Strategies to improve fairness in artificial intelligence: A systematic literature review

Abstract: Decisions based on artificial intelligence can reproduce biases or prejudices present in biased historical data and poorly formulated systems, presenting serious social consequences for underrepresented groups of individuals. This paper presents a systematic literature review of technical, feasible, and practicable solutions to improve fairness in artificial intelligence classified according to different perspectives: fairness metrics, moment of intervention (pre-processing, processing, or post-processing), research area, datasets, and algorithms used in the research. The main contribution of this paper is to establish common ground regarding the techniques to be used to improve fairness in artificial intelligence, defined as the absence of bias or discrimination in the decisions made by artificial intelligence systems.

Keywords: Artificial intelligence, Fairness, Fairness techniques, Fairness metrics, Systematic literature review.

1 Introduction

Artificial Intelligence (AI) is increasingly present in our daily decisions, influencing nearly every field such as marketing, finance, justice, medicine, sports, and libraries (Cortez et al., 2022; Pimenta et al., 2023; Yan et al., 2023). This growing presence impacts the lives of thousands, shaping the opportunities we encounter, our perceptions of the world, and our understanding of ourselves. A notable recent development in AI is ChatGPT, a Large Language Model (LLM). This generative AI model stands out from traditional AI models due to its unique methodologies, capabilities, and applications, such as learning from unlabeled data, utilizing large-scale datasets to identify patterns, and generating new data that resembles its training data, as well as its advanced natural language understanding and generation skills.

However, AI algorithms may learn incorrectly from data, which might result in the spread of false information online. In addition to various technical approaches underlying the implementation of fairer algorithms, this must be complemented with information literacy and information ethics programs and services (Onifade, 2023). If they malfunction or are abused, they will present serious risks to people and society. People's lives may depend on the effectiveness and safety of these algorithms. As they became more sophisticated, it may be particularly hard to understand the complex nature of how, when, and why these learning algorithms would fail. This concern is so critical that some authors argue that these algorithms should not be available without authorization from a government agency, which should act as a centralized expert regulator that creates standards, guidelines and expertise in collaboration with industry (Tutt, 2020).

Drawing on Cathy O'Neil's concept in “Weapons of Math Destruction” we're constantly categorized and analyzed, with our lives reduced to vast datasets encompassing everything from zip codes to online habits, purchases, and social media connections. These mountains of data feed algorithms that, if flawed, can create vicious cycles. Such algorithms reinforce existing biases, perpetuating issues like racism, sexism, and classism (O’Neil, 2016). Big tech companies have already been involved in algorithmic discrimination controversies. For instance, Amazon used a recruitment system that penalized women's resumes, and Google Photos misidentified black people as gorillas (Pessach & Shmueli, 2020). Bias against dark-skinned people is commonly reproduced, with examples cited by Cardenas & Vallejo-Cardenas (2019). These include black pedestrians being less likely to be identified by autonomous cars and therefore more prone to accidents, soap dispenser sensors working with greater difficulty on black skin, and facial recognition systems in retail stores incorrectly identifying black individuals as potential attackers more often than whites. An excellent example is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system in the USA. This tool assessed the risk of re-offending, but a crucial flaw emerged: black defendants were flagged as high-risk at twice the rate of white defendants with similar criminal histories. This racial bias in COMPAS led to harsher sentences for many black individuals (Mehrabi et al., 2021).

Given the relevance of the topic and its social impact, many papers are frequently published addressing the subject (Dash et al., 2019; Edizel et al., 2020; Hu et al., 2019; Lin et al., 2021; Mokhtari et al., 2020; Obaidat et al., 2021; Rudin & Ustunb, 2018; Sahu & Singh, 2019), and major conferences such as ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT); Fairness, Accountability, and Transparency in Machine Learning (FAT/ML); and International Workshop on Equitable Data & Technology (FairWare) bring together academics and practitioners interested in exploring ways to build fairer ethical and transparent systems. This paper presents an investigation, through a Systematic Literature Review (SLR), of studies that contain technical approaches or tools capable of addressing the reasons that enable biases and prejudices to be introduced in AI systems, improving their fairness. It is structured as

follows: after the introduction section, a brief introduction to the theme is given in section 2, section 3 presents the methodology used, section 4 presents the results obtained and discussion and in section 5 some final considerations about the study are presented in the conclusion.

2 Background

The landscape of artificial intelligence has undergone a radical transformation since the publication of Alan Turing's seminal paper, “Computing machinery and intelligence” (Von Hohendorff & Kaini Lazzaretti, 2021). The era of Big Data and the Internet of Things (IoT) in which about 79 zettabytes of data were generated in 2021, and with projections of 180 zettabytes by 2025 (DOMO, 2022). This data explosion is what has fueled algorithms that recognize patterns and reveal regularities present in such data, on which the decision-making process can rely.

2.1 Discrimination in AI

Race, gender, religion, and age are examples of the so-called protected variables, or sensitive attributes, and a specific group of individuals may be harmed due to the use of such variables when they lead to direct discrimination, or disparate treatment (Mehrabi et al., 2021). Some decisions, however, are not explicitly based on sensitive variables, but still generate results that disproportional harm or benefit a certain group of individuals, and this disadvantage is called indirect discriminations or disparate impact. Such decisions may be based on variables strongly correlated with some sensitivity variable, such as postal code and social class, or salary and gender. These correlated variables are called proxies (Mehrabi et al., 2021). Unlike the disparate treatment, the disparate impact itself is not illegal, and some permissions can be made according to, for example, the business needs or hiring decisions, thus this is the main object of investigation in algorithmic fairness (Feldman et al., 2015).

An algorithm is only as good as the data used by it (Solon Barocas & Selbst, 2016). A model based on ML is trained to behave as the examples it was exposed to, so when model training data contains biases, they will influence the learned rules and its rationale. Thus, for historical reasons, human biases are introduced into ML systems, harming those who are under- or over-represented in that data (Solon Barocas & Selbst, 2016). Many types of biases can be found in algorithm (Srinivasan & Chander, 2021), and the categorization of such biases are important, because they can motivate future solutions according to each specific type (Mehrabi et al., 2021).

Solon Barocas & Selbst (2016) explained that the five main reasons that may allow discrimination in ML are: 1) incorrect definitions of the objective variable and the class labels; 2) the lack of representativeness of groups and class labelling errors in the training data; 3) the lack of understanding and selection of the variables involved; 4) the lack of understanding and selection of proxies involved; and 5) the masking of decision makers’ prejudice views, that may or may not be intentional. All these reasons are stages in a subjective process of business understanding and problem definition, where the data used are reductive representations of a real-world phenomenon, which is infinitely more complex and specific, and these representations may not capture all the details involved in the issue we wish to solve.

Another important factor to be observed is the minimization of average errors that tend to adjust to the major group, that is, when the distribution of variables between the different group is different, these variables will have different relationships with the objective variable, therefore, when training a classifier that does not distinguish groups to minimize the overall error, it will only fit the majority population group, and this leads to a different and larger distribution of mean error in the minority group, making the model learn less about the minority group (Chouldechova & Roth, 2018).

2.2 Definition of fairness in AI

Based on the literature (Ferrer et al., 2021; Giovanola & Tiribelli, 2023; Mehrabi et al., 2021), fairness in AI can be defined as the absence of bias or discrimination in the decisions made by AI systems.

2.3 Improving fairness in AI

In AI, fairness is recognized and enforced through the establishment of metrics, that is, mathematical equations inserted at certain stages of the model development process that must be met. It is important to keep in mind that when seeking to achieve fairer systems, we may compromise the performance or the accuracy of the system, since the nature of the model that defines it is to discover and reproduce the patterns identified in the data, which patterns identified in the data, which in turn may be carrying biases and prejudices. By meeting the metric, the system will end up admitting some errors in its performance. These errors, however, should not significantly compromise the accuracy of the model and vice versa, so we must always seek the best trade-offs between fairness and accuracy in a model.

