Software Unfairness Detection in Machine Learning-based Systems: A Systematic Mapping Study

Roa Alharbi * and Noureddine Abbadeni

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0038.v1

Keywords: machine learning-based systems; software fairness; bias detection; fairness testing; systematic mapping study

 doi:10.20944/preprints202604.0038.v1

Article

Software Unfairness Detection in Machine Learning-Based Systems: A Systematic Mapping Study

Roa Alharbi [1,2,*] and Noureddine Abbadeni [1]

[1] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[2] Next Gen Connectivity & Wireless Sensors Institute, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

* Correspondence: rhalharbi@kacst.edu.sa

Abstract

Machine learning-based systems are increasingly deployed in high-stakes domains such as healthcare, finance, law, and e-commerce, where their predictions directly influence critical decisions. Although these systems offer powerful data-driven support, they also introduce serious concerns related to fairness, bias, and discrimination. As a result, detecting and addressing unfairness in machine learning software has become a central research challenge. This study presents a systematic mapping of research on software unfairness detection in machine learning systems, with the aim of consolidating existing fairness definitions, identifying major problem types, examining testing approaches, reviewing commonly used datasets, and highlighting open research gaps. A structured search was conducted across five major digital libraries and additional sources, covering publications from 2010 to 2025. From 1,805 initially identified records, 67 primary studies met the inclusion and quality assessment criteria. The findings show that research activity has grown significantly since 2019, reaching a peak in 2022. Most studies were published at conferences, followed by journals and workshops. The literature addresses various themes, including analysis of existing fairness methods, bias mitigation strategies, testing techniques, and evaluation frameworks. Fairness testing was performed at unit, integration, and system levels, with integration testing being the most common. Frequently used datasets include COMPAS, Adult Census Income, and German Credit. Widely adopted tools such as IBM AI Fairness 360, Themis, and Aequitas were also identified. Overall, the mapping highlights progress made in fairness research while emphasizing the need for stronger integration of fairness into practical machine learning development.

Keywords: machine learning-based systems; software fairness; bias detection; fairness testing; systematic mapping study

1. Introduction

Machine learning-based systems (MLS) have become deeply embedded in contemporary decision-making across sensitive domains such as healthcare, government services, education, business, finance, and security. Their capacity to process large-scale data and extract actionable patterns has positioned MLS as critical decision-support instruments in high-stakes contexts. Because MLS outputs can significantly affect people and groups making those predictions and recommendations credible, transparency of process, transparency of outcome, fairness, and equitable treatment is not simply desirable for societal trust and responsible/ethical deployment. When MLS produce biased outcomes, they will socially in-justice that can multiply historical social inequities and perpetuate discrimination on hiring, lending, and medical priority where the procedural justice and substantive outcomes are relevant.

Machine learning itself can be broadly categorized into supervised, unsupervised, semi-supervised, and reinforcement learning methods, each with distinctive ways of learning from data

 doi:10.20944/preprints202604.0038.v1

and making predictions [1]. Furthermore, systematic approaches for conducting structured reviews in software engineering and MLS research have been established and guide studies like this one [2].

Software fairness has been articulated through multiple complementary definitions. One widely employed perspective states that "an algorithm is fair if it gives similar predictions to similar individuals. Any two individuals who are similar with respect to a similarity metric defined for a particular task should be classified in the same way." [3] A second, frequently cited view holds that fairness is achieved when protected attributes (e.g., race, gender, age) are excluded from influencing outcomes [4]. These definitions motivate a family of operational metrics that enable empirical assessment, including disparate impact, demographic parity, equalized odds, and fairness through awareness, each capturing a distinct notion of equitable treatment and outcomes [5]. No one metric can ever be adequate in every situation but, rather, metrics need to be considered and understood in relation to the application context and the kinds of harms deferred. The presence of multiple metrics illustrates both indicators of the conceptual complexity of fairness and the practically challenging nature of applying fair treatment in real systems [68].

Fairness challenges can manifest at all stages of the MLS lifecycle. In the preprocessing stage, fairness work targets the training data (e.g., addressing imbalance and representational issues) before model induction. In the in-processing stage, algorithm design and optimization incorporate fairness considerations directly into learning procedures or objective functions. In the post-processing stage, predictions are examined and, if needed, adjusted to avoid disparate impacts across groups [6]. Separating out these stages of the design, development, and evaluation of AI systems shows the importance of taking fairness into account across the entire process, not just as a one-off check but as ingrained practice from start to finish [76].

