight, the new dataset will be bias-free.
Required Input The required input is composed by the original dataset D, the protected attribute g, and the target class Y. There are no further requirements for the input.
Output Expected The expected output is composed of a mitigated dataset D' and a weight vector W.
Parameters This technique has no parameters to set.
Issues and Limitations The main issue of this technique is the output: the mitigated dataset D' is equal to the original one D, but every tuple will have a different weight based on the protected attribute and the classification label. To apply the technique, the researcher should use the weight vector W. This vector should be used in the training and testing phases of the new model, in order to apply W to D' to obtain a mitigated solution.
Demonstration of Application A demonstration of *Reweighing* can be found here: https://github.com/Trusted-AI/AIF360/blob/master/examples/demo_reweighing_preproc.ipynb.
Source Code The source code of *Reweighing* is available here: https://github.com/Trusted-AI/AIF360/blob/main/aif360/algorithms/preprocessing/reweighing.py.
Documentation The documentation of *Reweighing* is available here: https://aif360.readthedocs.io/en/stable/modules/generated/aif360.algorithms.preprocessing.Reweighing.html.

5. Evaluation Pipeline

This section presents an innovative approach to systematically test and compare various fairness mitigation techniques from the literature. The proposed solution is designed to work with any tabular dataset. Currently, research work of this kind focuses on a few standard datasets from the literature (e.g., Adult Census Income, German Credit Risk Score, Compas Recidivism Score), and the considered fairness measures are typically limited. For example, many papers exclusively employ mitigation techniques implemented by *AIF360*, evaluating the results using only the predefined fairness metrics provided by the tool. The pipeline proposed in this paper allows to perform an evaluation using a larger set of metrics, each one derived from a different fairness definition, not bounded to the specific technique under analysis.

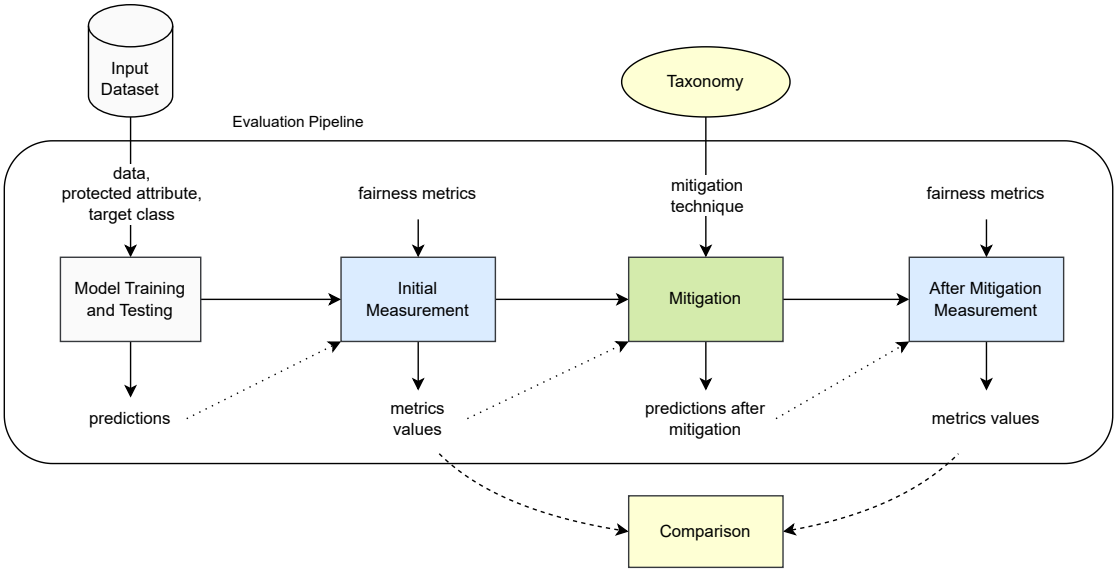

Figure 1: Proposed evaluation pipeline for testing and comparing fairness mitigation techniques.

Our pipeline stands out by considering a wide range of the most used fairness metrics, distinguishing this approach from available off-the-shelf tools. Similarly, we encompass all the mitigation techniques described in our taxonomy, not confining the evaluation to those offered by a single system. The overall goal is to fill a gap in the literature by providing a systematic evaluation of these techniques. This contribution is in addition to the systematic screening provided by the taxonomy, which is a key first step in better understanding unfairness mitigation techniques.

Figure 1 illustrates the proposed pipeline to quantitatively assess and compare the effectiveness of the mitigation techniques outlined in the taxonomy. Our pipeline, inspired by, and enriched from [5], starts with the input dataset D, in which a protected attribute g is identified. We assume the dataset to be pre-processed by the researcher, to ensure data quality.

The initial stage is *Model Training and Testing*, where a classification model c is trained on an input dataset D with target class Y. During testing, the model's predictions $\hat{Y}$ are computed. D, Y and $\hat{Y}$ represent the input to the *Initial Measurement* step to assess fairness. During this step, the identified fairness metrics are evaluated and analyzed. If any unfair behavior is detected, the *Mitigation* step is executed. Fairness mitigation techniques, selected on the basis of the taxonomy specified in Section 4, are applied. The result is a new set of predictions $\hat{Y}'$ for each technique applied. Finally, a second measurement step (*After Mitigation Measurement*) is executed to compute the new values for each fairness metric.

