

Reconciling Statistical and Causal Metrics of Fairness in Machine Learning on Data-Driven Systems

Chiara Criscuolo[†], Davide Martinenghi and Jing Huang

Politecnico di Milano - Department of Electronics, Information, and Bioengineering
Via G. Ponzio 34/5, 20133 Milano, Italy

Abstract

In the digital age, machine learning (ML) algorithms are becoming increasingly important in decision-making processes across a wide range of domains, including criminal justice, healthcare, and finance. While these algorithms provide significant benefits, they also pose the risk of perpetuating and exacerbating societal biases, especially when fairness is not taken into account during their design and implementation. We address the critical issue of fairness in machine learning, with a focus on combining statistical and causal fairness metrics to provide a more comprehensive approach to evaluate and ensure fairness by selecting the most suitable metric. To tackle this problem, we developed a research methodology aimed at systematically reviewing the existing literature while focusing on four research questions targeting the relationship between statistical and causal fairness metrics, which drove our analysis and categorization of papers. Based on the results of this review, we built a new fairness decision tree that integrates both types of metrics, which can guide users to choose the most suitable metric.

Keywords

Fairness, Statistical Fairness Metrics, Causal Fairness Metrics, Machine Learning

1. Introduction

In the digital age, machine learning (ML) systems have become essential to many aspects of daily living and social functions. Machine learning covers a wide range of critical applications, from criminal justice and credit scoring to healthcare diagnostics. However, a serious concern for justice has emerged along with the technological advancements that machine learning has brought forth. It is imperative to ensure that these systems do not perpetuate or exacerbate societal inequalities and biases that already exist [1, 2]. Machine learning algorithms, being data-driven, may inadvertently encode human bias. One striking example is the COMPAS Risk Assessment Tool [3] based on information about a defendant's criminal record, type of offense, record of contact with the community, and history of failing to appear in court to assist the judge in making bail decisions. Regarding this last aspect, the ProPublica team found that the software used by U.S. courts incorrectly labeled Black defendants as high-risk, almost twice as likely as White people [3]. Similar biases have been identified in other domains, such as e-commerce, where differentiated pricing strategies unfairly target returning customers based on their online behavior [4].

These examples show the need for a systematic approach to assessing and mitigating bias in machine learning systems. The problem's complexity is increased by the fact that fairness in ML can be understood from multiple perspectives, including statistical fairness, which focuses on ensuring equitable outcomes, and causal fairness, which aims to understand and address the underlying causal mechanisms that lead to biased results.

In this paper, we study the topic of fairness in machine learning, with a focus on the difference between statistical and causal fairness metrics; in particular, we study the possibility of incorporating them into a common vision. Our key contributions are as follows:

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

*Corresponding author.

✉ chiara.criscuolo@polimi.it (C. Criscuolo); davide.martinenghi@polimi.it (D. Martinenghi); jing1.huang@mail.polimi.it (J. Huang)

ORCID 0000-0002-1345-2482 (C. Criscuolo); 0000-0003-1542-0326 (D. Martinenghi)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- **Research Methodology:** We develop a systematic research methodology to identify and categorize the relevant literature.
- **Results of Literature Analysis:** Our systematic analysis of selected papers brings significant insights into the distinction between statistical and causal fairness, and the datasets most commonly used.
- **Fairness Decision Tree:** We present a new fairness decision tree framework that integrates both statistical and causal fairness metrics.

The rest of this paper is organized as follows. Section 2 introduces some preliminary concepts. Section 3 presents the adopted research methodology. Section 4 describes the analysis of the results. Section 5 presents the fairness decision tree. Section 6 concludes the paper.

2. Preliminaries

Fairness can be defined as the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics [5], and it holds significant relevance within the domain of Machine Learning (ML). In this context, ML algorithms embody a decision-making paradigm characterized by impartiality and the absence of bias. It is important to acknowledge that existing biases in the data can significantly influence the performance and outcomes of these algorithms, rendering the data and results unfair [1]. In terms of fairness, binary classification plays a crucial role in decision-making systems where outcomes significantly impact individuals, such as loan approvals, hiring decisions, and medical diagnoses. It maps input features to one of two possible outcomes: positive (1) or negative (0). These predictions are then compared to the actual outcomes to evaluate the model’s performance. Ensuring fairness in these models is essential to prevent discrimination [6].

Existing fairness definitions in ML algorithms can be classified into two categories: statistical fairness and causal fairness. Statistical fairness focuses on frequency statistics, ensuring equitable outcomes across demographic groups. In contrast, causal fairness explores causal relationships between attributes and outcomes, intervening to eliminate biases rooted in causal mechanisms. Statistical-based fairness metrics are categorized in [7, 8] and the Fairness Decision Tree presented by Baresi et al. [8] is designed to identifying the most suitable fairness interpretations for ML-based systems.

While statistics offer the tools to identify patterns and correlations within data [9], Judea Pearl’s work on causality challenges us to understand the “why” behind these patterns. This means that causal fairness differs from statistical fairness in that it is not entirely determined by observed data and necessitates the introduction of additional cause-and-effect assumptions. Causal fairness is a definition of fairness based on a causal connection between protected attributes and decisions. Causal graph models have limitations in their very structure, derived from domain knowledge, and inconsistencies in assumptions may occur [10]. Based on observed data, causal graph models often suffer from model non-uniqueness, which refers to the possibility that multiple different causal graph models can fit the same set of observed data equally well. This non-uniqueness implies that there may be multiple plausible explanations for the causal relationships in the data.

Based on Pearl’s structural causal models [11, 9], a structural equation-based mathematical object that describes the causal mechanisms of a system. Each causal model is associated with a causal graph for visualizing in a more user-friendly way the causal inference, where causal effects are carried by the causal paths that trace arrows pointing from the cause to the effect [12]. To better illustrate these notions, we introduce the Ladder of Causation taken from [13], a causal hierarchy presented by Pearl, which affirms that causation has three levels: association, intervention, and counterfactual.

