

BIAS AND UNFAIRNESS IN MACHINE LEARNING MODELS: A SYSTEMATIC LITERATURE REVIEW

Tiago Palma Pagano¹, Rafael Bessa Loureiro¹, Fernanda Vitória Nascimento Lisboa¹,
Gustavo Oliveira Ramos Cruz¹, Rodrigo Matos Peixoto¹, Guilherme Aragão de Sousa Guimarães¹,
Ewerton Lopes Silva de Oliveira², Ingrid Winkler¹, Erick Giovani Sperandio Nascimento^{1,3*}

¹SENAI CIMATEC, Salvador - BA

²HP Inc., Porto Alegre - RS

³Surrey Institute for People-Centred AI, School of Computer Science and Electronic Engineering,
Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom

*Corresponding author: erick.sperandio@fieb.org.br, erick.sperandio@surrey.ac.uk

ABSTRACT

One of the difficulties of artificial intelligence is to ensure that model decisions are fair and free of bias. In research, datasets, metrics, techniques, and tools are applied to detect and mitigate algorithmic unfairness and bias. This study aims to examine the latest existing knowledge about bias and unfairness in machine learning (ML) models with the RSL methodology and a bibliometric analysis. A Systematic Literature Review found 45 eligible articles published between 2017 and 2022 in the Scopus, IEEE Xplore, Web of Science, and Google Scholar knowledge bases. The results show numerous bias and unfairness detection and mitigation approaches for ML technologies, with clearly defined metrics in the literature, and varied metrics can be highlighted, we also address bias mitigation techniques, supporting tools, as well as more common datasets involving bias and unfairness identification and mitigation work, with binary and multi-class targets.

Keywords Bias · Unfairness · Machine Learning · Artificial Intelligence

1 Introduction

In industry, prediction-based decision algorithms are widely utilized by governments and organizations that are rapidly embracing them [1]. These techniques are already commonly used in lending, contracting, and online advertising, as well as in criminal pre-trial proceedings, immigration detention, and public health, among other areas [2]. With the rise of these techniques, worry emerged about the biases embedded in the models and how fair they are, defining their performance for issues related to sensitive social aspects such as race, gender, class, etc [3].

Systems that have an influence on people's lives raise ethical concerns about making judgments in a fair and unbiased fashion. As a result, bias and unfairness challenges have been extensively investigated taking into account the limits imposed by corporate practices, regulations, social traditions, and ethical obligations. [4]. Recognizing and reducing bias and unfairness are difficult tasks, since unfairness is defined differently in different cultures. As a consequence, user experience, cultural, social, historical, political, legal, and ethical considerations all have an impact on the unfairness criterion [5].

A definition for justice was given by [6], according to the author injustice is defined as "systematic and unfair discrimination or prejudice of certain individuals or groups of individuals in favor of others." He further states that injustice is commonly caused by social or statistical biases, the former referring to the divergence between how the world should be and how it actually is, the latter is the discrepancy between how the world is and how it is encoded in the system.

In [7] he differentiated the concepts of justice and bias, pointing out that usually works in the area use the two terms interchangeably. He defined that justice is a social concept of value judgment, being subjective, varying across cultures and societies and in the context of organizations such as schools, hospitals, or corporations. Biases, on the other hand, are related to systematic errors that alter human behaviors or judgments about others because of their membership in a group determined by distinguishing characteristics such as gender or age.

As a result, new approaches from data science, artificial intelligence (AI), and machine learning (ML) are needed to take into account the aforementioned constraints on algorithms [8].

The challenge worsens if key technological applications do not yet have ML models associated with the explainability of the decisions made, or those can only be evaluated by the team that created them, which leaves researchers unable to obtain these explanations and conduct experiments [9]. Given the millions of parameters analyzed by the machine, obtaining a transparent algorithm is quite challenging. Another option is to interpret it without knowing each step taken by the algorithm [10].

Analyzing bias and unfairness collaborates with model explainability, so explainability is intrinsic to the study. According to [11] explainability involves (i) defining model explainability, (ii) formulating explainability tasks to understand model behavior and developing solutions to those tasks, and finally, (iii) designing measures to evaluate model performance. Note that analyzing bias and unfairness directly addresses these topics, just as explainability promotes transparency.

Some solutions, such as AIF360, [12], FairLearn [13], Tensorflow Responsible AI [5] [14] [15] and Aequitas [16] are specific tools for dealing with bias and injustice.

However, the development approach to identify and mitigate bias and unfairness in ML models is left entirely to the developer, who often does not have adequate knowledge about the problem and must also consider aspects of fairness as a key element for the quality of the final model, confirming the need for a methodology to help deal with the problem [9].

Another challenge is that most existing solutions for mitigating bias and unfairness are one-off applications for a specific problem or use case (UC). There are numerous approaches to identifying bias and unfairness, known as fairness metrics, and this wide range makes it difficult to select the right evaluation criteria for the issue you want to mitigate [17, 18].

Given the contextualization done so far, this study aims to examine the most recent existing knowledge on bias and unfairness in machine learning (ML) models with the RSL methodology and bibliometric analysis.

The present work consists of the study involving 45 papers, as noted in Section 2. There are other literature review papers, such as [19, 20, 21, 22, 23, 24, 25]. However, our search did not consider papers prior to 2017, aiming to present a more recent overview of bias and unfairness in machine learning models, since, as shown in Figure 3, there was a spike in publications in the year 2018 and previously there was a smaller number, so papers after this period should present more recent overviews consistent with the most effective and less embryonic solutions of the topic. Given the above, it is important to point out that [19, 22, 20, 21, 24] have conducted their research without specifying a recent period, allowing already outdated surveys.

The work of [19] deals exclusively with datasets for bias and unfairness studies, without delving into mitigation aspects, while in [23] there is more unpacking of mitigation aspects, but the focus is on data management, highlighting the unfairness topics for this area compared to the others presented. While in [20] the focus is mainly on classification problems, highlighting that other technics are opportunities that should emerge in the coming years, in [24] is reinforced that work on algorithmic justice concentrates on single classification tasks, this finding were also identified by our work, however still [24] does not do an in-depth on bias mitigation methods, as well as [25] does not cite recent papers, having only one paper from 2019 and all others prior to this period.

Whereas [22] performs a more simplified analysis of fairness metrics and mitigation techniques, without addressing issues related to reference datasets for studies in the area. In the work of [21] they demonstrate why fairness is an important issue with examples of the possible harms in the real world, and they also examined the definitions of fairness and bias already proposed by researchers in different fields, such as general machine learning, deep learning and natural language processing, however the methodology for selecting the papers was not specified.

In [18] also reviews the literature related to algorithmic bias and makes recommendations for future research, however it focuses only on theoretical aspects of justice. It asserts that the mechanisms by which technology-driven biases translate into decisions and behaviors have been largely ignored. He brings definitions on how a context classifies, which can be individual, task, technology, organizational, and environmental, and can influence the perceptual and behavioral manifestations of bias in the model. His work seeks a reflection on the behavior of people impacted by model decisions in order to use them as an element of influence in model decisions. In [6] emphasizes that addressing

fairness in recommender systems provide different approaches, and, many gaps considering different users, research, the most important including: gender, age, ethnicity or personality.

As a differential, our work promotes a careful search methodology for the selection of the papers contained in Table 1, in order to extract the concepts and techniques discussed in the period when the theme was most debated among the scientific community.

The paper primarily seeks, methods of bias and unfairness identification and mitigation for ML technologies, through fairness metrics, bias mitigation techniques, supporting tools, as well as more common datasets, involving work addressing bias and unfairness identification and mitigation, with binary and multiclass targets. The unfolding of each of these elements will be addressed in the following sections.

This paper is organized as follows: Section II describes the research method and the advantages of using RSL, Section III examines the results and addresses elements such as the Types of Bias, the Identified Datasets with the main problems for identifying and reducing bias and unfairness, the Justice Metrics for measuring the models bias and unfairness in different ways, and from the identification of bias it is possible to approach the Techniques and models for bias and unfairness Mitigation, either by manipulating the data (pre-processing), the model itself (in-processing) or the prediction (post-processing), some techniques mainly of in-processing promote the identification of the sensitive attribute, important to train models independent of them. Section IV presents our final considerations and suggestions for further research.

2 Method

A Systematic Literature Review (SLR) aims to consolidate research by bringing together elements for understanding it [4]. Literature reviews are a widely used methodology to gather existing findings into a research field [26].

This systematic review followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [27] and was conducted using the method described in [28] and in [29], which encompasses five steps: Planning, Scoping, Searching, Assessing, Synthesizing.

During the Planning step, the knowledge bases that will be explored are defined [29]. The search for document patents was undertaken in the following knowledge bases:

- IEEE Xplore (www.ieeeexplore.ieee.org/)
- Scopus (www.periodicos.capes.gov.br)
- Web Of Science (www.periodicos.capes.gov.br/)
- Google Scholar (www.scholar.google.com)

These bases were chosen because they are reliable and multi-disciplinary knowledge databases of international scope, with comprehensive coverage of citation indexing, allowing the best data from scientific publications.

The Scope Definition step ensures that questions relevant to the research are considered before the actual Literature Review is carried out [29]. A brainstorming session was held with an interdisciplinary group composed of eleven experts on machine learning models, which selected two pertinent research questions to this systematic review address, namely:

Q1: What is the state of the art on the identification and mitigation of bias and unfairness in ML models?

Q2: What are the challenges and opportunities for identifying and mitigating bias and unfairness in ML models?

The Literature Search involves exploring the databases specified in the Planning step in a way that aims to solve the questions defined in the scope [29].

Initially, the keywords were used to search the knowledge bases noted in Figure 1. In addition to studies on bias or sensitive attributes using fairness or mitigation strategies for machine learning, it should include studies using the "AIF360", "Aequitas" or "FairLearn" tools for ML. This inclusion in the initial search aims to relate tools for identifying and mitigating bias and unfairness to the optimized search criteria, including the most important tools in the literature. These criteria defined the 'initial search', with 99 publications selected. Only review papers, research papers and conferences were considered.