2.3.1 Fairness metrics for AI

There is a wide debate about which definitions or metrics of fairness are preferred (Verma & Rubin, 2018). Table 1 summarize the main fairness metrics found in the literature, taking the work developed by Verma & Rubin (2018) as a starting point.

Table 1 Fairness metrics adapted and extended from Verma & Rubin (2018)

Group	Metric
Based on predicted outcome	Group fairness; Statistical parity; Equal acceptance rate; Demographic Parity) (Shimao et al., 2022).
	Conditional Statistical Parity
Based on predicted outcome and real label	Predictive Parity
	False Positive error rate balance; Predictive equality
	False Negative error rate balance; Equal Opportunity; True positive parity
	Equalized odds; Disparate mistreatment
	Conditional use accuracy equality
	Overall accuracy equality
	Treatment equality
Based on predicted probabilities and real label	Test-fairness; Calibration;
	Well-calibration
	Balance for positive class
	Balance for negative class
Distance metrics	Decision boundary covariance (Zafar et al., 2017)
	μ_1 and μ_2 (Edizel et al., 2020)
Individual or Similarity-based metrics	Causal discrimination
	Fairness through unawareness
	Fairness through awareness
	Sample Distortion (Calmon et al., 2017)
	Entropy index (Speicher et al., 2018)
	Theil Index (Ahmed et al., 2021)
Casual reasoning	Counterfactual fairness
	No unresolved discrimination
	No proxy discrimination
	Fair inference
Fair representation	Consistency (Zemel et al., 2013)

2.3.2 Fairness in the AI process stages

Generally, techniques to mitigate fairness can be implemented in the following stages of the AI process: preprocessing, processing (also designated the in-processing stage) and postprocessing. In the preprocessing stage, fairness-focused techniques aim to balance the training dataset for protected and unprotected groups. This can involve resampling, adjusting weights assigned to data points, or even modifying class labels. The goal is to eliminate correlations between variables and sensitive attributes (S. Barocas et al., 2019). The approaches that occur during the processing phase, act on the algorithm itself, reformulating the problem and incorporating the discrimination behavior of the model in the objective function through regularizations and restrictions. In the post-processing phase, the focus is on the results produced by the model, adjusting the trained classifier to satisfy the fairness constraints. For black-box models, one approach to fairness is to ensure proportional outcomes for protected and unprotected groups. This can involve adjusting classifications near decision thresholds, either promoting or demoting them to achieve balance (Ntoutsi et al., 2020).

2.3.3 Fairness improving techniques for AI

The techniques for fairness improvement vary according to the combination of fairness metrics, the base algorithms, and the processing stages at which they are executed. This paper summarize important techniques proposed in the literature according to some important works (Caton & Haas, 2020; Friedler et al., 2019; Mehrabi et al., 2021; Pessach & Shmueli, 2020). Feldman et al. (2015) proposed a data preprocessing method that achieves fairness without altering the training labels. They modified the data's attributes to ensure similar distributions for privileged and unprivileged groups. This was measured using the disparate impact metric. By achieving this balance, classification algorithms become less likely to make decisions based on group affiliation. On the other hand, Kamiran and Calders (2012) processed the data by adjusting label weights, with this method being particularly effective near decision thresholds in the training set. These samples are more susceptible to misclassification, especially when dealing with imbalanced datasets. Assigning higher weights to these labels increases their influence on the model, ultimately leading to fairer decision boundaries. The technique is based on a scoring algorithm that measures fairness based on a demographic parity (DP) metric. Calders and Verwer (2010) used a Naive Bayes algorithm, which trained two different models for the sensitive attribute values, and through small changes in the observed probabilities sought to reduce the demographic parity measure. The models were then recycled, forming a third model, i.e., changing the operation of the algorithm to achieve demographic parity. Hardt et al. (2016) proposed reversing some final decisions of classifier algorithms to increase equalized chances and equal opportunities, metrics proposed in their own study. There's a well-established consensus in the field that there is no single fairness metric to universally prioritize. Like the lack of a one-size-fits-all approach for intervention timing, the optimal choice depends on several factors. These include the trade-offs between fairness and performance, the training data's characteristics, the chosen algorithm, and most crucially, the legal, ethical, and social context where the model will be deployed.

3 Methodology

To document the evidences found in the research in a transparent, reproducible and systematic way the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021) was used.

3.1 Search strategy

The search terms used were "technique", "fairness" and "artificial intelligence". To ensure a comprehensive and high-sensitivity search, we also included variations and synonyms of these words. The search was conducted for publications from the last 5 years (2019-2023) on the Web of Science (WoS) and SCOPUS database. Boolean logical operators and the asterisk symbol * were used to truncate the suffix, and the search strategy was restricted to the title in the case of techniques and fairness related words and extended to the abstract and keywords in the case of the AI related words.
The search expression for used for the WoS was:

```
TI = ((technique* OR method* OR tool* OR way* OR action* OR framework* OR approach* OR strateg* OR syste*) AND (*fair* OR *bias* OR *justice*) AND ("machine learning" OR "artificial intelligence" OR "big data" OR "data mining" OR "LLM" OR "large language model") )
```

The search expression for used for the SCOPUS was:

```
TITLE (technique* OR method* OR tool* OR way* OR action* OR framework* OR approach* OR strateg* OR syste*) AND TITLE (*fair* OR *bias* OR *justice*) AND TITLE ("machine learning" OR "artificial intelligence" OR "big data" OR "data mining" OR "LLM" OR "large language model")
```

3.2 Eligibility criteria

For this review, articles describing primary studies with technical approaches that seek to improve fairness in AI were chosen. Given the contemporaneity of the subject, a search was made for articles published between the years 2019 to 2023 of any methodological type, and without area restriction. However, due to the particularity and relevance of the topic, the type of publication was limited to peer-reviewed scientific journal articles, ruling out any grey literature such as reports or traditional press articles. The search comprised only English language articles, as this is the predominant language of the publications. Other exclusion criteria include papers referring to secondary research or other narrative and systematic reviews, duplicate results, and articles that were not available online; papers that addressed fairness and

AI in other contexts that are not of interest to this study, such as resource allocation or computer system component tasks. It were also excluded papers that addressed and discussed the subject but did not suggest any intervention or proposal to improve the problem, papers in which the proposal for mitigating biases and promoting algorithmic fairness focused on socio-technical efforts, such as normalizations, guidelines, and ethical principles.

3.3 PRISMA Flowchart

The search in the WoS database returned 185 records, while in Scopus it returned 236 records. Of these records, a total of 27 papers were selected as eligible, determined through various steps in the research process illustrated in Figure 1, the PRISMA flowchart (Page et al., 2021).