1. **Association** [13] can be inferred directly from the observed data using conditional probabilities and conditional expectations, which correspond to statistically-based fairness metrics.
2. **Intervention** [13] involves not only seeing what is but also changing what we see. Interventional questions deal with expressions of the type $P(y|do(x), z)$, which denote “The probability of event $Y = y$, given that we intervene and set the value of X to x and subsequently observe event $Z = z$ ”.

It can be estimated experimentally from randomized trials or analytically using causal Bayesian networks.

3. **Counterfactual** [13] deals with expressions of the type $P(y_x|x', y')$ which denote "The probability that event $Y = y$ would be observed had X been x , given that we actually observed X to be x' and Y to be y' ". It can be computed only when the model is based on functional relations or is structural.

The majority of causal-based fairness notions are defined in terms of the non-observable quantities U of interventions and counterfactuals, so their applicability depends heavily on the identification of those quantities in the data. An overview of the principal causal-based fairness metrics is presented by [2].

1. **No unresolved discrimination** [7]: Requires that there exists no path from the protected attribute A to the predicted outcome \hat{Y} .
2. **Total Causal Effect** [11]: Is defined as the effect of changing the sensitive attribute A from $a = 0$ to $a = 1$ on decision $Y = y$ along all causal paths from A to Y , it is considered to be fair if the difference between the conditional distributions is within the fair threshold. It is defined as follows:

$$TE(y) = P(y|do(a = 1)) - P(y|do(a = 0))$$

3. **Path-specific Effect** [11]: Given a causal path set, the path-specific effect is defined as the value change of the sensitive attribute A from $a = 0$ to $a = 1$ on decision $Y = y$ along specific causal path π , it is considered to be fair if the difference is within the fair threshold. It is defined as follows:

$$PE_{\pi}(y) = P(y|do(a = 1|\pi, a = 0|\hat{\pi})) - P(y|do(a = 0))$$

where $P(y|do(a = 1|\pi, a = 0|\hat{\pi}))$ represents the post-intervention distribution of Y where the effect of intervention $do(a = 1)$ is transmitted only along π while the effect of reference intervention $do(a = 0)$ is transmitted along the other paths.

4. **No proxy discrimination** [14]: Requires there exists no path from the protected attribute A to the predicted outcome \hat{Y} that is blocked by a proxy variable R . It is defined as follows:

$$P(Y|do(R = r)) = P(Y|do(R = r')) \quad \forall r, r' \in dom(R)$$

5. **Counterfactual Fairness** [15]: Requires that the predicted outcome \hat{Y} in the graph does not depend on a descendant of the protected attribute A . This means that an outcome Y achieves counterfactual fairness towards an individual if the probability of $Y = y$ for such an individual is the same as the probability of $Y = y$ for the same individual but belongs to a different sensitive group. It is defined as follows:

$$P(y_{a=1}(U)|X = x, a = 0) = P(y_{a=0}(U)|X = x, a = 0)$$

Where X is the subset of observed variables O except sensitive variables and decision variables. Any context $X = x$ represents a certain sub-group of the population.

6. **Individual direct discrimination** [16, 2]: It aims to discover the direct discrimination at the individual level. It is based on situation testing, by comparing the individual with similar individuals from both groups (protected and unprotected). This means for a target individual i , select top-K individuals most similar to i from group $a = 1$, denoted as S^+ and top-K individuals most similar to i from group $a = 0$, denoted as S^- . The target individual is considered as discriminated if the difference observed between the rate of positive decisions in S^- and S^+ is higher than a predefined threshold (typically 5%).

Causal inference is used to define the distance function $d(i, i')$ to measure similarity between individuals: given a causal graph, only the variables that are direct parent nodes of the decision variable are considered to compute the similarity between individuals, which are denoted as $Q = P_a(Y) \setminus \{A\}$. The formal definition of $d(i, i')$ is:

$$d(i, i') = \sum_{k=1}^{|Q|} |CE(q_k, q'_k) * VD(q_k, q'_k)|$$

Where $VD(q_k, q'_k)$ is a distance function proposed by Luong et al. in [17] and $CE(q_k, q'_k)$ represents the causal effect of each of the selected variables $q_k \in Q$ on the actual outcome. In particular, $CE(q_k, q'_k)$ is defined as follows:

$$CE(y) = P(y|do(Q)) - P(y|do(q'_k, q \setminus q_k))$$

Where $P(y|do(Q))$ is the effect of the interventions that forces Q to take the set of values q , and $P(y|do(q'_k, q \setminus q_k))$ is the effect of the intervention that forces Q_k to take value q'_k and other attributes in Q to take the same values as q .

7. **Equality of Effort** [18]: It detects discrimination by comparing the effort required to reach the same level of the actual outcome of individuals from advantaged and disadvantaged groups who are similar to the target individual. A treatment variable T is selected and used to address the question: "To what extent the treatment variable T should change to make the individual (or a group of individuals) achieve a certain outcome level?"

Equality of effort notions are defined based on the potential outcome framework into individual γ -Equal effort and system γ -Equal effort. Both criteria can be used to measure the effort discrepancy between protected and unprotected groups.

Let's consider Y_i^t as the potential outcome for individual i had T been t , $E[Y_i^t]$ the expected outcome under treatment t for individual i , and consider, similar to *Individual direct discrimination*, S^+ and S^- as two sets of similar individuals of group $a = 0$ and $a = 1$ respectively.