These papers were used to optimize the search criteria using the litsearchr [30] library, which assembles a word co-occurrence network to identify the most relevant words. The optimized search yielded 16 selected papers, which can also be seen in Figure 1.

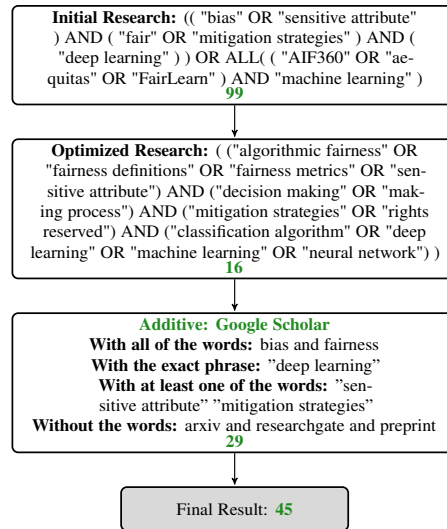


Figure 1: Papers Selection Process with the number of papers

In addition, a Google Scholar search was performed, and 29 publications were selected based on their Title and Abstract fields. The search was based on the string used in the databases, applying the same keywords in the advanced search criteria, as can be seen in Figure 1. The search in Google Scholar aims to select papers that might not have been indexed in the knowledge bases.

The Assessing the Evidence Base step selects the most relevant articles based on bibliometric analysis and reading the article abstracts.

Initially, searches in the four knowledge bases retrieved 99 articles, with the fields Title, Abstract, and Keywords serving as search criteria. We included only Review Articles, Research Articles, and Conference Proceedings published between 2017 and 2022, as shown by the bibliometric analysis in Figure 3. The red line represents the average difference in the number of articles over the previous five years, with a decrease in the final year due to the time span covered by the search.

A graph of the relationship between the keywords obtained in the search was generated using the biblioshiny tool [31] from the bibliometrix package [32] in the R language. Figure 2 illustrates some clusters that exemplify themes addressed in the RSL papers. The red cluster relates to "machine learning" and decision-making in models, the green cluster considers "fairness" and its economic and social impacts. It is also worth highlighting aspects related to transparency, interpretability, and the relationship of these keywords with the state of the art.

As a result of the final search, 45 articles were selected for discussion as shown in Figure 1.

The Synthesis and Analysis step consists in reading and evaluating the selected articles to identify patterns, differences, and gaps that might be studied further in future research on bias and unfairness in machine learning models.

3 Results

This section presents and analyses the 45 selected studies, which are included in Table 1, according to the research questions Q1 and Q2 set in the Scope Definition step. The results are organized in five sections: Types of Bias, Identified Datasets, Mitigation techniques and models, Technique for identification of the sensitive attribute and Fairness metrics. Those sections represent fundamental aspects of the discussion of bias and fairness.

The studies examined revealed issues that support the concern about bias and fairness in ML models. [9] addresses issues such as the lack of transparency of ML models, organizations such as Facebook and Telegram's lack of commitment to revealing the measures being taken in this effort, and even the constraints of resources, whether human or computer.

In the same manner, [6] highlights the importance of responsible AI, although there is still no clear and globally accepted definition of responsibility for AI systems. This should include fairness, security and privacy, explainability, safety and reproducibility. Specifically on fairness it highlights that the regulation should emphasize obligations to



Figure 2: Keyword co-occurrence network

"... minimize the risk of erroneous or biased decisions in critical areas such as education and training, employment, important services, law enforcement, and the judiciary."

Initially [9], criticizes the complexity of comprehending ML models, which can only be examined by the team that developed them, and which frequently does not understand all of the model's features or why it made certain judgments. Furthermore, the more complex the model, the more difficult it is to analyze its decision-making process.

The study [33] aims to provide an overview and a systemic view regarding recent criteria and processes in machine learning development, and to conduct empirical tests on the use of these for credit scores. The authors selected which fairness criteria best fit for these scores and cataloged state-of-the-art fairness processors, using them to identify when loan approval processes are met. Using seven datasets of credit scores, they performed empirical comparisons for different fairness processors.

The [34] study found security and transparency issues with automated decision systems (ADS), warning and urging data engineers to develop a more fair and inclusive procedure. For the authors, ADS must be accountable in the following areas: development, *design*, application, and usage, as well as rigorous regulation and monitoring, so that they do not perpetuate inequality.

The advantages and disadvantages of transparency in machine learning models, defining bias, fairness and arguing that a transparent algorithm is extremely difficult to obtain given the millions of parameters analyzed by a machine [10]. An alternative is a transparent output that can be analyzed and understood without having to understand every step made by the algorithm. To define transparency in [10] two categories have been defined: process transparency and result transparency.

The term "process transparency" refers to an understanding of the algorithm's underlying characteristics, such as the attributes it weighs in its decisions. The term "result transparency" refers to the capacity to understand decisions and patterns in classification process answers. In addition, the model must meet two requirements: global and local explanation. The Local explanation includes a detailed examination of which characteristics were most important in reaching a particular decision, whereas the Global explanation evaluates all decisions based on certain metrics. The author suggests a mental model of the main system for this evaluation, and if it can predict what the classification of the main model is, it is on the correct course to transparency. Finally, it is stated that a premise of *white-box* and *black-box* models might bring out implicit and explicit features of the models and facilitate auditors' job [10].

ML models, whether classification or regression, can be of type *White-box* or *Black-box*, depending on their availability and constraints:

- **White-box:** these are machine learning models that deliver easy to understand outcomes for application domain specialists. Typically, these models provide a good tradeoff between accuracy and explainability [35] and hence have less constraint and difficulties for structural adjustments. The structure and functioning of this model category are simple to grasp.

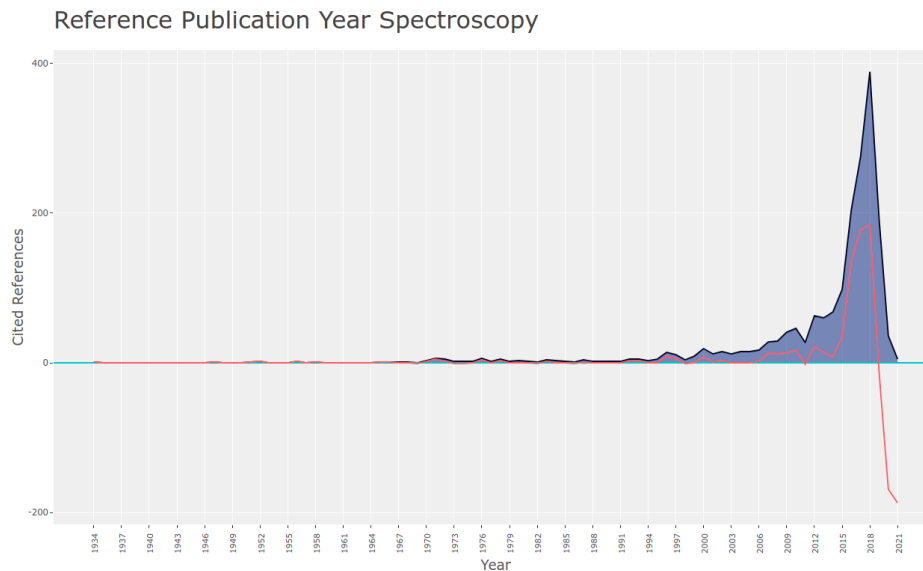


Figure 3: Year of the references cited in the papers

- **Black-box:** ML models that, from a mathematical perspective, are extremely difficult to explain and comprehend by specialists in practical areas [35]. Changes to the structure of models in this category are restricted, and it is difficult to grasp their structure and functioning.

The work [2] corroborates the argumentation of [9], emphasizing that when dealing with people, even the finest algorithm will be biased if sensitive attributes are not taken into consideration. One of the first issues raised is that the prejudice and justice literature is often confined to addressing the situation of a group or individual experiencing injustice in the present time. In this case, one must broaden the search and analyze how the individual's effect impacts his or her community and vice versa. Dataset and people behavior is fluid and can diverge dramatically over a few years, but the algorithm may retain a bias in its training and be unable to adapt to this shift. A group that is mistreated in the actual world would almost certainly be wronged by the algorithm, and that this type of bias just reflects reality rather than being a biased dataset.

In the work of [70] a problem is raised in the example of a model that evaluates the personality of patients using automated video interviews. If men have higher scores than women, this could be considered bias. On the other hand, if the annotations for agreement indicate higher scores for men than for women, the model reproduces this pattern and cannot be considered biased. There is an ambiguity in that the model would be fair because its measures reflect observable reality, while at the same time unfair because it gives unequal results to the group. This confusion occurs because of the lack of knowledge about the identification of the model's group bias, which uses right and wrong predictions criteria on the target provided by the dataset. Therefore, the model's right and wrong rates should be the same for different groups. There are numerous relationships that use these rates, as will be seen below.

With a similar opinion, [56] opposes the use of models for decision-making, defining the use of tools for risk assessment in models for pre-judgment as a justification. The authors argue that the implementation of these tools can introduce new uncertainties, disruptions, and risks into the judgment process. By conducting empirical experiments with unfair models, they conclude that the process of implementing these tools should be stopped.

Furthermore, [39] states that while there are various fair models for classification tasks, these are restricted to the present time, and because they embed the human bias, there is a propensity to repeat and escalate the segregation of particular groups through a vicious cycle. Whereas a classifier that gives a group a higher number of good ratings will give it an advantage in the future, and vice versa for negative ratings.

Meanwhile, [9] also claims that algorithms frequently disregard uncommon information, framing the act as censorship, such as Islamism and terrorist content. Because of this issue, decision-making algorithms tend to be biased toward more common occurrences in their case-specific databases.