An analysis of the fairness metrics values before and after the mitigation step allows to compare the outcomes of the mitigation techniques. The analysis can be accompanied by the generation of plots and graphs. These visualizations provide researchers with a comprehensive understanding of the impact of each mitigation technique on the input dataset.

6. Conclusions and Future Work

This paper proposed an approach to help data scientists for the selection of the most suitable unfairness mitigation techniques when biased datasets are employed. In particular, we proposed two contributions: a taxonomy and an evaluation pipeline. The taxonomy allows the classification and qualitative structured description of mitigation techniques. The evaluation pipeline allows the quantitative evaluation and comparison of the effectiveness of the techniques. This paper lays the basis for a more extensive method for the evaluation of fairness mitigation techniques.

Future work consists in considering how the techniques' effectiveness is impacted by the type of data used and the type of machine learning model applied. The consideration of these variables is key for a deeper understanding of the efficacy of the fairness mitigation techniques. Additionally, the selection of the fairness metrics used in *Initial Measurement* and *After Mitigation Measurement* is central for a sound analysis. Unfortunately, off-the-shelf tools provide a limited, and heterogeneous, amount of metrics. To overcome this limit, we intend to implement a comprehensive set of metrics based on different fairness definitions.

Finally, the pipeline presented in this paper defines a general approach, and therefore can be extended and further detailed in each of the presented steps. For instance, the *Mitigation* step is a complex process that varies according to the mitigation technique applied, while the *Measurement* steps require a thorough analysis of each fairness metric. The *Comparison* step is critical since it requires a complex analysis by the data scientists. As a future work, all the mentioned steps can be improved to provide the scientists with: *(i)* support for the execution of the mitigation techniques, *(ii)* guidelines for the selection of the fairness metrics, and *(iii)* guidelines for the comparison of mitigation results. Overall, the evaluation pipeline can be automated for a faster and more effective measurement.

The contribution of this paper allows for an informed choice of fairness mitigation techniques. It will guide data scientists on the selection process obviating the existing theoretical and methodological gaps that, until today, lead to unacceptable errors and inaccuracies of machine learning models caused by biased datasets.

Acknowledgments

Thanks to professor Letizia Tanca for her contribution: her precious comments and feedback have significantly improved this research work.

References

[1] N. A. Saxena, et al., How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations, Artif. Intell. 283 (2020) 103238.

[2] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the FairWare@ICSE, 2018, pp. 1–7.

[3] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, R. Ghani, Aequitas: A Bias and Fairness Audit Toolkit, CoRR abs/1811.05577 (2018).

[4] L. Baresi, C. Criscuolo, C. Ghezzi, Understanding fairness requirements for ml-based software, in: IEEE RE, IEEE, 2023, pp. 341–346. doi:10.1109/RE57278.2023.00046.
[5] C. Criscuolo, T. Dolci, M. Salnitri, Towards assessing data bias in clinical trials, in: Heterogeneous Data Management, Polystores, and Analytics for Healthcare - VLDB Workshops, Poly 2022 and DMAH 2022, Springer, 2022, pp. 57–74. doi:10.1007/978-3-031-23905-2_5.
[6] B. d'Alessandro, C. O'Neil, T. LaGatta, Conscientious classification: A data scientist's guide to discrimination-aware classification, Big data 5 (2017) 120–134.
[7] S. Caton, C. Haas, Fairness in machine learning: A survey, ACM Comput. Surv. (2023). doi:10.1145/3616865, just accepted.
[8] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2022) 115:1–115:35.
[9] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (2022). doi:10.1145/3494672.
[10] A. Balayn, C. Lofi, G.-J. Houben, Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems, The VLDB Journal 30 (2021) 739–768.
[11] K. Orphanou, J. Otterbacher, S. Kleanthous, K. Batsuren, F. Giunchiglia, V. Bogina, A. S. Tal, A. Hartman, T. Kuflik, Mitigating bias in algorithmic systems—a fish-eye view, ACM Comput. Surv. 55 (2022). doi:10.1145/3527152.
[12] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of FAT, FAT* '19, ACM, 2019, p. 329–338. doi:10.1145/3287560.3287589.
[13] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (2012) 1–33.
[14] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the ACM SIGKDD, 2015, pp. 259–268.
[15] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: ICML, PMLR, 2013, pp. 325–333.
[16] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, NeurIPS 30 (2017).
[17] S. Bird, et al., Fairlearn: A toolkit for assessing and improving fairness in ai, Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).
[18] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the AAAI/ACM, 2018, pp. 335–340.
[19] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-aware classifier with prejudice remover regularizer, in: ECML PKDD. Part II 23, Springer, 2012, pp. 35–50.
[20] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: ICML, PMLR, 2018, pp. 60–69.
[21] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration, NeurIPS 30 (2017).
[22] F. Kamiran, A. Karim, X. Zhang, Decision theory for discrimination-aware classification, in: IEEE IICDM, IEEE, 2012, pp. 924–929.

[23] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, NeurIPS 29 (2016).
[24] R. K. Bellamy, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4–1.
[25] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (2019) 56–65.