In consequence, $E[Y_{S^+}^t]$ is the expected outcome under treatment t for the subgroup S^+ . And the needed minimal value of treatment variable T to achieve γ -level of outcome within the subgroup S^+ is defined as follows:

$$\Psi_{S^+}(\gamma) = \operatorname{argmin}_{t \in T} E[Y_{S^+}^t] \geq \gamma$$

For a certain outcome level γ , individual γ -Equal effort is satisfied for individual i if:

$$\Psi_{S^+}(\gamma) = \Psi_{S^-}(\gamma)$$

When S^+ and S^- are extended to the entire group with sensitive attribute $a = 0$ and $a = 1$ respectively, D^+ is used to denote the first set and D^- denoted the second one. The System γ -Equal effort is satisfied for a sub-population if:

$$\Psi_{D^+}(\gamma) = \Psi_{D^-}(\gamma)$$

8. **Path-specific Counterfactual Fairness (PC Fairness)**[11]: Is defined to cover various causality-based fairness notions. Given a factual condition $X = x$ where $X \subseteq O$ and causal path set π , a predictor \hat{Y} achieves the PC fairness is it satisfies the following expression:

$$(P(\hat{y}_{a=1|\pi, a=0|\hat{\pi}}|X) - P(\hat{y}_{a=0}|X)) \leq \tau$$

Where τ is a predefined fairness threshold.

Consequently, for example, if we set π contains all causal paths and X an empty set, PC-Fairness corresponds to the *Total Causal Effect*.

Finally, regarding unfairness mitigation, depending on the stage of the ML algorithm, pre-processing, in-processing, and post-processing mechanisms can be used to intervene in the algorithm to achieve fair ML, respectively.

3. Research Methodology

This section defines the research questions and the methodological approach taken to address them, including the specific search techniques and keywords that were employed to locate pertinent material, discussing the inclusion and exclusion criteria employed to identify the most relevant studies.

3.1. Research Questions

The methodology is based on a set of structured research questions designed to explore fairness in ML, comprehending both statistical and causal dimensions. These questions guide the entire research process, from the initial literature review to the final analysis and synthesis. Here, we discuss each of these questions.

RQ1: What are the main concepts and differences between statistical-based and causal-based fairness?

This question is to provide a clear and simple review of the fundamental ideas and differences between statistical and causal approaches to fairness in ML. Understanding these differences is important for comprehending each perspective's specific advantages and limits.

RQ2: Which datasets are most commonly used in fairness research, and are there differences between those used in causal-based and statistical-based studies?

Identifying commonly used datasets in both causal and statistical fairness is crucial as it helps in understanding the contexts in which fairness metrics are tested and validated. This question attempts to figure out not just the most often used datasets in general, but also if there are differences in dataset utilization across causal and statistical fairness studies.

RQ3: Is it possible to have a common vision between causal and statistical fairness?

This question investigates the theoretical feasibility and applicability of combining causal-based and statistical-based fairness metrics.

RQ4: How to choose the most suitable metric considering both perspectives?

This final question discusses the research's practical consequences, suggesting an approach for selecting the most appropriate fairness metric while balancing causal and statistical aspects.

By addressing these questions, this work provides the foundation for a detailed investigation of fairness research, ensuring an in-depth knowledge of both the theoretical foundations and practical applications of the subject.

3.2. Search Strategy

We delineate the comprehensive search strategy employed to gather relevant literature, focusing on evaluating fairness in ML with a particular emphasis on fairness metrics.

The literature collection process is illustrated in Figure 1 and which delineates the steps from research question formulation to final paper selection that are: Keywords and Query Creation, Setting Database and Inclusion and Exclusion Criteria, Paper Analysis and Snowballing.

To ensure targeted and relevant database searches, keywords were identified from the research questions. These keywords, along with supportive terms, guide the search process effectively and are: *AI, Machine Learning, ML, Data, Fairness, Metric, Definition, Solution, Mitigation, Ethic, Measure, Causal, Statistical, Group, Individual, Counterfactual, Interventional*.

The query was constructed using boolean operators to refine search results and ensure relevance to our objectives. Furthermore, the search was conducted across prominent databases including Scopus, ACM Digital Library, IEEE Xplore, and Google Scholar, focusing on publications from the past five years in the field of computer science. The final query is the following:

```
(fair* OR discriminat* OR unfair* OR bias*) AND (causal* OR statistic* OR individual* OR group* OR counterfact* OR intervention* OR parity*) AND (metric* OR measur* OR defin* OR solut* OR mitigat*) AND ((machine learning) OR data* OR ethic* OR (artificial intelligence))
```

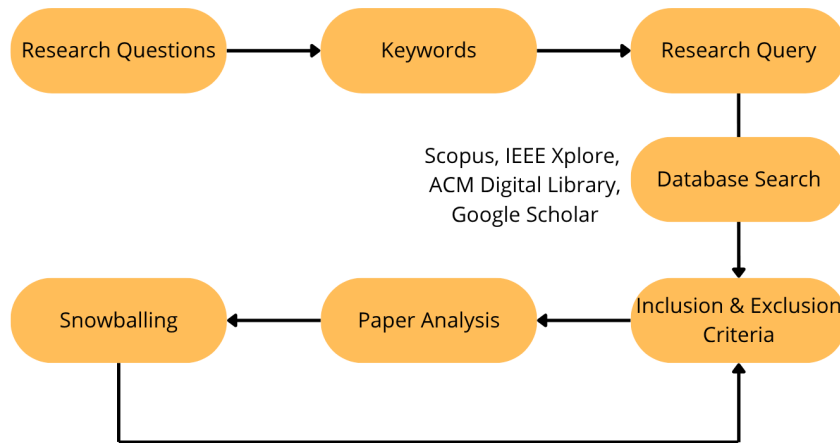


Figure 1: Literature collection process

The search provided a significant number of results: Scopus 5,270, ACM Digital Library 910, IEEE Xplore 2,335, Google Scholar 16,800.