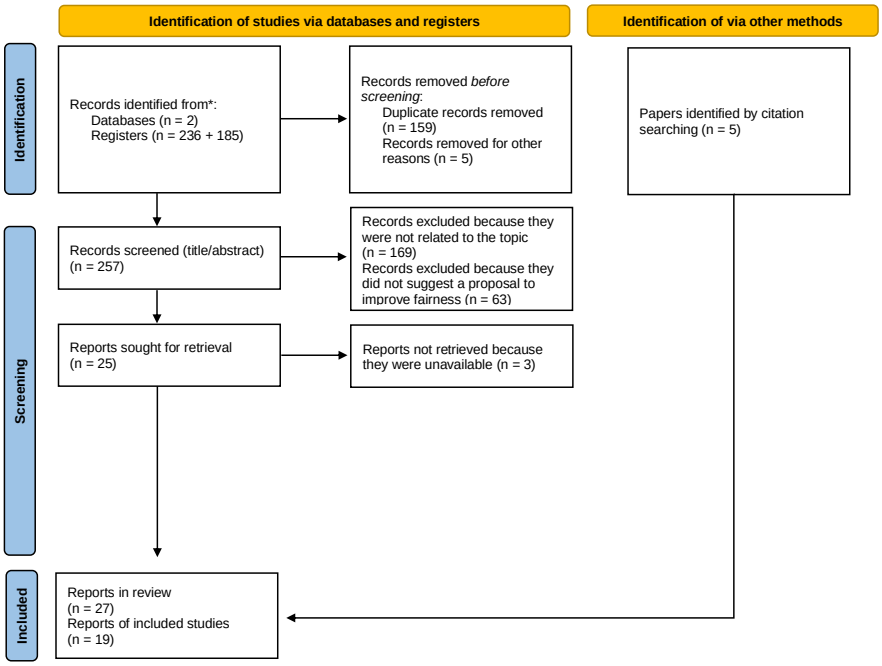

Figure 1. PRISMA flowchart of the SLR performed.

Of the selected papers, 12 papers were discarded upon full reading, as they did not focus on improving fairness. Furthermore, an additional 5 papers were identified from the references of the initial 27 papers. This process ended with a total of 19 papers included for the systematic literature review (SLR).

4. Results

The results of the selected papers will be described in a narrative form, summarized in a table, and discussed according to the characteristics they have in common.

4.1 Description of results

This study, as outlined in the methodology section, aims to identify technical solutions for enhancing fairness. In this sense, Lin et al. (2021) suggested a two-dimensional framework for evaluating existent AI-based interventions and exploring new and promising approaches relating to the recruitment and selection field: Blendoor, Eightfold, Entelo, Hire Vue, IBM Watson, Interviewing.io, Pymetrics, Textio Hire, Tengai, Equal Reality, and Vantage Point. They classified each tool according to its descriptive, predictive, and prescriptive information, and according to input-based and output-based interventions, or cognition-based interventions, thus drawing a map for identifying the best tool according to each specific case.

In order to promote the popularity of using fairness metrics and bias mitigation techniques, Bellamy et al. (2019) offered an open source toolkit in Python called AI Fairness 360. The package provides datasets and measurement classes, and pre-, in-, and post-processing algorithms that maintain code quality and make it easier to understand model validation. AI

Fairness 360 also offers an interactive web experience to explain concepts, documentation, guides and tutorials for developers and researchers.

The research of Ahmed et al. (2021) continued the previous work, using AI Fairness 360 to compare the fairness achieved before and after the experimentation of the techniques, on the US Employment Demographics dataset and confirmed the effectiveness of the package demonstrating good results achieved mainly with the post processing algorithms based on Random Forest (RF) and Logistic Regression (LogR).

The study of Dash et al. (2019) also presented pre, in and post-processing techniques for bias mitigation, but specifically for text summarization systems. The FairSumm algorithm was optimized to seek quality in the summarization of texts and meet the criteria of fairness applied as matroid constraint (concept used to generalize the notion of linear independence of the matrices), during the process of the model, while the ClasswiseSumm algorithm was optimized to group texts based on different classes and then to summarize each group separately, and the RefaSumm algorithm, to perform fair reclassification of texts based on some measure of fairness. The goal of the work was to perform quality summarization of texts and to ensure that all groups have their opinions represented in the summary.

Also seeking the protection of sensitive data, Hu et al. (2019) developed a framework for distributed learning, which relies on the participation of the party that holds the private property of the data (called in the study the third party), and the institution running the data center that develops the ML model. The data center generates fair random hypotheses from Gaussian distributions of the non-demographic data, obtains predictions, and sends them to the third party, which in turn owns the demographic data and estimates the correlation between the predictions and this data, confirming whether the hypothesis generated by the data center is fair or not. The next step then is to use these fair hypotheses to generate the private models. The study tested the proposed method to redesign four types of non-private algorithms previously used in private versions, these being: 1) Distributed Fair Ridge Regression (DFRR); 2) Distributed Fair Kernel Ridge Regression (DFKRR); Distributed Fair Logistic Regression (DFGR); and Distributed Fair PCA (DFPCA).

The proposal by Zafar et al. (2017) follows another line, presenting a mechanism to design fair classifiers based on LogR and Support Vector Machines (SVM) by promoting a new measure of unfairness decision boundary. This metric, calculated as the covariance between the sensitive attribute and the distance to the decision boundary, assesses the independence between the model's predictions and the sensitive variable. In simpler terms, it measures how closely the model's predictions align with the sensitive attribute, capturing potential bias at a group level. This metric derives from two complementary formulations of constraints for classifier training. While one formulation seeks to maximize accuracy under fairness constraint for compliance with discrimination policy or law (p% rule), the other seeks to maximize fairness under accuracy constraint for ensuring business necessity. This measure ensures fairness with respect to one or more sensitive attributes, for simultaneous treatment of direct and indirect discrimination.

Another proposal that explores the trade-offs between fairness and accuracy was made by Valdivia et al. (2021). The authors suggested a method based on a multi-objective algorithm, Non dominated Sorting Genetic Algorithm (NSGA-II) to guide a classifier, in this case decision trees, for being understandable and transparent (Valdivia et al., 2021). The trees seek to obtain the best trade-offs in accuracy and fairness by learning the best combination of the Hyper criteria parameters, maximum depth, minimum number of samples to split a node, total number of leaves and the weight of each class, and provides the best feasible solutions through a Pareto front (Valdivia et al., 2021).

The Super Sparse Linear Integer Model (SLIM) and Risk-calibrated Super Sparse Linear Integer Model (RiskSLIM) algorithms by Rudin & Ustunb (2018) also work with fairness constraints. SLIM is optimized for the trade-off between True Positives Rates (TPR) and the False Positive Rates (FPR), and predictions, and are based on whether the scores exceed a threshold value or not (if score total >1= yes, and if the score is <1= no). The choice of TPR and FPR depends on the application. For a medical screening for example, it is better to seek for a higher FPR, or false alarm. While RiskSLIM, is a risk score system, calibrated by risk, that is, the risk predicted by the model is the same existing in the data. RiskSLIM does not look for trade-offs between TPR and FPR, instead it seeks to achieve the best TPR. The authors argue that the indicated models achieve good performance results and have the advantage of being transparent and explainable, being good choices over black box models, where the decision rules are not explicit (Rudin & Ustunb, 2018).

A different approach is offered by Zhang et al. (2018), who used the performance of multiple networks with competing objectives for bias mitigation. The core of this system involves two models. The first, a predictor model, uses gradient-based methods to adjust weights and minimize losses while predicting the target variable Y based on independent variables X. This Y prediction then feeds into a concurrent network. This second network aims to predict the protected variable Z, but with a twist. Its additional inputs depend on the specific fairness metric we want to achieve. For instance, if we aim for Equalized Odds (EO), the concurrent network gets access to the true labels for Z. This allows it to learn the relationship between Y and Z, essentially forcing the predictor model to avoid decisions that disproportionately impact groups based on the protected variable. When the stipulated metric is reached, the training of the opponent model ends.

Two elected studies dealt specifically with movie recommendation systems. Sahu & Singh (2019) model brings an algorithm based on collaborative filtering, which uses user variables and movie variables, and mixes two types of recommendation. One that reinforces user preferences and one that is inversely correlated to those preferences, but with good quality. These variables are learned by a stochastic gradient descent extension and used to measure the Root Means Square Error (RMSE). The idea is to promote diversity and try to pierce the "bubble" of personal bias in recommended content by exposing the user to new possibilities (Sahu & Singh, 2019).