Finally, [50] brings together the perspectives of various experts, emphasizing opportunities from the usage of AI, evaluating its impact, challenges, and the potential research agenda represented by AI's rapid growth in various fields of

Study
1 One-Network Adversarial Fairness [36]
2 Cyber gremlin: Social networking, machine learning and the global war on al-qaida-and is-inspired terrorism [9]
3 Fairness research on deep learning [37]
4 Mdfa: Multi-differential fairness auditor for black box classifiers [38]
5 Algorithmic fairness: Choices, assumptions, and definitions [2]
6 Dynamic fairness – breaking vicious cycles in automatic decision making [39]
7 Recycling privileged learning and distribution matching for fairness [40]
8 We need fairness and explainability in algorithmic hiring blue sky ideas track [8]
9 Detecting bias: Does an algorithm have to be transparent in order to be fair? [10]
10 Fairness-Aware Instrumentation of Preprocessing Pipelines for Machine Learning [41]
11 Improving machine learning fairness with sampling and adversarial learning [42]
12 Responsible data management [34]
13 Using Machine Learning in Admissions: Reducing Human and Algorithmic Bias in the Selection Process [43]
14 Analysis bias in sensitive personal information used to train financial models [44]
15 VITAL-ECG: A de-bias algorithm embedded in a gender-immune device [45]
16 Dataset bias: A case study for visual question answering [46]
17 A causal bayesian networks viewpoint on fairness [47]
18 Constraining deep representations with a noise module for fair classification [48]
19 Algorithm Bias Detection and Mitigation in Lenovo Face Recognition Engine [49]
20 Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy [50]
21 Cost-sensitive hierarchical classification via multi-scale information entropy for data with an imbalanced distribution [51]
22 Fair adversarial gradient tree boosting [52]
23 Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy [53]
24 Privacy and Ethical Challenges in Big Data [54]
25 Singular race models: Addressing bias and accuracy in predicting prisoner recidivism [55]
26 Fairness in Credit Scoring: Assessment, Implementation and Profit Implications [33]
27 When Politicization Stops Algorithms in Criminal Justice [56]
28 A survey on bias and fairness in machine learning, [57]
29 Evolution and impact of bias in human and machine learning algorithm interaction [58]
30 Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics [59]
31 Fairness for image generation with uncertain sensitive attributes [60]
32 Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer [61]
33 Constructing a Fair Classifier with the Generated Fair Data [62]
34 Enforcing fairness in logistic regression algorithm [63]
35 Fairness metrics and bias mitigation strategies for rating predictions [64]
36 Fairness via Representation Neutralization [65]
37 Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings [66]
38 Fairness in Deep Learning: A Computational Perspective [67]
39 Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle [68]
40 Fair Outlier Detection Based on Adversarial Representation Learning [69]
41 Recommender Systems under European AI Regulations [6]
42 Psychological Measurement in the Information Age: Machine-Learned Computational Models [70]
43 Monitoring Fairness in HOLDA [71]
44 Algorithmic bias: review, synthesis, and future research directions [18]
45 Integrating Psychometrics and Computing Perspectives on Bias and Fairness in Affective Computing: A case study of automated video interviews [7]

Table 1: Papers returned by the search string

industry and society in general. Tastes, anxieties, and cultural proximity seem to induce bias in consumer behavior, which will impact demand for AI goods and services, which is, according to the study, an issue that is yet under research.

Inferring patterns from large datasets in an unbiased environment and developing theories to explain those patterns can eliminate the need for hypothesis testing, eradicating the bias in the analysis data and, consequently, in the decisions [1].

In the [50] paper, general issues around ML are addressed, where governments are increasingly experimenting with them to increase efficiency in large-scale personalization of services based on citizen profiles, such as predicting viral outbreaks and crime hotspots, and AI systems used for food safety inspections. Bias in this context implicates governance issues, which pose dangers to society because algorithms can develop biases that reinforce historical discrimination, undesirable practices, or result in unexpected effects due to hidden complexities. Other related themes include ethics, transparency and audits, accountability and legal issues, fairness and equity, protection from misuse, and the digital divide and data deficit. These aspects are reinforced by [53] when it states that the discussion should expand to include technology diplomacy as a facilitator of global policy alignment and governance, for developing solutions to avian flu, for example. It also discusses the importance of implementing fundamental ethical concepts in AI, such as beneficence, non-maleficence, decision-making, fairness, explainability, reliable AI, suggested human oversight, alternative decision plans, privacy, traceability, non-discrimination, and accountability.

The work [54] addresses the issue of data privacy, as well as other ethical challenges related to Big Data research, such as transparency, interpretability, and fairness of algorithms based on this data. It is critical to explore methods to assess and quantify the bias of algorithms that learn from Big Data, particularly in terms of potential dangers of discrimination against population subgroups, and to suggest strategies to rectify unwarranted bias.

It also deals with the difference between individual justice and group justice, where the former states that individuals who are similar except for the sensitive attribute should be treated similarly and given similar decisions. This relates to the legal concept of unequal treatment when the decision-making process is based on sensitive attributes. However, individual justice is only relevant when the decision-making process causes discrimination, and cannot be used when the goal is to address biases in the data. Group justice, on the other hand, depends on the statistics of the outcomes of the subgroups indexed in the data and can be quantified in various ways, such as demographic parity and equalized odds, and thus can have bias addressed in the data [54]. In [7] states that group fairness, considers that groups contain useful information to adjust predictions, making them more accurate, highlighted that the metrics statistical parity, group fairness, and adverse impact are all concerned with equality of acceptance rates across groups.

3.1 Types of Bias

The types of biases are the pre-existing ones when they exist independently of an algorithm itself and have their origins in society. The technical bias, on the other hand, is occurring because of the systems developed, and it can be treated, measured, and its cause understood. He also defined the type of bias called emergent, which occurs when a system is designed for different users or when social concepts change [34].

For [57] bias can be classified into data bias, algorithm bias, and user interaction. The first considers that bias is present in the data, such as unbalanced data for example. The second one addresses the bias caused exclusively by the algorithm, caused by optimization functions, regularization, among other causes. The third type of bias is caused by the interaction with the user, since the interface allows the user to impose his/her behavior for a self-selected biased interaction.

The work [58] brings the concept of iterated algorithmic bias, features present in recommendation systems, with the types: filtering bias, the active learning bias, and the random-based bias. The first occurs when the goal is to provide relevant information or preferences. The second occurs when it aims to predict with the user's preferences. And the last one is based on an unbiased approach and used as a baseline for no user preference.

The study [34] defines three types of bias: pre-existing, technological, and emergent. The pre-existing bias category refers to data that reflects inequalities absorbed by the algorithm, hence spreading them. The technical category relates to bias worsening pre-existing prejudice caused by one of the algorithm's internal decision processes, and this may be addressed rather simply in comparison to the others. Finally, the emergent category refers to bias that occurs as a result of the usage of one or more users. For example, if a manager assigns higher performance to male employees, the algorithm is likely to begin favoring them and/or incorrectly rating women in the same division of the organization.

The paper [57] goes further in this concept, listing the 23 most common sources of bias, and these are divided into three categories organized in order to consider the feedback loop, they are: data, algorithm, and user interaction. So we have some examples of biases:

- Historical and social: coming from the data;
- Emerging and popularity: coming from the algorithm;

- Behavioral and presentation bias: caused by interaction with the user.

In the article [58] the authors proposed a framework to analyze bias and concluded that filtering bias, prominent in personalized user interfaces, can limit the discoverability of relevant information to be presented. In addition, they address the importance and damage caused by feedback loops and how algorithm performance and human behavior influence each other by denying certain information to a user, impacting long-term performance.

The work [68] proposed a methodology to identify the risks of potential unintended and harmful biases in ML. They therefore developed a practical risk assessment questionnaire to identify the sources of bias that cause unfairness and applied it to cases such as criminal risk prediction, health care provisions, and mortgage lending. The questionnaire was validated with industry professionals, and 86% agreed it was useful for proactively diagnosing unexpected issues that may arise in the ML model. Note that this work allows you to identify causes that may bias the models in a theoretical way.

3.2 Identified Datasets

A survey of the datasets used in the papers was conducted, listed in Table 2. These datasets mostly are known to include demographic annotations, allowing to assess unfairness and bias in their data, and can be used to test and validate techniques aimed at resolving these issues. Other datasets do not have demographic data, as it aims to identify bias and unfairness in image generation, reconstruction, enhancement, and super-resolution, not necessarily associated with demographic sensitive issues [60].

Metrics	References
COMPAS	[37, 41, 62, 36, 40, 48, 63, 38, 69]
Communities	[38]
FDOC	[55]
FDLE	[55]
Student	[69]
Bank	[37, 44, 48]
German	[37, 62, 48, 38]
Credit	[69]
Adult	[37, 41, 62, 44, 59, 36, 40, 42, 48, 63, 65, 69, 38]
Boston	[37]
MEPS	[37, 62, 65]
Heart	[37]
MIMIC II	[45]
Weight	[69]
Drug	[69]
AFHQ Cats and Dogs	[60]
LFW	[61]
CelebA	[61, 65]
MOPRH	[61]
MovieLens IM	[64]
CI-MNIST	[59]
VQA	[46]
VizWiz	[46]
CLEVR	[46]
Sintético	[38]

Table 2: Table with the datasets present in each paper.

Some datasets address crime-related issues such as Propublica Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), Communities and Crime (Communities), and Florida Department of Corrections (FDOC).

The COMPAS [72] dataset describes a binary classification task, which shows whether an inmate will reoffend within two years, has sensitive attributes such as race, age and gender. This is one of the most widely used datasets for bias and fairness experiments, with a controversial and relevant topic.

Similar in purpose to COMPAS, the Communities dataset [73] compares various socioeconomic situations of US citizens in the 1990s with the crime rate, identifying the per capita rate of violent crime in each community.