For the selection strategy, we present the inclusion and exclusion criteria utilized during the selection process for identifying relevant literature. These criteria serve as guidelines to ensure the systematic and targeted inclusion of papers that align with our objectives while excluding those that do not meet the specified thematic requirements. The **Inclusion Criteria** are:

- **Article or Paper:** Articles, conference papers, or research papers that contribute to the discourse on fairness in ML.
- **More than 5 citations:** Papers cited more than five times, indicating the paper's impact and relevance.
- **Discussion on fairness:** This includes any study that addresses fairness in the context of ML, covering both theoretical and practical aspects.
- **New mitigation technique:** Studies that suggest methods to reduce or eliminate biases in ML models.
- **Tool:** Studies that introduce software or tools designed to assess, measure, or enhance fairness in ML models.
- **New perspective:** Papers that provide innovative viewpoints or conceptual frameworks for understanding fairness.
- **New fairness metric:** Studies that develop and validate new metrics for evaluating fairness in ML.

The **Exclusion Criteria** are:

- **Theses or reports:** Academic theses and technical reports, as these often serve as preliminary or non-peer-reviewed documents.
- **Surveys:** Summary of existing research rather than contributing new findings or perspectives.
- **Papers that primarily focus on techniques without discussing fairness:** Studies focusing on techniques or algorithms without addressing their fairness implications.
- **Papers not written in English:** Studies written in languages other than English.

Through the selection process based on inclusion and exclusion criteria, a total of 26 papers were initially identified as relevant. Subsequently, employing a snowballing technique to expand the pool of selected papers, an additional 3 papers were incorporated, bringing the total number of selected papers to 29. Snowballing technique, also known as snowball sampling or iterative citation searching [19], is a technique commonly used in research to identify additional relevant studies beyond those initially retrieved. It involves reviewing the reference lists of selected papers to identify additional sources that

may not have been captured in the initial search by examining the references of retrieved papers and identifying relevant citations. Thus, the incorporation of additional papers through snowballing enriches the research process by capturing potentially overlooked or lesser-known studies that contribute to a more comprehensive understanding of the subject matter. Finally, the paper analysis is designed to collect all the information to address the research questions from the selected papers.

4. Results of Literature Analysis

The following data were systematically collected from the papers found (besides paper’s title, authors, year of publication, venue and URL/DOI):

- **Fairness Metrics (Causal or Not or Both):** Whether the paper discusses causal fairness metrics or statistical fairness metrics or both.
- **Analysis Type:** The type of analysis conducted in the paper: Classification, Evaluation, Definition, or Solution Proposal (a new tool, perspective or algorithm).
- **Content:** The paper’s main findings and contributions.
- **Context:** The specific domain or application area of the research.
- **Experimental Datasets:** The datasets used in the paper, if any.
- **Methods:** Mitigation techniques type, if pre-, in- or post-processing.

This comprehensive data collection approach ensures that each paper’s relevant information is captured accurately and thoroughly, facilitating further analysis and synthesis of findings in subsequent stages of the research process. A portion of our findings is encapsulated in Table 1.

4.1. Answering RQ1-RQ2

In this section, we are going to address the research questions RQ1 and RQ2 by analyzing the selected papers. The distribution of causal and non-causal papers among selected papers is: 70% focus on causal fairness metrics and 30% on statistical fairness metrics.

RQ1: What are the main concepts and differences between statistical-based and causal-based fairness?

Statistical-based fairness metrics are grounded in ensuring that the observed outcomes of a machine-learning model are distributed equitably. These metrics focus on the fairness of the model’s predictions without delving into the underlying causal mechanisms that generate the data. Causal-based fairness metrics, in contrast, emphasize the importance of understanding and addressing the causal relationships between variables. These metrics, including Counterfactual fairness and Interventional fairness, seek to identify and mitigate biases that arise from the causal influence of protected attributes on the outcomes. By examining the causal pathways, thus, causal fairness metrics aim to ensure that decisions are not only fair in an observational sense but also in a causal sense, addressing deeper, more systemic biases.

To summarize, the fundamental difference between statistical and causal fairness lies in their approach. While statistical fairness focuses on the distribution of outcomes, causal fairness delves into the root causes of biases, examining how and why these biases occur.

Q2: Which datasets are most commonly used in fairness research, and are there differences between those used in causal-based and statistical-based studies?

The visualization in Figure 2 provides the number of datasets used across papers and shows a detailed comparison of the frequency of dataset usage in these studies. From the chart, it is evident that the Adult dataset [41] is predominantly used in both causal and statistical studies, highlighting its relevance and utility in fairness research, given its rich demographic features and applicability to numerous fairness metrics. The Compas [3] dataset is widely used in causal studies, which underscores the necessity of understanding fairness within specific domains, such as criminal justice, where biases can have significant societal implications. Similarly, the German Credit dataset [42] is commonly used in statistical studies to evaluate fairness in financial decision-making, such as credit scoring models,

REF.	FAIRNESS	ANALYSIS	DATE	DATASETS	METHODS
[14]	Causal	Perspective	2017	/	/
[15]	Causal	Definition	2017	Law School	/
[20]	Both	Tool	2017	Adult, German	Pre, Post
[21]	Causal	Perspective	2018	Berkeley admissions	/
[22]	Causal	Definition	2018	Adult, Compas	Pre
[23]	Not Causal	Tool	2018	Criminal Justice, Public Health, Public Safety and Policing	In
[7]	Both	Classification	2018	German	/
[10]	Causal	Definition	2019	Synthetic, Adult, NYC Stop and Frisk	Pre
[24]	Causal	Algorithm	2019	Adult, Compas	Pre, Post
[12]	Causal	Algorithm	2019	Adult, Dutch	Pre
[25]	Causal	Tool	2019	Law school, NHS	Pre, Post
[11]	Causal	Algorithm	2019	Synthetic, Adult	Post
[26]	Causal	Tool	2020	Adult, Berkeley	Pre, Post
[27]	Causal	Tool	2020	Dunhumby	Pre
[18]	Causal	Definition	2020	Adult	/
[28]	Causal	Perspective	2020	BlogCatalog, Flickr	Pre, Post
[29]	Not Causal	Classification	2020	German	/
[30]	Not Causal	Algorithm	2020	ML-20M	In
[31]	Both	Classification	2020	Adult, Compas	Pre, Post
[32]	Both	Tool	2020	/	Pre, In, Post
[33]	Causal	Algorithm	2021	Adult, Credit Approval	Pre
[34]	Causal	Algorithm	2021	Synthetic, Adult, COMPAS, Nutrition	Pre
[35]	Causal	Evaluation	2021	/	/
[36]	Causal	Perspective	2021	MovieLens, Insurance	Pre, Post
[37]	Not Causal	Classification	2021	Adult, Compas	Pre, In, Post
[38]	Causal	Definition	2022	Synthetic, Bail, Credit defaulter	Pre
[39]	Both	Classification	2022	/	/
[40]	Not Causal	Evaluation	2023	Adult, Folk, Credit, Heart	/
[8]	Not Causal	Classification	2023	German	/