As AI systems do not use conventional code rules, but data, to predict future behavior, Obaidat et al. (2021) said that such systems are more susceptible to tampering with adversaries' data, which can flood the system with false data and consequently generate unreliable decisions. To address the problem, the authors proposed the Minimization AI bias applying Random Sampling Technique (MAIRST) method, which combines random sampling to train data in convolutional neural network (CNN) algorithms. The proposed approach involves a dynamic testing process to simulate an adversary attack. Here, a neural network (NN) is exposed to test data that has been randomly tampered with. This simulates a real-world scenario where an attacker might try to infiltrate the system and negatively impact its performance. To mitigate this threat, a second line of defence is implemented. MAIRST, a filtering technique, is applied to the tampered test set. This process aims to identify and remove as much of the compromised data as possible. The result, a cleaned test set, is then used for the final evaluation of the model's ability to recognize and classify garment images (Obaidat et al., 2021).

Calmon et al. (2017) suggested a framework for probabilistic data transformation for discrimination reduction. Through random mapping, the original data set is transformed into a new data set, which is used to train the model and similarly transform the data on which the model is applied. This random mapping must satisfy the discrimination control, distortion limit and utility preservation detailed in the study.

Krasanakis et al. (2018), on the other hand, proposed an Adaptive Sensitive Reweighting (ASR) scheme that uses a Convex Underlying Label Error Perturbation (CULEP) model to estimate underlying label distributions with which it adapts the weights to achieve good trade-off between accuracy and elimination of direct and indirect discrimination. This method operates under the assumption that the training data contains hidden class labels beyond the ones directly available. By predicting these underlying labels, the model aims to achieve partial classification while adhering to a fairness objective.

The model of Edizel et al. (2020) focuses on an algorithm for post-processing the recommendation matrix. FaiRecSys seeks to solve the Fair Recommendation Matrix (FRM) problems using a binary sensitive attribute vector and a stipulated fairness level to compute a new recommendation matrix, which must respect the metrics proposed in the study μ_1 e μ_2 proposed in the study.

H. Zhang et al. (2021) presented OmniFair, a system designed to enforce fairness constraints across various machine learning models. The main findings emphasized that OmniFair outperformed existing methods in balancing accuracy and fairness, achieving a smaller accuracy drop while maintaining fairness constraints. OmniFair covers a wider range of bias levels compared to other methods, providing flexibility in the accuracy-fairness trade-off. It also demonstrates superior performance in reducing False Discovery Rate (FDR) differences and supports customized fairness metrics. Furthermore, OmniFair is significantly faster than comparable in-processing methods and offers efficient hyper-parameter tuning. Empirical results across multiple datasets show that OmniFair consistently achieves high-quality results with minimal accuracy loss.

Tae & Whang (2021) presented a method to improve ML model accuracy and fairness by selectively acquiring data. Instead of uniformly collecting data, a proposed framework called Slice Tuner acquires varying amounts for different data slices, optimizing the balance of model performance across all slices. This approach addresses the inefficiencies and biases that arise from over-representing certain data segments. By maintaining learning curves and using convex optimization, Slice Tuner ensures fairer and more accurate outcomes. Implementing such a framework can significantly enhance fairness in ML models by equitably distributing data acquisition efforts.

Shimao et al. (2022) introduced the concept of Strategic Best-Response Fairness (SBR-fairness) to address the limitations of existing fair-ML algorithms that do not consider the behavioral responses of prediction subjects. The authors argued that without accounting for how individuals might strategically alter their behavior in response to ML predictions, fairness efforts may be undermined. They proposed an innovative algorithm that removes discrimination while considering these strategic responses, thus improving fairness in ML predictions by ensuring that both the prediction outcomes and the incentives for individuals to change their behavior are aligned with fairness goals.

Pagano et al. (2023) investigated improving fairness in ML models with a focus on gender. They emphasized the importance of context-sensitive fairness metrics, showing that metrics tailored to specific sensitive attributes are more effective across various domains such as computer vision, NLP, and recommendation systems. By applying the gender

attribute, the study identifies statistical parity, PPR disparity, and error disparity as consistent and reliable fairness indicators. Comprehensive evaluations using accuracy, precision, recall, F1-score, and fourteen fairness metrics reveal nuanced insights into bias and fairness. Case studies with diverse datasets and models like VGG19, BERT, and Wide Deep architectures highlight the effectiveness of these metrics. The findings stressed that sensitive attributes significantly influence metric selection, advocating against a one-size-fits-all approach.

Wan et al. (2023) presented several strategies to improve fairness in ML models. They emphasized the need for fairness constraints to be integrated into the training process to mitigate biases. The authors proposed a novel regularization method that adjusts model predictions based on context-specific fairness criteria. Experimental results demonstrated that this approach significantly reduces disparate impact and other fairness-related discrepancies across various datasets. By incorporating fairness regularization directly into the optimization objective, the technique ensures that the trade-off between accuracy and fairness is effectively managed. The study also highlighted the importance of selecting appropriate fairness metrics based on the context of the application.

4.2 Summary of results

The selected papers in SLR are shown in Table 2, which presents their main characteristics, namely the year of publication, authors' names, stage of intervention, fairness and performance metrics, techniques, and data set used in the experiments for practical proof of the proposals.

Table 2 SLR Selected papers

Reference	Stage(s)	Fairness Metric(s)	Performance Metric(s)	Technique(s)	Dataset(s)
Zafar et al. (2017)	In	Decision Boundary Covariance	Accuracy	LogR and SVM customized by the fairness metric.	UCI Adult; UCI Bank marketing
Calmon et al. (2017)	Pre	Sample Distortion	ROC	Probabilistic data transformation onto LogR and RF	ProPublica COMPAS; UCI Adult
Rudin & Ustunb (2018)	In	Statistical Parity	TPR; FNR; ROC	SLIM; RiskSLIM	Obstructive Sleep Apnea; Seizure Prediction; Recidivism
B. H. Zhang et al. (2018)	In	Statistical Parity; EO; Equal Opportunity;	FPR; FNR	Adversarial Learning in LogR	UCI Adult
Krasanakis et al. (2018)	Pre	p% rule; Difference between sensitive and non-sensitive FPR and FNR	Accuracy	ASR + CULEP	UCI Adult; UCI Bank marketing; ProPublica COMPAS
Hu et al. (2019)	Pre	Statistical Parity	Error Classifier	DFRR; DFKRR; DFGR; DFPCA.	ProPublica COMPAS; UCI default of credit card; Community Crime
Bellamy et al. (2019)	Pre; In and Post	Statistical Parity; EO; Equal Opportunity; Consistency; Sample distortion; Theil Index.	Accuracy	LogR and RF Re-weighing; NN Adversarial debiasing; LogR Prejudice remover; LogR and RF Optimized pre-processing; LogR and RF EO postprocessing; LogR and RF Disparate Impact Remover; LogR and RF Calibrated EO postprocessing; LogR and RF Learning Fair Representations; LogR and RF Reject option classification	UCI Adult; UCI Stat log German Credit Data; ProPublica COMPAS
Sahu & Singh (2019)	In	Diversity and Quality in Recommendations	For bias score; and against bias score	Collaborative filtering.	Movie Lens
Celis et al. (2019)	In	Statistical Parity; TPR; Accuracy Rate; FDR;	Accuracy	Algo 1-SR; Algo 1-FDR; Algo 1-SR+FDRCOV; FPR-COV; FNR-COV; SHIFT; and REDUCTION;	UCI Adult; German Credit Data; ProPublica COMPAS
Dash et al. (2019)	Pre; In; and Post	Statistical Parity	ROUGE; Recall and F1 score	ClasswiseSumm; FairSumm; RefaSumm.	Claritin tweets; US Election 2016 tweets; Me Too tweets
Edizel et al. (2020)	Post	μ_1 e μ_2	Precision and Recall	FaiRecSys	Movie Lens; Reddit
Obaidat et al. (2021)	Pre	-	Accuracy	Minimize AI bias applying the Random Sampling Technique (MAIRST).	Fashion MNIST