The FDOC [55] dataset, on the other hand, contains sentences with the types of charges, which can be violent charges (murder, manslaughter, sex crimes, and other violent crimes); robbery; burglary; other property charges (including

theft, fraud, and damage); drug-related charges; and other charges (including weapons and other public order offenses). The dataset uses Florida Department of Law Enforcement (FDLE) criminal history records for recidivism information within 3 years. They have the attributes such as the major crime category, the offender's age of admission and release, time served in prison, number of crimes committed prior to arrest, race, marital status, employment status, gender, education level, and if recidivist whether or not they were supervised after release.

Addressing issues concerning the selection process and approval of individuals, the Student [74] dataset has the data collected during 2005 and 2006 in two public schools in Portugal. The dataset was built from two sources: school reports, based on sheets of paper including some tributes with the three grades of the period and number of school absences; and questionnaires, used to complement the previous information. It also includes demographic data with mother's education, family income, social/emotional situation, alcohol consumption, variables that can affect student performance.

Another theme found in the selected datasets involves financial issues of bank credit such as Bank marketing (Bank), German credit (German) and Credit. Wage forecasting with the Adult dataset and product pricing with the Boston housing price (Boston) dataset.

The Bank dataset is related to the marketing campaigns of a Portuguese bank between the years 2008 to 2013. The goal of the classification is to predict whether or not a customer will make a deposit subscription [44].

With similar purpose, the German [73] dataset has 1000 items and 20 categorical attributes. Each entry in this dataset represents an individual who receives credit from a bank. According to the set of attributes, each individual is evaluated on his or her credit risk.

The Credit [75] dataset, on the other hand, contains payment data from a Taiwanese bank (a cash and credit card issuer) for the purpose of identifying the bank's credit card holders who would potentially receive a loan. It has demographic annotations such as education level, age, and gender.

One of the most widely used datasets, Adult [73] has 32,561 full cases representing adults from the 1994 US census. The task is to predict whether an adult's salary is above or below \$50,000 based on 14 characteristics. The sensitive attribute 'gender' is embedded in the samples.

For real estate pricing, the Boston dataset has data extracted from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970 and each of the 506 samples represents data obtained on 14 characteristics for households. The classification of this model aims to predict the property value of the region using attributes such as crime rate, proportion of residential land, average number of rooms per household, among others [37].

Another characteristic of the datasets found highlights applications in the health area, either to predict patients' financial expenses, as in the dataset Medical Expenditure Panel Survey (MEPS), or to identify possible health risks for patients as in the datasets: MEPS, Heart Disease (Heart), Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II), Weight, and Drugs.

The MEPS [76] dataset contains data on families and individuals in the United States, with their medical providers and employers, with information on the cost and use of health care or insurance.

To identify and prevent diseases the Heart [37] dataset contains 76 attributes, but all published experiments refer to the use of a subset of 14 of them. The target attribute refers to the presence of heart disease in the patient and can be 0 (no presence) to 4. Experiments aim to classify the presence or absence of heart disease.

In the same vein as Heart, the MIMIC II [45] dataset contains vital signs captured from patient monitors and clinical data from tens of thousands of Intensive Care Unit (ICU) patients. It has demographic data such as patient gender and age, hospital admissions and discharge dates, room tracking, dates of death (in or out of hospital), ICD-9 codes, unique code for healthcare professional and patient type, as well as medications, lab tests, fluid administration, notes and reports.

The Weight [77] dataset contains data for estimating obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. It has 17 attributes and 2,111 samples, labeled with the level of obesity which can be Low Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. The sensitive attributes are gender, age, weight, height, smoking, among others.

To predict narcotic use, the Drug [78] dataset was collected from online survey including personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information. The *dataset* contains information on the use of 18 central nervous system psychoactive drugs such as amphetamines, cannabis, cocaine, ecstasy, legal drugs, LSD, and magic mushrooms, among others. It has demographic attributes such as gender, education level, and age group.

In the area of image enhancement and face recognition, bias may not be associated with demographic features, the datasets that have demographic information were identified, among them: Labeled Faces in the Wild (LFW), Large-scale CelebFaces Attributes (CelebA), MORPH Longitudinal Database (MORPH), MovieLens 1M and Visual Question Answering (VQA). The dataset Animal FacesHQ (AFHQ) deals with the identification of animals, and the bias is associated with implicit features of the images, as well as the dataset Correlated and Imbalanced MNIST (CI-MNIST). Synthetic datasets were also found as an alternative.

The LFW [61] dataset contains 13,233 images of faces of 5749 distinct people and 1680 individuals are in two or more images. LFW is applied to face recognition problems and the images were annotated for demographic information such as gender, ethnicity, skin color, age group, hair color, eyeglass wearing, among other sensitive attributes.

The CelebA [79] dataset contains 202,599 face images with 10,177 individuals and 40 annotated attributes per image such as gender, Asian features, skin color, age group, head color and eye color, among other sensitive attributes, just as LFW is also used for face recognition problems.

The MORPH dataset contains over 400,000 images of almost 70,000 individuals. The images are 8-bit color and sizes can vary. MORPH has annotations for age, sex, race, height, weight, and eye coordinates.

The MovieLens 1M [64] dataset contains a set of movie ratings from the MovieLens website, a movie recommendation service of 1 million reviews from 6,000 users for 4,000 movies, with demographics such as gender, age, occupation, and zip code, plus data from the movies and the ratings.

The VQA [46] dataset contains natural language questions about images. It has 250,000 images, 760,000 questions and about 10 million answers. The questions have a sensitive criterion from the point of view of the questioner, and can be a simple question or a very difficult one, creating a bias. The images can also be very complex, making it difficult to identify the question element. The VizWiz dataset has the same proposal as the VQA for object recognition and assistive technologies, collected from users with visual impairment. CLEVR has a similar proposal to VQA and VizWiz, but was generated automatically by algorithms containing images with three basic shapes (spheres, cubes and cylinders) in two different sizes (small and large) and eight different colors, and includes questions and answers with the elements contained in the images. The combination of VQA, VizWiz and CLEVR gave origin to another dataset of questions and answers, annotated with the sensitive attribute of the visual conditions of the user who asked the question, which could be normal vision, visually impaired or robot.

The AFHQ [60] dataset is a dataset of animal faces consisting of 15,000 high-quality images at 512×512 resolution. It includes three domains of cat, dog and wildlife, each providing 5,000 images, it also contains three domains and several images of various breeds (larger than eight) for each domain. All images are aligned vertically and horizontally to have the eyes in the center. Low quality images were discarded. The work by [60] used only images of cats and dogs.

The Correlated and Imbalanced MNIST (CI-MNIST) [59] dataset is a variant of the MNIST dataset with additional artificial attributes for eligibility analysis. For an image, the label indicates eligibility or ineligibility, respectively, given that it is even or odd. The dataset varies the background colors as a protected or sensitive attribute, where blue denotes the non-privileged group and red denotes the privileged group. The dataset is designed to evaluate bias mitigation approaches in challenging situations and address different situations. The dataset has 50,000 images for the training set, 10,000 images for validation and testing with the eligible images representing 50 percent of each of these. Various background colors, colored boxes added at some top of the image of varying sizes were used to allow the impact of the colors, positions and sizes of the elements contained in the image to be analyzed.

Another alternative for the dataset is to create it synthetically [38], in which case it follows a normal distribution for the data. In it was created a binary sensitive attribute with Bernoulli distribution for its occurrence.

The use of the datasets presented can be seen in Section 3.5 associated with mitigation techniques.

3.3 Fairness metrics

In the research [39], claims that machine learning models increasingly provide approaches to quantify bias and inequality in classification operations as a methodology for measuring bias and fairness. While many metrics have been developed, when it comes to long-term decisions, the models and scientific community have produced poor outcomes. Some existing metrics for measuring model bias are insufficient, either because they only evaluate the individual or the group, or because they are unable to predict a model's behavior over time. The authors offer the metric Demographic Parity as a solution, which when applied to a model ensures that the average classification of individuals in each group converges to the same point, achieving a balance between accuracy, bias, and fairness for the groups classified by the model.

In [40], Demographic Parity, which assures that decisions are unconnected to sensitive attributes, was one of the metrics used to evaluate the model. Equalized Odds, to guarantee parity between positive and negative evaluations, and Equality

of Opportunity, to ensure that individuals meet the same criteria and are treated equally. Each of these metrics assures that groups are treated fairly and that the model's quality does not deteriorate or become biased over time, as addressed in [39].

The metrics for assessing fairness should apply the same treatment to multiple groups, however if one of the metrics identifies bias, other metrics can charge that the model is fair.

Five metrics for assessing fairness were established from the review of the papers: Equalized Odds, Equality of Opportunity, Demographic Parity, Individual Differential Fairness, and MDFA.

As a basis for the fairness metrics, it is important to define true positive (TP), false positive (FP), true negative (TN), and false negative (FN). These values are obtained from the rights and wrongs of the model's prediction relative to the target or ground truth provided by the dataset. Positive values are defined as the positive class that the model should predict, as opposed to negative values. For example, if the model should predict whether an individual will reoffend, the positive class will be 1, which indicates that the individual will reoffend, and the negative class will be 0. Therefore, if the positive classes are correct, they will be computed in TP, while the errors will be computed in FP. On the other hand, hits for negative classes will be computed in TN and errors in FN.

For a multiclass problem there is no positive and negative class, just consider the values for each individual class, observe Figure 4.

		True Class		
		Class 1	Class 2	Class 3
Predicted Class	Class 1	TP	FP	FP
	Class 2	FN	TN	TN
	Class 3	FN	TN	TN

Figure 4: Confusion Matrix Multiclass

In the example, the scenario for calculating the values of Class 1 is illustrated, TP is the value of the correct prediction, consistent with the target. The TN are the sum of the classes that do not involve class 1, neither in the prediction, nor in the target. The PF are the sum of the classes falsely predicted as class 1, while the FN are the sum of the classes predicted as other classes that should have been predicted as class 1.