Table 1

Summary Table with papers on statistical and causal fairness

where use reflects the importance of ensuring equitable treatment in financial services, an important area where biases can impact individuals' economic opportunities. The chart also highlights the use of synthetic datasets in causal studies, which are artificially created by researchers, and they are essential for testing causal metrics because they allow researchers to design experiments that can isolate and examine the effects of specific variables on fairness, providing insights that might not be possible with real-world data.

5. Fairness Decision Tree

In this section, we are going to present our new fairness decision tree (Figure 3), which aims to extend and update the original proposal by Baresi et al. [8]. Our goal is to assist people in selecting the most appropriate fairness definitions for their machine learning (ML) systems by combining both statistical and causal fairness metrics. The proposed decision tree is open and can further be extended if new needs and definitions arise.

First, we will provide a detailed summary of the original decision tree, covering nodes A to F and metrics 1 to 12. Following that, we will introduce the new causal part of the tree (nodes G, H, I, L, M, and metrics 13 to 19), which is the original contribution of this work.

The tree begins with the question, "Is past knowledge relevant?" (A). If the answer is yes, the tree

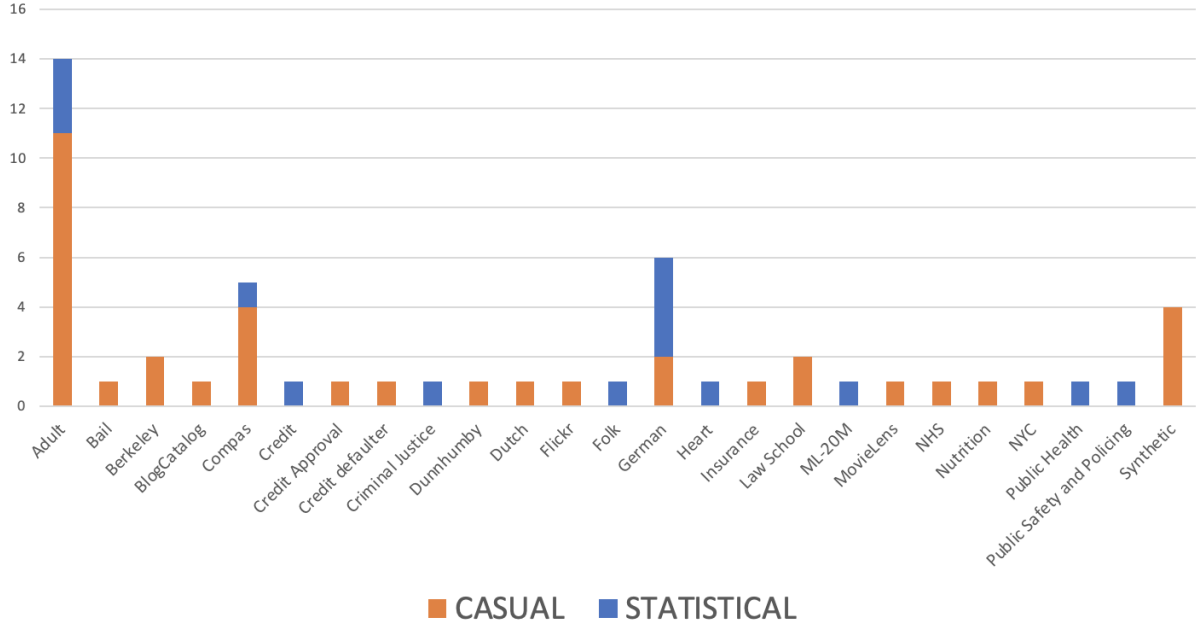


Figure 2: Number of utilized datasets for experiment across causal and statistical papers

helps experts decide about the importance of past decisions compared to new predictions. If the answer is no, the focus shifts to predictions and legitimate attributes (B). Based on the responses, the tree suggests different statistical fairness definitions. When past knowledge is relevant, the next question (C) asks which type of predictions the expert is interested in: wrong, correct, or both. If the focus is on wrong predictions, the tree further asks whether the interest is in negative or positive predictions (D) and how conservative the decisions should be (E). Based on the responses, the tree suggests specific statistical fairness definitions. If the expert is interested in correct predictions, the next question (F) asks how to balance predictions and past decisions, leading to other three possible statistical definitions.

The introduction of causal fairness metrics expands the tree with new questions (G, H, I, L, M). These additions address the limitations of purely statistical approaches by also considering the causal relationships. The new causal fairness metrics section starts with the question, "Are you interested in modifying the predictions?" (G). This question is important because modifying predictions involves engaging with causal relationships, it suggests an interest in not just observing what is but also in altering and understanding the causal relationships that lead to specific outcomes. If the answer is no, we move on (C), returning to the statistical part of the tree. If the answer is yes, the tree explores the possibility of adding or modifying the cause-effect relationships through the question "Do you want to add or modify cause-effect relations?" (H). Here, it is essential to distinguish between two paths that follow from the response to this question. This leads to two different nodes depending on whether the expert chooses to add (Counterfactual fairness metrics) or modify (Interventional fairness metrics) cause-effect relationships. In both cases, the decision tree differentiates between analyzing by group or by individual with the question "Do you want to analyze by group or by individual?" (I), which separates the path into group-level and individual-level analysis. This distinction is significant as it acknowledges that fairness can be assessed by considering the overall impact on groups, where the focus is on collective fairness, or by individual circumstances, which demands a more granular, personalized assessment of fairness.