Valdivia et al. (2021)	In	FPR parity	Geometric mean (G-mean)	NSGA-II and Decision Trees	UCI Adult; UCI Stat log; German Credit Data; ProPublica COMPAS, ProPublica Violent; Ricci
Ahmed et al. (2021)	Post	Statistical Parity; EO; Equal Opportunity; Theil Index	Accuracy	LogR and RF Equal odds postprocessing; LogR and RF Calibrated equal odds postprocessing.	US Employment Demographics
H. Zhang et al. 2021)	In	Statistical Parity; FPR Parity; FNR Parity; False Omission Rate Parity; FDR Parity; and Misclassification Rate Parity	Accuracy	LogR, RF, XGBoost (XgB), NN and CMA-ES	UCI Adult; ProPublica COMPAS; Law School Admission Council (LSAC); Bank telemarketing data (Bank)
Tae & Whang (2021)	Pre	Average Equalized Error Rates (Avg. EER); Maximum Equalized Error Rates (Max. EER)	Accuracy	CNN	Fashion-MNIST; Mixed-MNIST; UTKFace; UCI Adult.
Shimao et al. (2022)	Pre	p% rule	Accuracy	DIR (DP-Based); PRR (DP-Based); CEO (EO-Based); Strategic Best-Response Fair Discriminant Removed (SBF-DR)	German Credit Data;
Pagano et al. (2023)	Pre	FNR Disparity; PPR Disparity; FPR Disparity; Error disparity; TPR disparity; Predictive Darity; Statistical Parity; FOR Difference (FORD); FNR Difference; Average Odds Difference (AOD); AUPRC; EO; Predictive Equality; Equal of Opportunity	Accuracy; Precision; Recall; F1 score; AUPRC.	Visual Geometry Group (VGG19); Bidirectional Encoder Representations from Transformers (BERT); and Wide and Deep models	Face Recognition–FairFace Challenge (FIFC); Jigsaw Unintended Bias in Toxicity Classification (JUBTC); Movie Lens.
Wan et al. (2023)	In	Demographic Parity; Equal Opportunity; EO; Overall Accuracy Equality; Treatment Equality; Equalizing Disincentives; Rawlsian Max-Min fairness principle; Fair clustering; Fairness through awareness; Counterfactual fairness; Differential fairness	Accuracy	Adversarial Learning	UCI Adult;

5. Discussion

This section discusses the different aspects surrounding fairness improvement in ML, such as metrics, algorithms, process steps, and datasets most used in the papers identified in the SLR. Rudin & Ustunb (2018) sought to develop fairness in ML systems related to health and justice while Lin et al. (2021) presented tools that can be used specifically to address fairness during recruitment and selection process. Some authors dealt with more unique systems, such as Edizel et al., (2020) and Sahu & Singh (2019) who sought to improve fairness in content recommendation after processing the results of the initial recommendation, and Dash et al. (2019) who offered a solution for summarizing texts that equally represents all groups presented in the data. Some authors sought solutions to specifically improve the use and availability of demographic data, offering distributed learning solutions (Hu et al., 2019). Obaidat et al. (2021) sought to achieve an even more particular objective, the minimization of biases and prejudices arising specifically from intrusions into ML systems. Finally, note to Tae & Whang (2021) and Pagano et al. (2023) who tried to improve fairness in images datasets.

5.1 Fairness metrics

Regarding the metrics, or notions of fairness used in each proposal, we can see a clear preference for the use of statistical parity, as shown in Figure 2. Ten authors sought to achieve this metric, also called disparate impact, equal acceptance rate, Demographic parity or just group fairness, and two authors sought to meet the *p% rule (*Table 2*)*, which is considered a derivation from statistical parity. This measure seeks to equalize the outcomes between protected and unprotected

groups, as is independent of real labels, which is an advantage when such labels are not available, is in the case of sensitive attributes. For Friedler et al. (2016), fairness metrics tend to correlate, and in general, the combination of class sensitive errors rates, plus statistical parity measure, can guarantee a good job when seeking to impose fairness in AI. However, although this measure seems to be desirable and simple, it was criticized in famous paper in the field "Fairness through awareness" (Dwork et al., 2012), that argue the theory of self-realization in which in a recruitment case for example unqualified members of a protected group are chosen in order to justify future discrimination against that same group. They also argue that statistical parity is not guaranteed to subgroups of the protected group and that it would have reduced utility in certain situations. There were two second most used metrics, presented in five studies, the metric Equalized Odds, which stipulates that both group have the same TPR and PR, and Equal Opportunities where the groups should only have the same TPR, both proposed by Hardt et al. (2016). According to Pessach & Shmueli (2020) the effectiveness of the Equalized Odds metric was proven in the Pro Publica data set, famous COMPAS case, where it was observed that although the accuracy was similar between the two groups (African Americans and Caucasians), the FPR among African Americans was twice as high as the FPR among Caucasians, proving that the system was twice as wrong in predicting recidivism for blacks than for white Americans. Speicher et al. (2018) also mention that any notion of group fairness, which encompasses both, statistical and equalized odds, does not consider the size of different groups, so it should not be considered ideal.

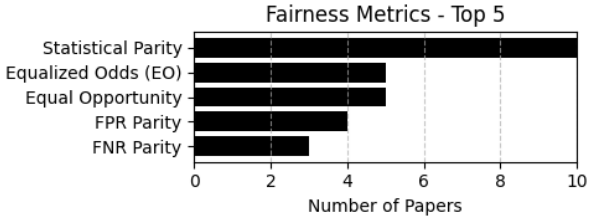

Figure 2. Fairness Metrics – Top 5.

Regarding performance metrics, a certain preference for accuracy was observed, as 11 papers chose to measure the model's performance in this way, as shown in Figure 3. In addition to this metric, other confusion matrix metrics such as F1 Score (2), FNR (2), FPR (1), Precision (2), Recall (3) and TPR (1) were used by the authors. Three authors also used fairness and performance metrics inherent to the purpose of each study. In Dash et al. (2019), fairness was imposed by statistical parity, while performance was assessed using ROUGE measurement, which evaluates the quality of summaries produced by the model. In Sahu & Singh (2019), performance was measured by the quality of recommendations, considering user evaluations, while fairness was understood as the diversity introduced in the final recommendations, which aimed to alleviate personal bias when recommending movies. Hu et al. (2019) and Valdivia et al. (2021) focused on measuring the systems by the errors.