This process should be performed for all classes and the overall TP, FP, TN and FN of the model should be averaged over the individual values.

To understand the justice metrics, which use the TP, FP, TN and FN, the statistical metrics must also be defined as per Table 4.

The objective of the metric Equalized Odds is to ensure that the probability that an individual in a positive class receives a good result and the probability that an individual in a negative class wrongly receives a positive result for the protected and unprotected groups are the same. That is, the TPR and FPR of the protected and unprotected groups must be the same. [21].

$$EO = \frac{1}{2} * \left(\left| \frac{FP_p}{FP_p + TN_p} - \frac{FP_u}{FP_u + TN_u} \right| + \left| \frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u} \right| \right) \quad (1)$$

In contrast, the metric "Equality of Opportunity" must satisfy equal opportunity in a binary classifier (Z). As a result, the probability of an individual in a positive class receiving a good outcome must be the same for both protected and unprotected groups. That is, the TPR for both the protected and unprotected groups must be the same.[21].

Statistical Metrics	References	Equation
Positive Predictive Value (PPV)	[3]	$PPV = TP / (TP + FP)$
False Discovery Rate (FDR)	[3]	$FDR = FP / (TP + FP)$
False Omission Rate (FOR)	[3]	$FOR = FN / (TN + FN)$
Negative Predictive Value (NPV)	[3]	$NPV = TN / (TN + FN)$
True Positive Rate (TPR)	[3, 40, 61]	$TPR = TP / (TP + FN)$
False Positive Rate (FPR)	[3, 40, 55]	$FPR = FP / (FP + TN)$
False Negative Rate (FNR)	[3, 55]	$FNR = FN / (TP + FN)$
True Negative Rate (TNR)	[3]	$TNR = TN / (FP + TN)$

Table 3: Table of statistical metrics

$$EOO = \frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u} \quad (2)$$

According to the justice metric Demographic Parity (DP), also known as Statistical Parity, the probability of an outcome being positive [21]. For this, the formula below should be applied.

$$DP = \frac{TP + FP}{N} \quad (3)$$

The justice metric Disparate Impact (DI) compares the proportion of individuals who receive a favorable outcome for two groups, a protected group and an unprotected group. This measure must equal to 1 to be fair.

$$DI = \frac{\frac{TP_p + FP_p}{N_p}}{\frac{TP_u + FP_u}{N_u}} \quad (4)$$

The K-Nearest Neighbors Consistency (KNNC) justice metric, on the other hand, is the only individual justice metric used by [62], it measures the similarity of sensitive attribute labels for similar instances [44].

$$KNNC = 1 - \frac{1}{n} \sum_{i=1}^n \left| \hat{y}_i - \frac{1}{k} \sum_{j \in \mathcal{N}_k(x_i)} \hat{y}_j \right| \quad (5)$$

Difference metrics were used as fairness metrics by [62], they are: Absolute Balanced Accuracy Difference (ABAD), Absolute Average Odds Difference (AAOD), Absolute Equal Opportunity Rate Difference (AEORD) and Statistical Parity Difference (SPD). The Differences metrics are calculated from the difference of the 'Disparity' metrics between two classes.

The ABAD is the difference in balanced accuracy in protected and unprotected groups, defined by Equation 6.

$$ABAD = \left| \frac{1}{2} [TPR_p + TNR_p] - [TPR_u + TNR_u] \right| \quad (6)$$

The AAOD is the absolute difference in TPR and FPR between different protected groups, defined by Equation 7.

$$AAOD = \left| \frac{(FPR_u + FNR_p) - (TPR_u + TPR_p)}{2} \right| \quad (7)$$

AEORD is the difference in recall scores (TPR) between the protected and unprotected groups. A value of 0 indicates equality of opportunity, defined by Equation 8.

$$AEORD = |TPR_p - TPR_u| \quad (8)$$

Finally, SPD is the difference in SD between a protected and an unprotected group, defined by Equation 9.

$$SPD = \frac{TP_p + FP_p}{N_p} - \frac{TP_u + FP_u}{N_u} \quad (9)$$

In addition to fairness metrics, some works use classification metrics such as accuracy, precision, recall and F1-score [46] as criteria for identifying bias. In addition to fairness metrics, some works use classification metrics such as accuracy, precision, recall and F1-score [46] as criteria for identifying bias. Measures of bias linked to the accuracy of model predictions are designed to check for unexpected differences in accuracy between groups. A less accurate prediction for one group compared to another contains systematic error, which disproportionately affects one group over the other [7]

Accuracy is the ratio between the number of true negatives and true positives to the total number of observations. Precision is the proportion of correct positive identifications. Recall is the proportion of true positives correctly identified. The F1-score is the weighted average of Precision and Recall. The formulas for each can be seen in Equations 10, 12, 13, 13

$$accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (10)$$

$$precision = \frac{TP}{TP + FP} \quad (11)$$

$$recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1\text{-score} = \frac{2 * (recall * precision)}{(recall * precision)} \quad (13)$$

Other cases used the number of positive (NIP) and negative (NIN) instances as the criteria for fairness metrics, as well as the base rate (BR) also known as prior probabilities are the unconditional probabilities, it is a probability with respect to all samples (N) [44]. The formulas for each can be seen in Equations 14, 14, 16

$$NIP = TP + FP \quad (14)$$

$$NIN = TN + FN \quad (15)$$

$$BR = NIP/N \quad (16)$$

All reported fairness metrics can be seen in Table 4.

Metric Fairness	References
EO	[21, 61, 59, 63, 65]
EOO	[21, 59, 42, 63, 69]
DP	[21, 60, 59, 42, 65, 69]
DI	[44]
KNNC	[44]
ABAD	[62]
AAOD	[62]
AEORD	[62]
SPD	[62]
accuracy	[46]
precision	[46]
recall	[46]
F1-score	[46]
NIP	[44]
NIN	[44]
BR	[44]

Table 4: Metrics used as fairness criteria

3.4 Techniques for Bias Analysis

For bias analysis and identification, a methodology was proposed by [71] for models trained with federated learning with the HOLDA architecture, checking the influence of biased individuals on unbiased individuals. Whenever a user updates its internal state by replacing the previous best model, when that model has a better generalization performance on the local validation data, the system evaluates the fairness of that new model. They performed an experiment training an ANN with 200 neurons in the hidden layer. The sensitive attribute used was "Gender". They concluded that local models trained with unbiased customers have little influence on the model, while biased customers impact the model unfairness. In this way the biased customers end up influencing the unbiased ones, but local models trained only with an unbiased customer tend to be slightly unfair. The dataset used was Adult. The fairness metrics used were DP, EO and EOO.

Also aiming to identify unfairness and promote explainability of model decisions, the technique suggested by [10] includes a model that combines white-box and black-box features for local and global explanations, respectively. Local explanation involves determining which features contributed the most to the classification of a given data sample, which can be achieved with a visualization tool or algorithm that can simulate and explain the decisions of the original model. In terms of overall explanation, the model decisions perform comparisons with the classifications obtained by each group, using decile risk scores to demonstrate whether there is bias in the model. The experiments were performed with the COMPAS dataset.

3.5 Mitigation techniques and models

As noted earlier, bias and unfairness mitigation techniques can be of the types: pre-processing, in-processing, and post-processing. According to [6] pre-processing mitigation techniques focus on rebalancing the data. In-processing mitigation, on the other hand, focuses on the model and its regularization with a bias correction term in the loss function or implicit in the model as with adversarial networks, where the model predicts the sensitive attribute.

The preprocessing mitigation technique aims to alter the dataset in a way that positively impacts the fairness metrics, as can be seen in [41]'s study, where FairDAGs library is proposed as an acyclic graph generator that describes the data flow during preprocessing. Its purpose is to identify and mitigate bias in the distribution and distortions that may arise with protected groups, while allowing direct observation of changes in the dataset. The four types of treatment are: bias by filtering the data, standardizing missing values, changes in the proportion of the dataset after replacement of NaN values, and, for natural language processing (NLP) systems, filtering out extraneous names or words that the computer may not recognize. The results showed that DAG was able to identify and represent differences in the data that occurred during preprocessing, as well as correct imbalances in the datasets examined.

In the work of [44] the preprocessing has a different purpose, as it aims to remove sensitive data from the model for a banking system, ensuring the removal of customer data after the output without affecting the ML model. The goal is the generation of synthetic data from the representation of the original data in order to preserve privacy while maintaining the usefulness of that original data. The synthetic data is generated by the Trusted Model Executor (TME) which is an AIF360 tool. At the end, the bias in the synthetic dataset was evaluated by comparing it with the original datasets in order to validate the TME.

Also using AIF360 to perform preprocessing operations, the study by [45] addresses that smartwatches distinguish between men and women in the identification of cardiovascular problems, evaluating more characteristics of the former group than the latter. In view of the above there should be a correction to fit the needs of both genders the removal of sensitive data, with the rebalancing of the dataset distribution and processing operations. It also adjusts non-representative data for accurate assessment of user health. The mitigation technique in preprocessing used was Reweighting. At the end the Vital-ECG was developed, a watch-like device that detects heart rate, blood pressure, skin temperature and other body variables without distinction of gender and with superior predictions.

Still in the area of data generation, the work [62] generates a new dataset that has no disparity of distribution, quality or noise, ensuring that all classes are treated equally. To do this it used the VAE-GAN architecture which, although it showed great improvements in model impartiality, the use of synthetic data during training limited its ability to generalize real data, reducing accuracy and precision. To minimize the trade-off, the model trained with artificial data used transfer learning techniques to perform an adjustment of the weights with real data.