In the counterfactual case (ADD), if the interest is in the group analysis, we suggest using the Unresolved Discrimination [7] (13). This metric captures any discrimination that remains after accounting for all known causal pathways. For those interested in individual analysis, the Counterfactual fairness [15] (14) is proposed. This metric assesses fairness by comparing the actual outcome with the outcome that would have occurred in a counterfactual world where the protected attribute is different.

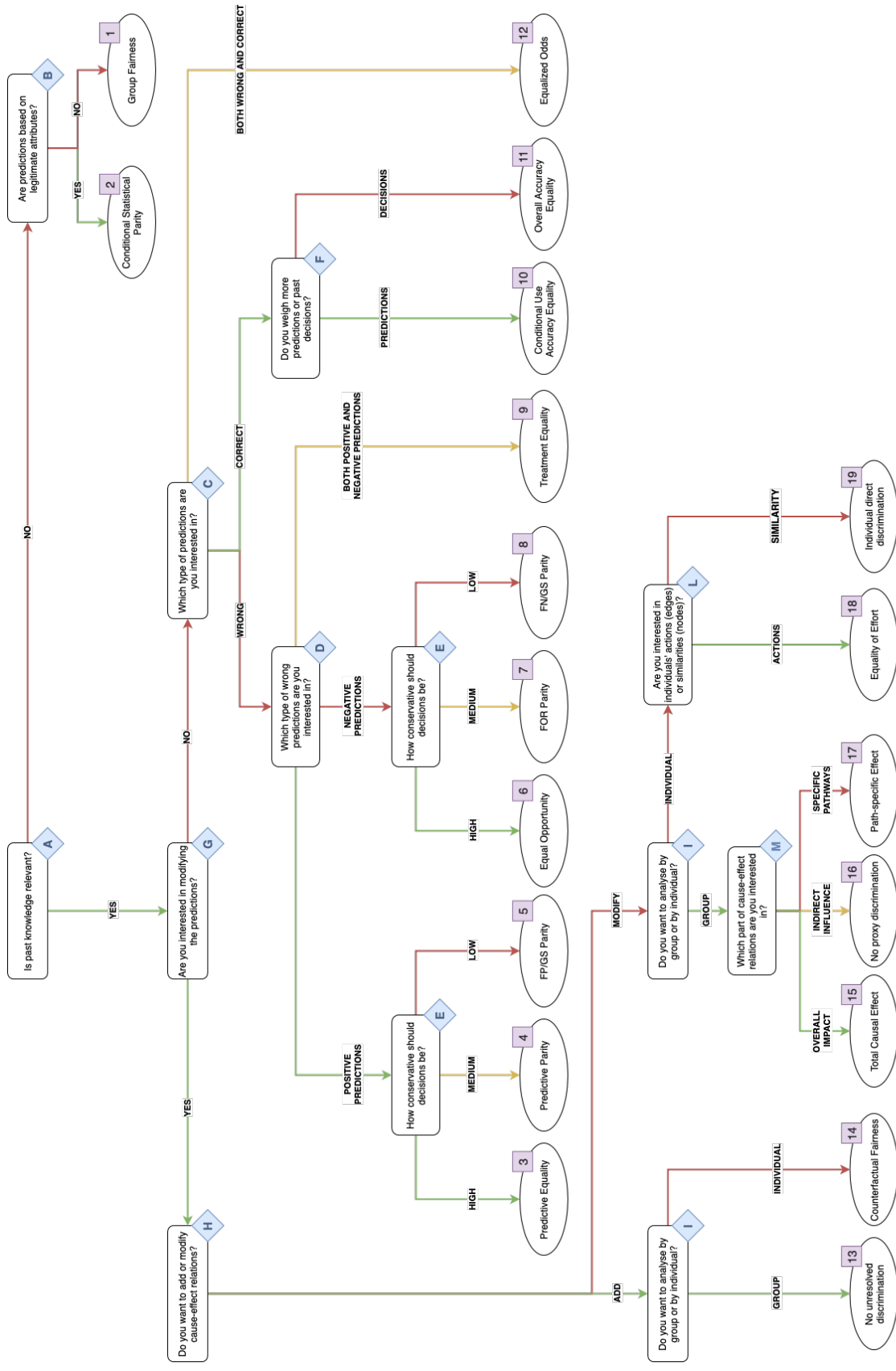


Figure 3: The new Fairness Decision Tree

In the Interventional case (MODIFY), for group analysis, if the interest is on modifying the overall impact of the causal relationships, the decision tree guides the expert to consider the Total Causal Effect [11] (15). This metric quantifies the total influence of a protected attribute on the outcome, considering all possible pathways. If the interest is in indirect influences, the No Proxy Discrimination [14] (16) metric is recommended. This metric ensures that the protected attribute does not influence the outcome indirectly through proxy variables. For those interested in specific causal pathways, the Path-Specific Effect [11] (17) is highlighted. This metric allows experts to dissect the causal graph and analyze the effect of the protected attribute through specific pathways. And when the analysis is at the individual level, the tree distinguishes between actions and similarities (L). This bifurcation addresses two different dimensions of individual-level fairness. Actions refer to the behaviors or efforts that individuals must undertake to achieve certain outcomes. For instance, the Equality of Effort [18] (18) metric is recommended. This metric assesses fairness by evaluating the level of effort required by different individuals to attain the same result, focusing on the processes or actions rather than just the outcomes. On the other hand, similarities refer to the comparison between individuals who have similar attributes. Here, the Individual Direct Discrimination [16] (19) metric is presented. This metric compares individuals with similar attributes (nodes) to ensure that decisions are not biased against similar individuals.