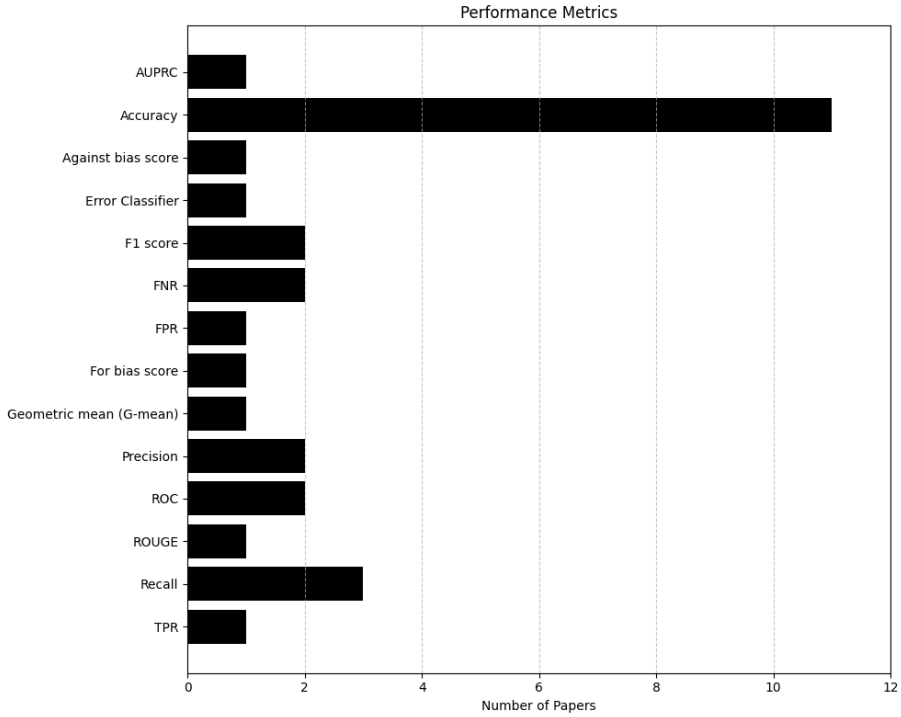

Figure 3. Performance Metrics.

5.2 Fairness improving techniques and process stages

Some papers presented multiple techniques across various stages of intervention to improve fairness, including pre-processing, in-processing, and post-processing stages. Each study is detailed according to the proposed algorithms and their intervention stage, allowing for better visualization of the solutions presented and the specific processing stages in which they were employed.

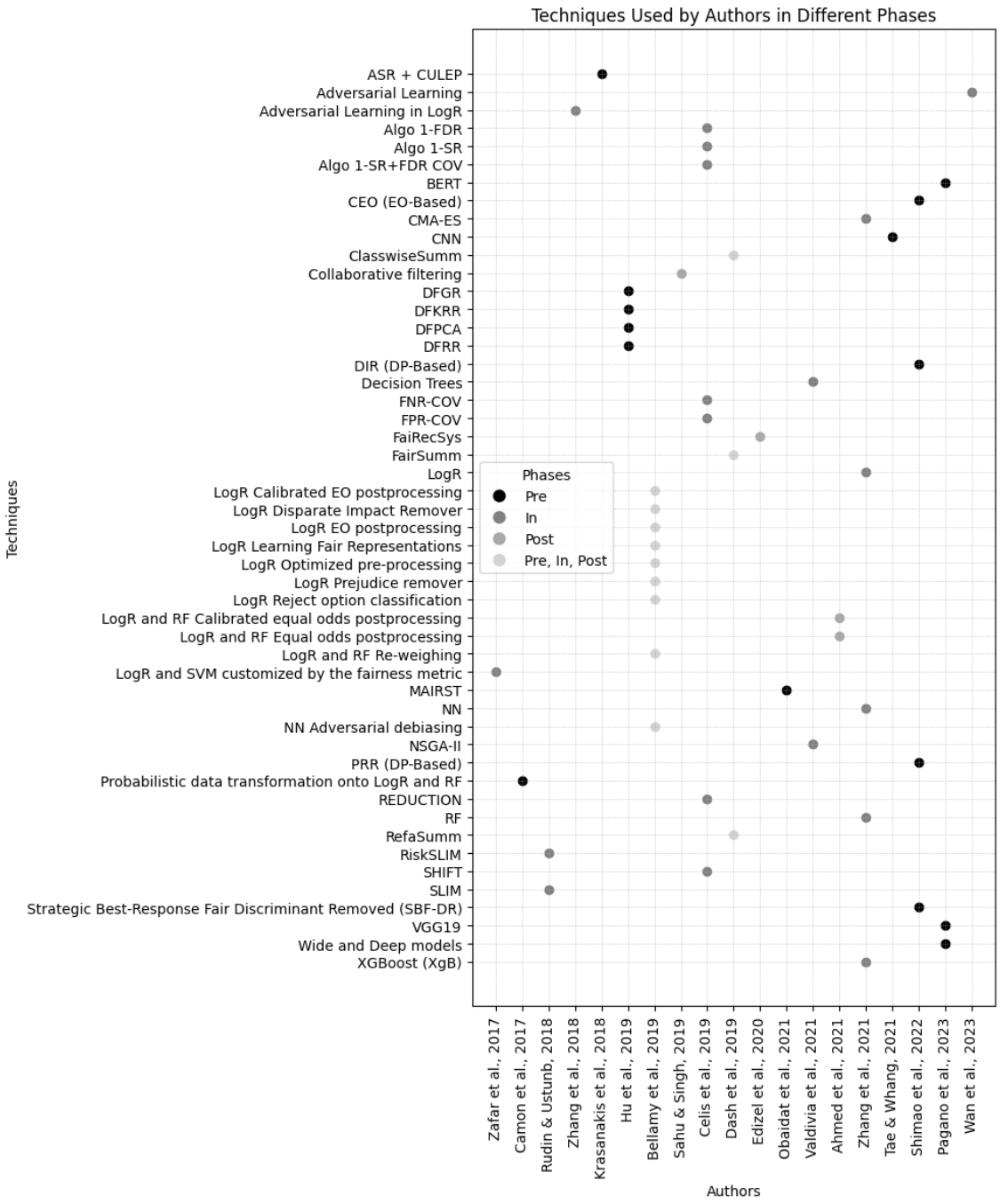

Figure 4. Technique and stage of processing.

Through the 19 selected papers, we identified 50 different techniques for improving fairness in AI, among which 11 of them work by adapting the algorithms and imposes fairness restrictions on the objective functions during the process stage (in-processing), 6 of them reclassify the outcomes so that the fairness metrics are met in post-processing, and 10 of them modify the dataset used to train the model during the pre-processing stage. In this sense, Pessach & Shmueli (2020) considered that pre and postprocessing techniques can be applied to any type of algorithm, they can harm the transparency and explanation of the results. This processing technique allows explicit control over the trade-off between accuracy and fairness. While this approach offers flexibility and can be tailored to specific algorithms, it also comes with added complexity in development. Furthermore, explaining the model's fairness-accuracy trade-off and how it searches for fairness during processing can be challenging, potentially hindering interpretability of the results. During the post processing specifically, two individuals who are similar in all characteristics except the group to which they belong, e.g., race, can be treated differently, requiring the decision maker to have sensitive information from each individual, which is prohibited by law (Pessach & Shmueli, 2020).

It is possible to verify, through the analysis of Figure 4, that the LogR and RF algorithms were the most used algorithms, 11 and 8 times respectively, and thus considered the most appropriate in the proposals of the different studies. Models based on LogR have the advantage of being interpretable and explainable, with less probability of overfitting and applicable for multiclass predictions, while models based on RF have high accuracy when compared to others algorithms but can be more difficult to interpret and require more complexity in training. Both of them are supervised learning algorithms (Camp, 2022). Only three papers cite the use of unsupervised algorithms: NN (Bellamy et al., 2019; H. Zhang et al., 2021), CNN (Obaidat et al., 2021). Lin et al. (2021) enriched our research with eleven commercial tools, already in use in the market, to improve fairness during recruitment and selection process, but these tools were not classified since they did not disclose which techniques were used in their commercial solutions.

5.3 Datasets

As the choice of dataset is essential for learning the model, another chart was developed with the most used datasets in eligible papers, shown in Figure 5.