In [61] work, in the area of computer vision, highlights that face recognition and analysis models generally exhibit demographic biases, even in models where accuracy is high. The reason is usually due to datasets with under-represented categories, whether for identifying identity, gender, or expressions of the human face. Biases can be in relation to age, gender, and skin tone. Therefore, a bias mitigation technique was proposed with a dataset of facial images, where to increase demographic diversity, a style transfer approach using Generative Adversarial Networks (GANs) was used to

create additional images by transferring multiple demographic attributes to each image in a biased set. The literature review study [37] highlights pre-processing techniques to mitigate data bias, such as Synthetic Minority Over-sampling Technique (SMOTE) and uses Data Augmentation, as well as [61]. In the end it defines open questions on the topic such as the fact that metrics can be conflicting, indicating a model that is fair in one metric and unfair in another.

Also in the area of computer vision, [60] presents several intuitive notions of group fairness, applied to image enhancement problems. Due to the uncertainty in defining the clusters, since, for the author, there are no ground truth identities in the clusters, and the sensitive attributes are not well defined.

Some conclusions about the impacts of fairness metrics on the pre-processing mitigation process could be obtained in [60]. It states that the metric *demographic parity* is strongly dependent on clusters, which is problematic for generating images of people in the data augmentation process, because the classes of the sensitive attribute 'race' are ill-defined. In CPR, implemented using *Langevin dynamics*, this phenomenon does not occur, and it can be seen in the results obtained that, for any choice of protected clusters, the expected properties are displayed.

The fairness metrics identified in the papers that addressed preprocessing are in Table 5, as are the datasets in Table 6.

Fairness Metrics	References
FPR	[41]
FNR	[41]
Demographic Parity	[60]
TPR	[61]
Equalized Odds	[61]
Absolute Balanced accuracy difference	[62]
Absolute average odds difference	[62]
Absolute equal opportunity rate difference	[62]
Statistical Parity Difference	[44]
Disparate Impact	[44]
K-Nearest Neighbors Consistency	[44]
Number of Positive Instances	[44]
Number of Negative Instances	[44]
Base Rate	[44]

Table 5: Fairness metrics used in preprocessing techniques

Datasets	References
Heart	[37]
Adult	[37, 41, 62, 44]
Bank marketing dataset	[37, 44]
Boston house price dataset	[37]
COMPAS dataset	[37, 41, 62]
German credit dataset	[37, 62]
Medical Expenditure Panel Survey (MEPS)	[37, 62]
FlickrFaces	[60]
AFHQ Cats and Dogs	[60]
LFW	[61]
CelebA	[61]
MOPRH	[61]
MIMIC II	[45]

Table 6: Datasets used in preprocessing techniques

The in-processing mitigation technique was identified in a larger amount of papers, and can be observed in [43, 63, 55, 59, 64, 36, 42, 40, 48]

An in-processing solution was proposed by [43] in the holistic and often subjective methods that may contain biases in the student selection process in schools. From this perspective, learning algorithms capable of admitting a diverse

student population were developed, even for groups with historical disadvantages. The study examined the impact of characteristics such as income, color, and gender on student admission rates.

The work [63] also presents an in-process mitigation solution for group bias. It used the logistic regression technique to develop the model. The solution used was Pareto Optimal, which aims to ensure a better accuracy loss function while keeping the fairness metrics at the threshold set at 80%. The author states that the in-processing solution, where the algorithm is adjusted during learning, would be a natural solution, because the pre-processing algorithms would be altering the original data, hurting ethical norms, however it is possible to work with data balance, without altering the users' data.

Another em-processing mitigation model was proposed by [55], with a new classification approach for datasets based on the sensitive attribute 'race', with the aim of increasing prediction accuracy and reducing racial bias in crime recidivism. The recidivism prediction models, were evaluated by the type of crime, including 'violent crimes', 'property', 'drug' and 'other'. For the 'all crimes', 'Caucasian data set', and 'African American data set' groups, the results still contained bias, although lower than the baseline data. The ratios obtained were 41:59, 34:66, and 46:54.

The [59] study focuses on bias mitigation in deep learning models for classification. The authors point to the need for a systematic analysis of different bias mitigation techniques in-processing with MLP and CNN. Using a dataset that allows the creation of different bias sets, they performed an analysis of the mitigation models recently proposed in the literature. Then they showed the correlation between eligibility and sensitive attributes, the possible presence of bias even without sensitive attributes, and the importance of the initial choice of architecture for model performance.

Whereas [64] presents a focus on the ways in which bias can occur in recommender systems, while addressing the lack of systematic mapping to address unfairness and bias in the current literature. In the experiments, sources of unfairness that can occur in recommendation tasks were mapped, while evaluating whether existing bias mitigation approaches successfully improve different types of fairness metrics. It also presents a mitigation strategy in which the algorithm learns the difference between predicted and observed ratings in subgroups, identifying which is biased and correcting the prediction. The results show that fairness increased in most use cases, but performance for MSE and MAE vary in each case.

The works of [43, 63, 55, 59, 64, 51] have in common the fact that their models were trained in order to mitigate bias from only adjusting the weights of their proposed models. In the work of [65] there is already an attempt to mitigate the bias by neutralizing the sensitive attribute in the model. It has been shown to be possible to make a classification model fairer by removing bias only in its output layer, in a process that occurs during its construction. To this end, a technique was developed where training samples with different sensitive attributes are neutralized, causing the model's dependence on sensitive attributes to be reduced. The main advantage demonstrated by the method is the small loss of accuracy in exchange for improved fairness metrics, without requiring access to the sensitive attributes in the database. In addition, the authors argue that it is possible to increase the quality of the technique by combining it with others, for example by using a fairer basis than the one used in the experiments.

In the solution proposed by [51] the classification detects the item with the highest probability of belonging to the 'target' class of the model; however, there are cases where numerous items have very close probabilities and bias the model, causing an error to propagate across multiple levels. To avoid this, there is a need for a threshold with a minimum degree for the data to be classified and triggers a recalculation of the maximum node probability. The sensitive cost then performs its own probability calculation on the data with the highest degree of membership. These calculations avoid bias caused by using a single probability or over-optimal adjustment caused by using data with no prior context. Hierarchical Precision and Hierarchical Recall, which evaluates the relationship between all descendants of the class and includes Hierarchical F1, Hierarchical Recall, and Hierarchical Precision, were used as metrics. The threshold is adaptive, without requiring user parameters, since metrics exist throughout the classification. Even with fewer samples, it produced results that were superior to the state of the art.

The other works the neutralization of sensitive attributes in an attempt to mitigate model bias is more direct as can be seen in [36, 42, 40, 48] by identifying it beforehand, similarly the investigation of [47] which addresses a new perspective on the concept of justice by determining whether an attribute is sensitive by evaluating it in a Causal Bayesian Networks model. This model examines the direct effects of one characteristic on another and determines whether a sensitive attribute 'A' influences the output 'Y' of a model, producing correlation plots that strive to understand whether or not decisions made were made fairly.

In [36] a pre-existing biased model must be updated to become fair, minimizing unfairness without causing abrupt structural changes. The study uses an adversarial learning technique with the distinction that the generating model is the original network, however the adversarial model comprises an extra hidden layer, rather than a second model, in order to predict which sensitive attribute influenced the generator's decision. The main element of this competition model is that if the discriminator finds the sensitive attribute that influenced the decision the most, it demonstrates dependence on the

generator model, suggesting bias. The generator moves away from the sensitive attributes and performs a classification that does not depend on them, eventually lowering the discriminator's hit rate until it completely loses its predictive ability. The network architecture has three parts: adding an adversarial layer on top of the network, balancing the distribution of classes across the minisets, and adapting sensitive attributes until they are no longer present.

The technique was developed for classification tasks, but can be used for any neural network with biases starting with sensitive attributes [36], achieved better results compared to the state of the art with the metrics addressed.

In the same way as [36], [42] also uses adversarial network for sensitive attribute identification and examines metrics and combinations of techniques for bias mitigation. The study was conducted using basic ANN models and a Split model, which forms the basic model by permuting attribute classes as training criteria in order to identify which one is sensitive. Another model based on the Classifier-Adversarial Network (CAN) architecture, in which the adversarial network predicts the sensitive attribute based on the output of the basic model. Finally, there is the CAN with Embedding (CANE) architecture, which takes as input the output of the basic model as well as the weights created in the penultimate layer. They demonstrated that the models from the Basic RNA architecture can improve accuracy, but not bias. Meanwhile, the models of the CAN and CANE architectures improved accuracy and reduce bias, with CANE being better than CAN.

Still involving adversarial network, in [69] the Adversarial Fairness Local Outlier Factor (AFLOF) method is proposed for outlier detection, combining adversarial algorithms with the Local Outlier Factor (LOF) algorithm, which returns a value indicating whether an instance is an outlier, aiming to achieve a fairer and more assertive result than LOF and FairLOF. It works with the sensitive attributes "Gender", "Age" and "Race". It also uses the AUC-ROC score to measure outlier detection. It results in a fairer and more assertive performance for outlier detection than the previous methods cited, thus achieving a breakthrough in the study of fairness. The work of [52], on the other hand, argues that research on fairness and bias in machine learning focuses only on neural networks, with few publications for other classification techniques. As a result, the author investigated Adversarial Gradient Tree Boosting to rank data and noted that while the adversary progressively loses the reference of the sensitive attribute that led to that prediction.

Another contribution is the adversarial learning method for generic classifiers, such as decision trees in [52]. Comparing numerous state-of-the-art models with the one provided in the paper, which covers two justice metrics. They used varied decision trees in the model given that they make rankings, which are then sent through a weighted average to an adversary, who predicts which sensitive attribute was significant to the final decision. While the adversary is able to detect the sensitive attribute, a gradient propagation occurs, updating the weights in the decision trees and trying to prevent the sensitive attribute from having a direct impact on the ranking.