5.1. Answering RQ3-RQ4

This integration of causal and statistical fairness metrics into a unique fairness decision tree answers the following research questions.

RQ3: Is it possible to have a common vision between causal and statistical fairness?

While statistical fairness and causal fairness each take a different approach, they share a common vision of fairness in ML models. Both approaches recognize biases and inequities that may exist in the models that process data, but they address these issues from different angles: statistical metrics assess the fairness of a model by comparing observed data using conditional probabilities and conditional expectations, focusing on the distribution of model outputs and ensuring that the outcomes are equitable. Causal ones, instead, delve into the underlying causal relationships that drive model decisions. This approach is more concerned with understanding and eliminating the effects of sensitive attributes on model outputs through causal pathways. Thus, it provides a deeper analysis by examining how changes in a protected attribute causally influence the outcome, considering both direct and indirect effects. Despite their different methodologies, both statistical and causal fairness strive to achieve the same goal: reducing unfairness in ML models. Therefore, it is possible and also beneficial to have a common vision between causal and statistical fairness, as they complement each other in addressing different facets of bias and discrimination in ML.

RQ4: How to choose the most suitable metric considering both perspectives?

Choosing the most suitable fairness metric requires an approach that balances insights from both statistical and causal perspectives. This decision is complex and context-dependent, necessitating a thorough understanding of the dataset, the model, and the specific fairness goals of the application. The fairness decision tree framework presented in this chapter provides a method for navigating these choices. Here is a detailed approach to selecting the most appropriate metric.

- **Deeper Understanding of the Current Dataset and Model:** The first step is to analyze the features and attributes contained in the dataset. It is crucial to focus on protected attributes that might trigger inequality, such as race, gender, and age. Understanding the distribution of these attributes is essential for selecting appropriate metrics, as it helps identify potential sources of bias.
- **Choose the Appropriate Fairness Metrics:** Use the fairness decision tree as a guide to navigate through various fairness metrics. The definition of fairness can vary depending on the context. Thus, it is also crucial to clarify the specific fairness objectives to align the choice of metrics with the desired outcomes.

- **Comprehensive Assessment and Comparison:** Use a combination of multiple metrics to perform a comprehensive assessment. Comparing the results of different metrics can provide a better understanding of fairness. Additionally, it is beneficial to compare the chosen metrics with those used in other studies to ensure consistency and validity.

This approach ensures a deeper understanding and mitigation of underlying biases, considering both perspectives.

6. Conclusions

We investigate the topic of fairness in machine learning (ML), with a focus on the difference between statistical and causal fairness metrics, in particular, the possibility to incorporate them into a common vision. Our approach was motivated by the need to investigate, understand, and unify existing fairness metrics, which usually concentrate solely on statistical outcomes or delve into causal mechanisms without considering both perspectives. We introduced the fairness decision tree, a novel solution that integrates the causal fairness metrics into the original tree proposed by Baresi et al. [8], offering a comprehensive view for evaluating fairness in ML models. This tree guides users through a structured process to select the most suitable fairness metrics based on the specific context of their application. This solution also shows that the statistical-based and causal-based fairness metrics could have a common vision since both perspectives have the same goal: reducing unfairness in ML models, although they address different facets of bias.

6.1. Limitations and Future Work

One limitation of this work is that the fairness decision tree framework has yet to be validated and thoroughly tested in real-world scenarios. This validation could be achieved through user testing, surveys, or field experiments to determine its practicality and effectiveness in a variety of applications. Another limitation is that the reviewed literature covers only the last seven years, potentially excluding older but still relevant studies. Thus, expanding the time frame to include earlier work could provide a more comprehensive understanding of how fairness metrics have evolved over time. Furthermore, the fairness decision tree may need refinement when applied to various contexts, as real-world applications could present challenges not fully anticipated by the current framework.

Future research may improve the findings of this work in a variety of ways. One direction could be the application of the tree to various datasets, potentially through experiments involving real-world scenarios or user interactions to validate its utility and robustness across different domains. Furthermore, as fairness in ML evolves, new metrics and perspectives are likely to emerge. Thus, future research should be adaptive, incorporating these advancements while continuing to investigate the relationships between various fairness perspectives.

Acknowledgments

We thank Professor L. Tanca and T. Dolci for their support. This work was supported in part by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU and from the Italian PRIN project 2022XERWK9 “S-PIC4CHU” – Semantics-based Provenance, Integrity, and Curation for Consistent, High-quality, and Unbiased data science.