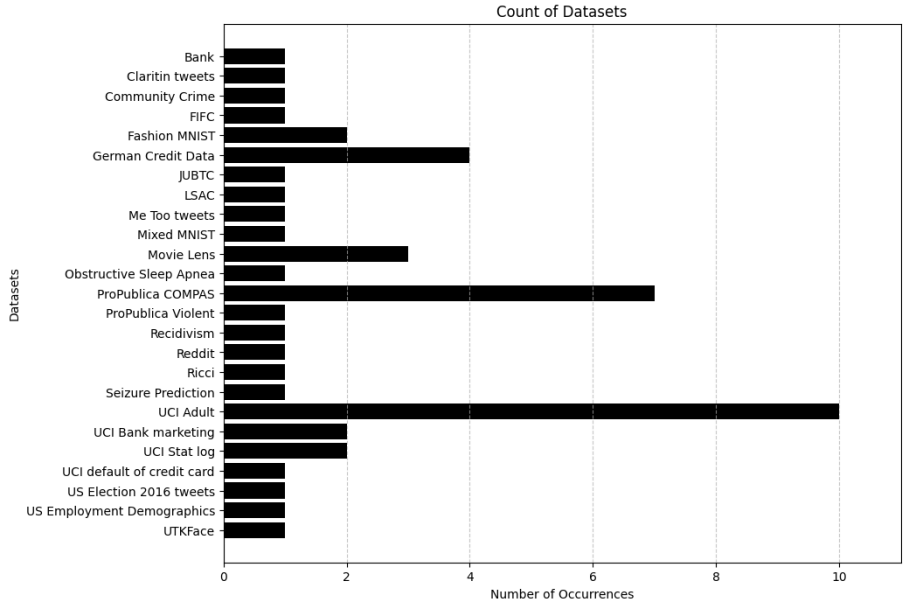

Figure 5. Used datasets.

The UCI Adult and the Pro Publica COMPAS datasets were used ten and seven times, respectively, among the papers that comprise the SLR, being the most popular and demonstrating a reference standard in the research of the subject. The UCI Adult data set (Center for Machine Learning and Intelligent Systems, 2022) comprises information taken from 1994 US census. The dataset is made up of fourteen attributes, including age gender, occupation, education, race, and income and is mainly used in fairness research to compare the inference of race and gender in income (Mehrabi et al., 2021). The ProPublica COMPAS dataset (ProPublica, 2022), on the other hand, brings the criminal history of defendants from Broward County in the state of Florida, United States, during 2013 and 2014, and it contains attributes such as prison time, age, address, race, gender, income, etc. And the risk score of the COMPAS system, used in the study of fairness mainly to understand the relationship of race in the risk attribute by the system. Obaidat et al. (2021) were the only ones to make use of a dataset with images, the Fashion MNIST, a dataset that contains 70 thousand images of garments, such as t-shirts, pants and boots. The authors sought to minimize the effects of data tampering on hacked systems and trained a CNN with the MAIRST method to recognized and classify the pieces.

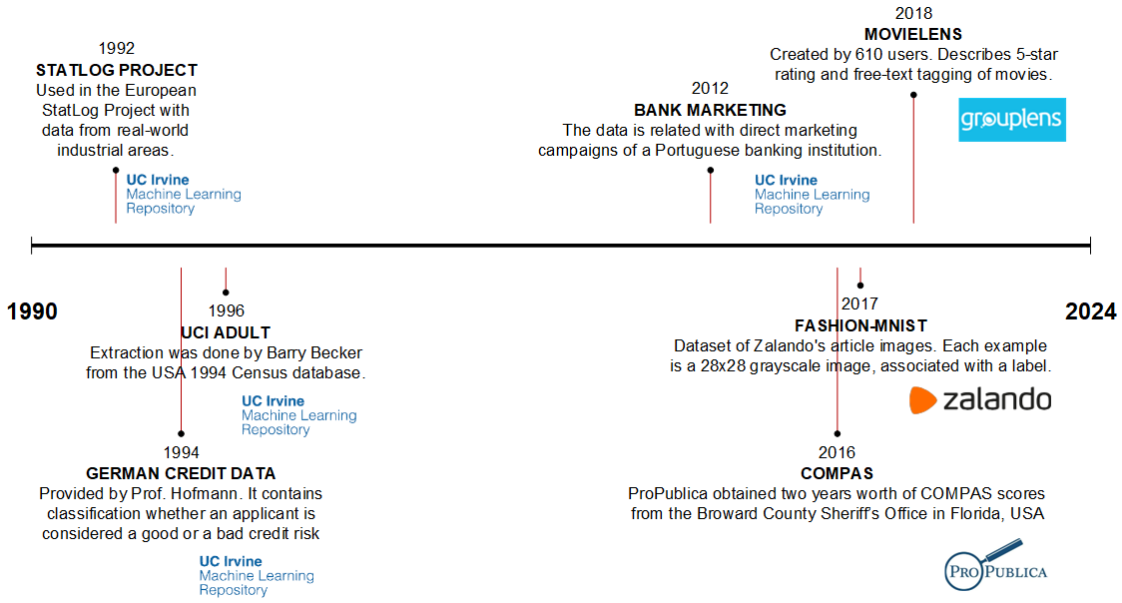

Figure 6. Chronology of the appearance of key datasets used to study fairness in artificial intelligence.

Figure 6 presents a chronology of the emergence of most used datasets in the twenty eligible papers on the SLR. The criteria for selecting a dataset to be represented in this chronology is that it has more than one use in eligible articles. The Adult UTI dataset is one of the longest available in this group, which may justify the fact that it is the most used. The COMPAS dataset was made available 20 years after the UCI Adult dataset and is already the second most used dataset, which may indicate that this dataset is very promising in terms of its possible future use.

6. Conclusion

In this research we discussed how prejudices are introduced in AI, presented several definitions of fairness metrics and techniques to improve fairness in AI. Our main objective was to present solutions to the fairness problem, so through a SLR we mapped 19 recent papers which suggested technical approaches of how to improve fairness in AI in different types of AI systems, such as classification, recommendation, and summarization, besides data tampering, data protection and data usage monitoring and compliance.

After describing and summarizing the papers, 50 techniques were identified in different stages of development, most of them during the preprocessing, where statistical parity was the most used fairness metric, as well as the algorithms LogR and RF. The UCI Adult from 28 years ago, and the ProPublica COMPAS from 8 years ago were also widely used when reaching the theme, and as we know the great importance of training data in the model. The COMPAS dataset appears to be very promising in terms of its possible future use in further studies. We also suggest that further experiments use different datasets for further explorations and research advancement and updating the fairness field in AI. Widespread adoption of these solutions hinges on a deep understanding of the system's context and the impacted groups.

Therefore, we aim to make this research accessible, that is, disseminate the subject and popularize the solutions. This SLR can serve as a foundation for establishing common ground in algorithmic fairness research. By promoting standardized metrics, datasets, and techniques, we can foster a unified perspective on key concepts and methods. This will provide a strong base for future research, guiding us towards established knowledge and areas ripe for further exploration. During the research, we recognize that the selected papers proved to be divergent in relation to the type of ML systems and the purpose each author sought to solve, and it made difficult to classify the studies to the parameters that we pre-defined as metrics, stages, and algorithms in which we seek to recognize references.

The work presented here is extremely important for the promotion of the need to improve fairness in AI. As a future work, we propose the organization of similar studies over time with a view to creating an observatory of fairness in AI and the construction of a practical guide for researchers with techniques to improve fairness in AI processes.