The [52] model called FAGTB performed well on accuracy and fairness metrics for the COMPAS, Adult, Bank, and Default datasets, outperforming other state-of-the-art models on several of them and considerably outperforming the network adversary. The study leaves certain questions unanswered for future research, such as an adversary using Deep Neural Decision Forests. If this method were used to retrieve the gradient, theoretically, the transparency of the model for the algorithm's decision would be apparent because it consists only of trees. As a final caveat, they acknowledge that the algorithm handles distinct groups well, but the EO and DP fairness metrics do not measure bias between individuals, and is an aspect for improvement.

Following varied work with adversarial learning, the model proposed by [40] called Privileged Information is a technique that trains the model with all the features of the original dataset, including sensitive attributes, and then tests it without these attributes. The model is an in-processing type adjusted with the goal of mitigating unfairness and independent of sensitive attributes, while maintaining its ability to produce accurate predictions, thus respecting the protected information for decision making. Note that in this case, the model fully fits the dataset in an attempt to mitigate bias. The author emphasizes the strength of his model in identifying the best predictor relative to other state-of-the-art work, having the sensitive attributes as optional, and still using Privileged Information.

Whereas [48] avoids model bias by using only data with minimal or, if possible, no sensitive attributes. By applying a noise conditioning operation to the data provided in the model, inducing the model to ignore sensitive attributes, reducing bias. The goal of the model is to create as accurate a representation as possible in the prediction, with fairness. The models used the techniques of logistic regression and Random Forest.

The justice metrics identified in the papers that addressed in-processing are in Table 7, as are the datasets in Table ??.

Mitigation solutions for post-processing were also found, as in [38, 46]. In [38] it proposes a solution for an already formed model, seeking to identify whether certain groups receive discriminatory treatment due to their sensitive attributes. With the identification of discrimination for a group, it is verified whether the sensitive attributes are impacting the model, even if indirectly. The model has a neural network with four fully connected layers of 8 neurons, expressing the weights as a function of the features in order to minimize the maximum average discrepancy function

Fairness Metric	References
Demographic Parity	[59, 42, 65, 69]
Equality of Opportunity	[59, 42, 63, 69]
Equalized Odds	[59, 63, 65]
accuracy	[59, 40, 42, 55]
Disparate Impact	[36]
TPR	[40]
FPR	[40, 55, 36]
FNR	[55, 36]

Table 7: Fairness metrics used in em-processing techniques

between the sensitive attribute classes promoting unfairness mitigation. He applied his mitigation model to a Logistic Classification model. The work allows black-box type models to be mitigated for unfairness, but also by understanding the assigned treatment.

In the [46] study it uses the identification of biases in models developed to recognize the user, where the user can be a human with normal vision, a blind person, or a robot. The identification takes place when answering a question, so NLP is applied. Its bias can be seen in the most frequently asked question "what is this object?", as well as the low image quality compared to the others. Initially, annotations were assigned to the content of the images such as "boy", "package", "grass", "airplane", and "sky". Random Forest, K-Nearest Neighbors, Nave Bayes, and Logistic Regression techniques were used to develop the models. Logistic Regression produced the best results, with 99% on all metrics. The authors found that the algorithms readily recognized the bias in each dataset and provided a means of tracing the origin of the questions and images.

The justice metrics identified in the papers that addressed post-processing are in Table ??, as are the datasets in Table 10.

Datasets	References
Sintetic (normal distribution)	[38]
COMPAS	[38]
VQA	[46]
VizWiz	[46]
CLEVR	[46]

Table 10: Datasets used in post-processing techniques

4 Discussion

All 45 studies examined addressed comparable techniques, case studies, datasets, metrics, and applications.

Adult datasets and COMPAS were used to address the most frequently reported bias identification and injustice mitigation.

In [37]’s work investigated the sources and implications of various types of bias, either in the datasets or in the model. The study investigates bias, offering methods for eliminating it, as well as constructing groups and subgroups that help understand the problem, and discusses general categories such as temporal, spatial, behavioral, posterior, transcendental, and group bias. Specific cases, such as the Simpsons paradox or social behavior bias, are grouped within these categories.

The forms of bias observed by [37] are categorized as follows: dataset bias, model bias, and emergent bias, or pre-processing, in-processing, and post-processing, as previously described. In order to go deeper into these categories, the study [37] splits them into eight broad and 18 particular categories, as well as providing metrics and strategies for resolving each of them.

A frequent concern about the individual-group interaction is that few ML models handle it. According to [2], if a model is biased in rejecting loans to black males, for example, it will increase its database with rejections for this group, reinforcing the bias and initiating a vicious spiral that will reassert itself with each loan denial.

The work [39] focuses on the topic of vicious loops in machine learning, claiming that models may be free of bias in the present but may be biased in the future. To overcome this, he suggests that the model fulfill the Demographic Parity metric, which ensures that the classification of varied groups is constantly converging and that no group is disadvantaged over time.

Except for [38] and [10], the model proposals were primarily white-box classification. The former proposes a model for bias elimination using Multi-Differential Fairness by integrating in-processing and post-processing, whereas the latter proposes that the focus of algorithm transparency should be on the output rather than the whole decision-making process of the algorithm.

According to the works reviewed, sensitive attributes are defined as elements that should not directly affect the prediction of a model, such as color, race, sex, nationality, religion, and sexual preference, among others. According to US laws such as the *Fair Housing Act* (FHA) and the *Equal Credit Opportunity Act* (ECOA) [80], sensitive attributes should never favor, harm, or alter the outcome of individuals and groups in decision-making processes such as hiring or a court sanction. There is also the fact that all techniques and tools confirm the importance of sensitive attributes in mitigating biases, because for the identification of bias there is the need for the indication of a sensitive attribute, and the mitigation of bias will be based on this identification, remembering that the identification is done through a justice metric.

As for the datasets, 25 datasets were identified, most of them with sensitive attributes such as demographic data, and the ones that did not have any were for studies in the area of image enhancement, when not associated with face recognition. The datasets address aspects related to criminality, the selection and approval process of individuals, financial issues of bank credit, product pricing, health and medical diagnosis, face recognition and image enhancement, and synthetic datasets.

About the metrics of justice, the most used are EO, EOO and DP, as observed in Table 4. Highlighting the importance of statistical metrics, difference metrics, and classification metrics, as several papers have used them as criteria for fairness.

Among the bias mitigation and identification tools: FairLearn and AIF360, weren't used in any practical studies. The topic of identifying bias in the data and the model was also addressed, with the Aequitas tool being the most frequently mentioned.

As for the mitigation techniques, pre-processing techniques for rebalancing the data were addressed; in the in-processing techniques, such as regularizing the model, addressing levels of elimination of the sensitive attribute, with some possible approaches, such as being identified before training the model, during model training, not using it in training or disregarding it completely, training with all attributes so that the model can adjust itself through a loss function. The post-processing techniques, on the other hand, aim to discover which sensitive attribute had an impact on the model result, rebalancing the prediction.

The most common justice metrics such as EO, EOO and PD are covered with the FairLearn, AIF360, Aequitas and Responsible AI tools, with the exception of EOO not covered by FairLearn.

In [6]'s work highlights some research gaps such as the wide varieties of justice metrics as a factor hindering which one best fits each case, lacking a comprehensive formal and comparative study of the strengths and limitations of each of the metrics. It also highlights that a formal study of the techniques with the strengths and limitations of each is lacking. It also addresses the need for state-of-the-art recommendation system techniques. It highlights that there is still an absence of studies on the economic and social consequences of biases in high-risk systems. The work of [18] attempts to elucidate some of these gaps, from the point of view of organizations and individuals, but without addressing the technical aspects of such solutions, when it highlights the importance of the socio-technical nature of biases in algorithms, the need to understand the social processes and contexts impacted by the use of biased information and algorithmic technologies.

Finally, all studies have addressed the algorithm's transparency, or the capacity to explain the decision-making process that caused the model to classify a certain individual or group the way it did. This method must fundamentally explain either the local decision, which includes the classification of a single individual, or the global decision, which verifies the whole algorithm process. The relevance of transparency is to make it explicit to a customer, company, or court that the model does not consider sensitive attributes and does not discriminate against a specific group, just as it becomes possible to attribute responsibility to the model's developers if the model is biased.

5 Final considerations

The objective of this study was to examine the latest existing knowledge on bias and unfairness in machine learning (ML) models with the RSL methodology and a bibliometric analysis. Thus, the paper was to answer questions Q1 and Q2 in the 2 section.

To answer question Q1, the findings demonstrate that there is a focus on bias and unfairness identification methods for ML technologies, with well-defined metrics in the literature, such as fairness metrics, featured in tools, datasets, and bias mitigation techniques. This diversity ends up not defining the most appropriate approach for each context given that different solutions can be observed for the same problem, leading to a lack of definition about which one would be the most appropriate, without a generic solution for the identification and mitigation of biases. The vagueness raised in Q1's answer opens up aspects to be considered in Q2's answer.

To answer question Q2, where the existing opportunities should be highlighted, there is very limited support for black-box models, which contrasts with the abundance of information for white-box models. The need for transparency and explainability of ML algorithms, as well as the defining and preservation of sensitive attributes was also emphasized, with the selected datasets acting as a basis for research addressing the identification and mitigation of bias and unfairness.

As opportunities for future work, we conclude that more research is needed to identify the techniques and metrics that should be employed in each particular case in order to standardize and ensure fairness in machine learning models. For a definition on which metric should be used for each use case, more specific studies should be conducted under different architectures and sensitive attributes. This analysis would allow the context to define the most appropriate metric to identify bias in protected groups, and whether the sensitive attribute can be a relevant element in defining the fairness metric for a given context. It was observed that, in a given dataset, the metrics do not present uniform results, pointing to different categories of bias and their context-related particularities.