Declaration on Generative AI

During the preparation of this work, GPT-4 was used in order to check grammar and spelling. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, *ACM computing surveys (CSUR)* 54 (2021) 1–35.
- [2] K. Makhoul, S. Zhio, C. Palamidessi, Survey on causal-based machine learning fairness notions, *arXiv preprint arXiv:2010.09553* (2020).
- [3] W. Dieterich, C. Mendoza, T. Brennan, Compas risk scales: Demonstrating accuracy equity and predictive parity, *Northpointe Inc* 7 (2016) 1–36.
- [4] A. Hannak, G. Soeller, D. Lazer, A. Mislove, C. Wilson, Measuring price discrimination and steering on e-commerce web sites, in: *Proceedings of the 2014 Conference on Internet Measurement Conference, IMC '14*, Association for Computing Machinery, New York, NY, USA, 2014, p. 305–318. URL: <https://doi.org/10.1145/2663716.2663744>. doi:10.1145/2663716.2663744.
- [5] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, Y. Liu, How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 99–106.
- [6] S. Caton, C. Haas, Fairness in machine learning: A survey, *ACM Computing Surveys* (2020).
- [7] S. Verma, J. Rubin, Fairness definitions explained, in: *Proceedings of the international workshop on software fairness*, 2018, pp. 1–7.
- [8] L. Baresi, C. Criscuolo, C. Ghezzi, Understanding fairness requirements for ml-based software, in: *2023 IEEE 31st International Requirements Engineering Conference (RE)*, IEEE, 2023, pp. 341–346.
- [9] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, USA, 2009.
- [10] A. Khademi, S. Lee, D. Foley, V. Honavar, Fairness in algorithmic decision making: An excursion through the lens of causality, in: *The World Wide Web Conference*, 2019, pp. 2907–2914.
- [11] Y. Wu, L. Zhang, X. Wu, H. Tong, Pc-fairness: A unified framework for measuring causality-based fairness, *Advances in neural information processing systems* 32 (2019).
- [12] L. Zhang, Y. Wu, X. Wu, Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms, *IEEE Transactions on Knowledge and Data Engineering* 31 (2018) 2035–2050.
- [13] J. Pearl, The seven tools of causal inference, with reflections on machine learning, *Communications of the ACM* 62 (2019) 54–60.
- [14] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, *Advances in neural information processing systems* 30 (2017).
- [15] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, *Advances in neural information processing systems* 30 (2017).
- [16] L. Zhang, Y. Wu, X. Wu, Situation testing-based discrimination discovery: A causal inference approach., in: *IJCAI*, volume 16, 2016, pp. 2718–2724.
- [17] B. T. Luong, S. Ruggieri, F. Turini, k-nn as an implementation of situation testing for discrimination discovery and prevention, in: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 502–510.
- [18] W. Huan, Y. Wu, L. Zhang, X. Wu, Fairness through equality of effort, in: *Companion Proceedings of the Web Conference 2020*, 2020, pp. 743–751.
- [19] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, EASE '14*, Association for Computing Machinery, New York, NY, USA, 2014. URL: <https://doi.org/10.1145/2601248.2601268>. doi:10.1145/2601248.2601268.
- [20] S. Galhotra, Y. Brun, A. Meliou, Fairness testing: testing software for discrimination, in: *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, 2017, pp. 498–510.
- [21] J. R. Loftus, C. Russell, M. J. Kusner, R. Silva, Causal reasoning for algorithmic fairness, *arXiv preprint arXiv:1805.05859* (2018).
- [22] R. Nabi, I. Shpitser, Fair inference on outcomes, in: *Proceedings of the AAAI Conference on*

Artificial Intelligence, 2018.

- [23] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, R. Ghani, Aequitas: A bias and fairness audit toolkit, arXiv preprint arXiv:1811.05577 (2018).
- [24] B. Salimi, L. Rodriguez, B. Howe, D. Suciu, Capuchin: Causal database repair for algorithmic fairness, arXiv preprint arXiv:1902.08283 (2019).
- [25] N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, R. Silva, The sensitivity of counterfactual fairness to unmeasured confounding, in: Uncertainty in artificial intelligence, PMLR, 2020, pp. 616–626.
- [26] J. N. Yan, Z. Gu, H. Lin, J. M. Rzeszotarski, Silva: Interactively assessing machine learning fairness using causality, in: Proceedings of the 2020 chi conference on human factors in computing systems, 2020, pp. 1–13.
- [27] M. Sato, S. Takemori, J. Singh, T. Ohkuma, Unbiased learning for the causal effect of recommendation, in: Proceedings of the 14th ACM conference on recommender systems, 2020, pp. 378–387.
- [28] R. Guo, J. Li, H. Liu, Learning individual causal effects from networked observational data, in: Proceedings of the 13th international conference on web search and data mining, 2020, pp. 232–240.
- [29] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 514–524.
- [30] M. Morik, A. Singh, J. Hong, T. Joachims, Controlling fairness and bias in dynamic learning-to-rank, in: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 429–438.
- [31] B. Salimi, B. Howe, D. Suciu, Database repair meets algorithmic fairness, ACM SIGMOD Record 49 (2020) 34–41.
- [32] A. Castelnovo, R. Crupi, G. Del Gamba, G. Greco, A. Naseer, D. Regoli, B. S. M. Gonzalez, Befair: Addressing fairness in the banking sector, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 3652–3661.
- [33] B. Van Breugel, T. Kyono, J. Berrevoets, M. Van der Schaar, Decaf: Generating fair synthetic data using causally-aware generative networks, Advances in Neural Information Processing Systems 34 (2021) 22221–22233.
- [34] W. Pan, S. Cui, J. Bian, C. Zhang, F. Wang, Explaining algorithmic fairness through fairness-aware causal path decomposition, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1287–1297.
- [35] A. Kasirzadeh, A. Smart, The use and misuse of counterfactuals in ethical machine learning, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 228–236.
- [36] Y. Li, H. Chen, S. Xu, Y. Ge, Y. Zhang, Towards personalized fairness based on causal notion, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1054–1063.
- [37] M. Du, F. Yang, N. Zou, X. Hu, Fairness in deep learning: A computational perspective, IEEE Intelligent Systems 36 (2020) 25–34.
- [38] J. Ma, R. Guo, M. Wan, L. Yang, A. Zhang, J. Li, Learning fair node representations with graph counterfactual fairness, in: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 695–703.
- [39] A. N. Carey, X. Wu, The causal fairness field guide: Perspectives from social and formal sciences, Frontiers in big Data 5 (2022) 892837.
- [40] S. Guha, F. A. Khan, J. Stoyanovich, S. Schelter, Automated data cleaning can hurt fairness in machine learning-based decision making, IEEE Transactions on Knowledge and Data Engineering (2024).
- [41] B. Becker, R. Kohavi, Adult, UCI Machine Learning Repository, 1996. DOI: <https://doi.org/10.24432/C5XW20>.
- [42] M. Lichman, et al., Uci machine learning repository, 2013, 2013.