References

Ahmed, S., Athyaab, S. A., & Muqtadeer, S. A. (2021). Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 557–

563. https://doi.org/10.1109/ICICT50816.2021.9358507
Barocas, S., Hardt, M., & Narayanan, A. (2019). *FAIRNESS AND MACHINE LEARNING Limitations and Opportunities*.
Barocas, Solon, & Selbst, A. D. (2016). Big Data ' S Disparate Impact. *California Law Review*, *104*(671), 671–732. https://doi.org/http://dx.doi.org/10.15779/Z38BG31
Bellamy, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4–5). https://doi.org/10.1147/JRD.2019.2942287
Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x
Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 3993–4002.
Camp, D. (2022). *Machine Learning Cheat Sheet*.
Cardenas, S., & Vallejo-Cardenas, S. F. (2019). Continuing the Conversation on How Structural Racial and Ethnic Inequalities Affect AI Biases. In *2019 IEEE International Symposium on Technology and Society (ISTAS), Technology and Society (ISTAS), 2019 IEEE International Symposium on* (pp. 1–7). IEEE. https://doi.org/10.1109/ISTAS48451.2019.8937853
Caton, S., & Haas, C. (2020). *Fairness in Machine Learning: A Survey*. 1–33.
Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with Fairness Constraints. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 319–328. https://doi.org/10.1145/3287560.3287586
Center for Machine Learning and Intelligent Systems. (2022). *UCI Machine Learnig Repository*.
Chouldechova, A., & Roth, A. (2018). *The Frontiers of Fairness in Machine Learning*. 1–13.
Cortez, A., Trigo, A., & Loureiro, N. (2022). Football Match Line-Up Prediction Based on Physiological Variables: A Machine Learning Approach †. *Computers*, *11*(40), 1–14. https://doi.org/10.3390/computers11030040
Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S., & Chakraborty, A. (2019). Summarizing User-generated Textual Content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359274
DOMO. (2022). *Data Never Sleeps 9.0*. https://www.domo.com/learn/infographic/data-never-sleeps-9
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255
Edizel, B. (1), Bonchi, F. (2), Panisson, A. (2), Hajian, S. (3), & Tassa, T. (4). (2020). FaiRecSys: mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics*, *9*(2), 197–213. https://doi.org/10.1007/s41060-019-00181-5
Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 259–268. https://doi.org/10.1145/2783258.2783311
Ferrer, X., Nuenen, T. van, Such, J. M., Cote, M., & Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80. https://doi.org/10.1109/MTS.2021.3056293
Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian, S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589
Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (im)possibility of fairness*. *im*, 1–16.
Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & SOCIETY*, *38*(2), 549–563. https://doi.org/10.1007/s00146-022-01455-6
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *Nips*, 3323–3331.
Hu, H., Liu, Y., Wang, Z., & Lan, C. (2019). A Distributed Fair Machine Learning Framework with Private Demographic Data Protection. In *2019 IEEE International Conference on Data Mining (ICDM), Data Mining (ICDM), 2019 IEEE International Conference on* (pp. 1102–1107). IEEE. https://doi.org/10.1109/ICDM.2019.00131
Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. In *Knowledge and Information Systems* (Vol. 33, Issue 1). https://doi.org/10.1007/s10115-011-0463-8
Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, *2*, 853–862. https://doi.org/10.1145/3178876.3186133
Lin, Y.-T., Hung, T.-W., & Huang, L. T.-L. (2021). Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias. *Philosophy & Technology*, *34*(1), 65–90.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine

Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607
Mokhtari, K. A., Benbernou, S., Ouziri, M., Lahmar, H., & Younas, M. (2020). A monitoring framework for transparency and fairness in big data plataform. *WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY*. https://doi.org/10.1002/cpe.6069
Ntoutsi, E. (1), Gadiraju, U. (1), Iosifidis, V. (1), Nejdl, W. (1), Staab 12,13), S. (1, Fafalios, P. (2), Vidal, M.-E. (3), Ruggieri, S. (4), Turini, F. (4), Papadopoulos, S. (5), Krasanakis, E. (5), Kompatsiaris, I. (5), Kinder-Kurlanda, K. (6), Wagner, C. (6), Karimi, F. (6), Fernandez, M. (7), Alani, H. (7), Berendt 9), B. (8, Kruegel, T. (10), … Tiropanis, T. (12). (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356
O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st ed.). Crown Publishing Group. https://www.amazon.co.uk/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815
Obaidat, M., Singh, N., & Vergara, G. (2021). Artificial Intelligence Bias Minimization Via Random Sampling Technique of Adversary Data. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference, CCWC 2021*, 1226–1230. https://doi.org/10.1109/CCWC51732.2021.9375929
Onifade, A. B. (2023). Looking beyond the impressions of algorithms and fact-checking in fighting online misinformation: A literature review. *Education for Information*, *39*(1), 33–49.
Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O. R., Peixoto, R. M., Guimarães, G. A. de S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Context-Based Patterns in Machine Learning Bias and Fairness Metrics: A Sensitive Attributes-Based Approach. *Big Data and Cognitive Computing*, *7*(1), 27. https://doi.org/10.3390/bdcc7010027
Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/bmj.n160
Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. *ArXiv:2001.09784*. https://doi.org/10.1257/pandp.20181018
Pimenta, D., Teles, M., Belfo, F., & Trigo, A. (2023). Medication recommendation in cancer treatment based on cell line similarity. *Procedia Computer Science*, *219*, 1493–1500. https://doi.org/10.1016/j.procs.2023.01.440
ProPublica. (2022). *ProPublica Data Store*. https://www.propublica.org/datastore/
Rudin, C., & Ustunb, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449–466. https://doi.org/10.1287/inte.2018.0957
Sahu, S., & Singh, S. K. (2019). Ethics in AI: Collaborative filtering based approach to alleviate strong user biases and prejudices. In *2019 Twelfth International Conference on Contemporary Computing (IC3), Contemporary Computing (IC3), 2019 Twelfth International Conference on* (pp. 1–6). IEEE. https://doi.org/10.1109/IC3.2019.8844875
Shimao, H., Khern-am-nuai, W., Kannan, K., & Cohen, M. C. (2022). Strategic Best Response Fairness in Fair Machine Learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 664–664. https://doi.org/10.1145/3514094.3534194
Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). *A Unified Approach to Quantifying Algorithmic Unfairness*. 2239–2248. https://doi.org/10.1145/3219819.3220046
Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM*, *64*(8), 44–49. https://doi.org/10.1145/3464903
Tae, K. H., & Whang, S. E. (2021). Slice Tuner. *Proceedings of the 2021 International Conference on Management of Data*, 1771–1783. https://doi.org/10.1145/3448016.3452792
Tutt, A. (2020). An FDA for algorithms. *Administrative Law Review*, *69*(1), 83–123.
Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, *36*(4), 1619–1643. https://doi.org/10.1002/int.22354
Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings - International Conference on Software Engineering*, 1–7. https://doi.org/10.1145/3194770.3194776
Von Hohendorff, R., & Kaini Lazzaretti, B. (2021). O uso da inteligênccia artificial na tomada de decisões judiciais: Uma Análise sob a Perspectiva da Crítica Hermenêutica do Direito. *RDUno: Revista Do Programa de Pós-Graduação Em Direito Da Unochapecó*, *3*(4), 15–32. https://doi.org/10.46699/rduno.v3i4.6072
Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, *17*(3), 1–27. https://doi.org/10.1145/3551390
Yan, R., Zhao, X., & Mazumdar, S. (2023). Chatbots in libraries: A systematic literature review. *Education for Information*, *39*(4), 431–449. https://doi.org/10.3233/EFI-230045
Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, *54*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, *28*(PART 2), 1362–1370.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Zhang, H., Chu, X., Asudeh, A., & Navathe, S. B. (2021). OmniFair: A Declarative System for Model-Agnostic Group Fairness in Machine Learning. *Proceedings of the 2021 International Conference on Management of Data*, 2076–2088. https://doi.org/10.1145/3448016.3452787