References

- [1] Yogesh K Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57:101994, 2021.
- [2] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8, 2021.
- [3] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.
- [4] David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020.
- [5] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.
- [6] Tommaso Di Noia, Nava Tintarev, Panagiota Fatourou, and Markus Schedl. Recommender systems under european ai regulations. *Communications of the ACM*, 65(4):69–73, 2022.
- [7] Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews. *IEEE Signal Processing Magazine*, 38(6):84–95, 2021.
- [8] Candice Schumann, Jeffrey S Foster, Nicholas Mattei, and John P Dickerson. We need fairness and explainability in algorithmic hiring. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1716–1720, 2020.
- [9] Jamil Ammar. Cyber gremlin: social networking, machine learning and the global war on al-qaida-and is-inspired terrorism. *International Journal of Law and Information Technology*, 27(3):238–265, 2019.
- [10] William Seymour. Detecting bias: does an algorithm have to be transparent in order to be fair? *Jo Bates Paul D. Clough Robert Jäschke*, page 2, 2018.
- [11] Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. Explainable ai in industry. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3203–3204, 2019.
- [12] R.K.E. Bellamy, A. Mojsilovic, S. Nagar, K.N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K.R. Varshney, Y. Zhang, K. Dey, M. Hind, S.C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, and S. Mehta. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4-5), 2019. cited By 26.

- [13] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.
- [14] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.
- [15] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122*, 2020.
- [16] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.
- [17] A. Nielsen. *Practical Fairness: Achieving Fair and Secure Data Models*. O’Reilly Media, Incorporated, 2020.
- [18] Nima Kordzadeh and Maryam Ghasemaghahi. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409, 2022.
- [19] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1452, 2022.
- [20] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.
- [21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.
- [22] Marley Bacelar. Monitoring bias and fairness in machine learning models: A review. *ScienceOpen Preprints*, 2021.
- [23] Agathe Balayn, Christoph Lofi, and Geert-Jan 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, 739–768, 2021.
- [24] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.
- [25] Harini Suresh and John Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1810.08810*, 2018.
- [26] Sascha Kraus, Matthias Breier, and Sonia Dasí-Rodríguez. The art of crafting a systematic literature review in entrepreneurship research. *International Entrepreneurship and Management Journal*, 16(3):1023–1042, 2020.
- [27] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews*, 10(1):1–11, 2021.
- [28] Tiago Palma Pagano, Victor Rocha Santos, Yasmin da Silva Bonfim, José Vinícius Dantas Paranhos, Lucas Lemos Ortega, Paulo Henrique Miranda Sá, Lian Filipe Santana Nascimento, Ingrid Winkler, and Erick Giovanni Sperandio Nascimento. Machine learning models and videos of facial regions for estimating heart rate: A review on patents, datasets, and literature. *Electronics*, 11(9):1473, 2022.
- [29] Andrew Booth, Anthea Sutton, and Diana Papaioannou. *Systematic Approaches to a Successful Literature Review*. SAGE, 2016.
- [30] Eliza M Grames, Andrew N Stillman, Morgan W Tingley, and Chris S Elphick. An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, 10(10):1645–1654, 2019.
- [31] SB Patil. Global library & information science research seen through prism of biblioshiny. *Stud. Indian Place Names*, 40:158–170, 2020.
- [32] Massimo Aria and Corrado Cuccurullo. bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of informetrics*, 11(4):959–975, 2017.
- [33] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094, 2022.
- [34] Julia Stoyanovich, Bill Howe, and HV Jagadish. Responsible data management. *Proceedings of the VLDB Endowment*, 13(12):3474–3488, 2020.

- [35] Octavio Loyola-Gonzalez. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 2019.
- [36] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2412–2420, 2019.
- [37] Chen Jinyin, Chen Yipeng, Chen Yiming, Zheng Haibin, Ji Shouling, Shi Jie, and Cheng Yao. Fairness research on deep learning. *Journal of Computer Research and Development*, 58(2):264, 2021.
- [38] Xavier Gitiaux and Huzefa Rangwala. mdfa: Multi-differential fairness auditor for black box classifiers. In *IJCAI*, pages 5871–5879, 2019.
- [39] Benjamin Paaßen, Astrid Bunge, Carolin Hainke, Leon Sindelar, and Matthias Vogelsang. Dynamic fairness – breaking vicious cycles in automatic decision making. In *ESANN*, pages 477–482, 2019.
- [40] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 2017.
- [41] Ke Yang, Biao Huang, Julia Stoyanovich, and Sebastian Schelter. Fairness-aware instrumentation of preprocessing pipelines for machine learning. *Workshop on Human-In-the-Loop Data Analytics (HILDA’20)*, 2020.
- [42] Jack J Amend and Scott Spurlock. Improving machine learning fairness with sampling and adversarial learning. *Journal of Computing Sciences in Colleges*, 36(5):14–23, 2021.
- [43] Barbara Martinez Neda, Yue Zeng, and Sergio Gago-Masague. Using machine learning in admissions: Reducing human and algorithmic bias in the selection process. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1323–1323, 2021.
- [44] Reginald Bryant, Celia Cintas, Isaac Wambugu, Andrew Kinai, and Komminist Weldemariam. Analyzing bias in sensitive personal information used to train financial models. *arXiv preprint arXiv:1911.03623*, 2019.
- [45] Annunziata Paviglianiti and Eros Pasero. Vital-ecg: a de-bias algorithm embedded in a gender-immune device. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pages 314–318. IEEE, 2020.
- [46] Anubrata Das, Samreen Anjum, and Danna Gurari. Dataset bias: A case study for visual question answering. *Proceedings of the Association for Information Science and Technology*, 56(1):58–67, 2019.
- [47] Silvia Chiappa and William S Isaac. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pages 3–20. Springer, 2018.
- [48] Mattia Cerrato, Roberto Esposito, and Laura Li Puma. Constraining deep representations with a noise module for fair classification. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 470–472, 2020.
- [49] Sheng Shi, Shanshan Wei, Zhongchao Shi, Yangzhou Du, Wei Fan, Jianping Fan, Yolanda Conyers, and Feiyu Xu. Algorithm bias detection and mitigation in lenovo face recognition engine. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 442–453. Springer, 2020.
- [50] Yogesh K Dwivedi, Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, et al. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, page 101994, 2019.
- [51] Weijie Zheng and Hong Zhao. Cost-sensitive hierarchical classification via multi-scale information entropy for data with an imbalanced distribution. *Applied Intelligence*, pages 1–13, 2021.
- [52] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fair adversarial gradient tree boosting. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1060–1065. IEEE, 2019.
- [53] Claudio Feijóo, Youngsun Kwon, Johannes M Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. Harnessing artificial intelligence (ai) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications policy*, 44(6):101988, 2020.
- [54] Sébastien Gambs. Privacy and ethical challenges in big data. In *International Symposium on Foundations and Practice of Security*, pages 17–26. Springer, 2018.
- [55] Bhanu Jain, Manfred Huber, Leonidas Fegaras, and Ramez A Elmasri. Singular race models: addressing bias and accuracy in predicting prisoner recidivism. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 599–607, 2019.
- [56] Pascal D König and Georg Wenzelburger. When politicization stops algorithms in criminal justice. *The British Journal of Criminology*, 61(3):832–851, 2021.

- [57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.
- [58] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. Evolution and impact of bias in human and machine learning algorithm interaction. *Plos one*, 15(8):e0235502, 2020.
- [59] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabanian, and Sina Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021.
- [60] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pages 4721–4732. PMLR, 2021.
- [61] Markos Georgopoulos, James Oldfield, Mihalīs A Nicolaou, Yannis Panagakis, and Maja Pantic. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307, 2021.
- [62] Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916, 2021.
- [63] Sandro Radovanović, Andrija Petrović, Boris Delibašić, and Milija Suknović. Enforcing fairness in logistic regression algorithm. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–7. IEEE, 2020.
- [64] Ashwathy Ashokan and Christian Haas. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management*, 58(5):102646, 2021.
- [65] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34, 2021.
- [66] Dana Pessach and Erez Shmueli. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert Systems with Applications*, 185:115667, 2021.
- [67] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36:25–34, 2021.
- [68] Michelle Seng Ah Lee and Jatinder Singh. Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 704–714, 2021.
- [69] Shu Li, Jiong Yu, Xusheng Du, Yi Lu, and Rui Qiu. Fair outlier detection based on adversarial representation learning. *Symmetry*, 14(2):347, 2022.
- [70] Sidney K D’Mello, Louis Tay, and Rosy Southwell. Psychological measurement in the information age: Machine-learned computational models. *Current Directions in Psychological Science*, 31(1):76–87, 2022.
- [71] Michele Fontana, Francesca Naretto, Anna Monreale, and Fosca Giannotti. Monitoring fairness in holda. In *Proceedings of conference on hibrid human-artificial intelligence*, page 66.
- [72] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. Compas dataset, 2016.
- [73] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- [74] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference*, pages 5–12. EUROSIS-ETI, 2008.
- [75] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.
- [76] Timothy B Creedon, Samuel H Zuvekas, Steven C Hill, Mir M Ali, Chandler McClellan, and Judith G Dey. Effects of medicaid expansion on insurance coverage and health services use among adults with disabilities newly eligible for medicaid. *Health Services Research*, 2022.
- [77] Eduardo De-La-Hoz-Correa, Fabio Mendoza Palechor, Alexis De-La-Hoz-Manotas, Roberto Morales Ortega, and Adriana Beatriz Sánchez Hernández. Obesity level estimation software based on decision trees. *Journal of Computer Science*, 2019.
- [78] Elaine Fehrman, Awaz K Muhammad, Evgeny M Mirkes, Vincent Egan, and Alexander N Gorban. The five factor model of personality and evaluation of drug consumption risk. In *Data science*, pages 231–242. Springer, 2017.
- [79] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3730–3738, December 2015.
- [80] US Congres. 15 u.s.c. 1691 : Equal credit opportunity act, 1974.