While skills related to fairness are very visible in AI work today, literature is still in its infancy. Previous surveys and summaries of how fairness was addressed have been helpful including what we have learned about metrics, mitigation methods, and algorithmic attitudes towards bias and discrimination. However, these works tend to focus on specific technical strata (e.g., metric definitions, bias mitigation strategies, or algorithm families) without providing a consolidated mapping that simultaneously: (a) synthesizes fairness definitions, (b) categorizes problem types tackled in the field (analysis, mitigation, testing, evaluation), (c) identifies approaches to fairness testing (including where they sit in the MLS lifecycle), (d) enumerates datasets/algorithms/models used to detect unfairness (and why they exhibit bias), and (e) surfaces research gaps and trends across time and publication venues. Consequently, it remains difficult to observe field-level structure, compare practices across subcommunities, and identify consistent trends that would support standardization and reproducibility.

This fragmentation is compounded by the dispersion of contributions across software engineering, machine learning, data science, and applied computing venues. Researchers in these communities often apply different fairness definitions and evaluation protocols, complicating cross-study synthesis. At the same time, the ecosystem's reliance on commonly used datasets and benchmarks, each with known limitations, further motivates careful, structured accounting of what has been studied, how it has been evaluated, and where persistent gaps remain. Taken together, these observations motivate an integrative perspective that moves beyond isolated algorithmic or metric-centric reviews toward a systematic mapping study (SMS) that organizes the field along multiple, interlocking dimensions (definitions, problem types, testing approaches, datasets/algorithms/models, and research gaps/trends) over a defined time window.

Accordingly, this study undertakes an SMS of software unfairness detection in MLS over the 2010–2025 period, using established SMS procedures to search, screen, and assess the literature before extracting and synthesizing data aligned with predefined research questions. Unlike narrower technical surveys, this mapping emphasizes breadth with structure: it documents where research is published (venues/years), what kinds of fairness problems are being addressed, how fairness is tested across the MLS pipeline, which datasets/algorithms/models are used to reveal or study unfairness (including reasons underpinning their biases), and what gaps and trends emerge from the collective

 doi:10.20944/preprints202604.0038.v1

evidence. By consolidating these elements into a single, organized account, the study provides a platform for understanding the field's current state and for charting principled directions for future work [37].

To situate this contribution within existing scholarship, it is useful to distinguish the present SMS from prior work. Earlier reviews offer valuable treatments of fairness metrics and mitigation strategies or discuss algorithmic bias from particular vantage points. However, they do not concurrently (a) map definitions, (b) classify problem types, (c) survey fairness testing approaches by lifecycle stage, (d) catalog datasets/algorithms/models used to detect unfairness with attention to bias sources, and (e) synthesize gaps and trends across venues and years within a single, unified framework. Nor do they provide a consolidated picture that speaks directly to software engineering concerns (e.g., testing levels and lifecycle integration) alongside machine-learning-centric perspectives. This study addresses that need through an SMS design explicitly structured to capture these dimensions and report their interrelationships.

Finally, although the Introduction deliberately avoids detailed tutorials on ML algorithms, data modalities, and deep learning architectures—topics that are covered comprehensively in the Background chapter—the present section maintains the essential foundations required to motivate a fairness-centered SMS. Specifically, it retains: (i) the core operational definitions of fairness and the standard metrics by which it is evaluated, [3–5] (ii) the three-stage view of where fairness work occurs in the MLS lifecycle, [6] and (iii) the overarching rationale for end-to-end fairness verification and validation in decision-critical contexts. Broader technical details (e.g., supervised vs. unsupervised learning, neural architecture, and dataset taxonomies) are reserved for the Background so the Introduction can maintain a focused funnel from societal context to fairness problem, to literature limitations, to the study's objectives and scope.

Main Objective

The goal of this study is to present the results of a systematic mapping study of software unfairness detection in MLS. The primary objective is to address predefined research questions with research-analyzed results and to present the frequencies of solutions, thereby identifying the kinds of research, their quantities, and their results in this field.

Specific Objectives

1. Identify the most valuable venues of papers in the field of unfairness detection in MLS.
2. Explore different software fairness definitions.
3. Recognize types of addressed problems (detection, analysis, or evaluation).
4. Find approaches for fairness testing (algorithms or tools) and explore fairness testing levels in MLS.
5. Provide researchers with datasets, algorithms, and models for detecting unfairness in MLS and explain the reasons behind their biases.
6. Investigate the gaps in software fairness in MLS research topics.

The rest of the report is organized as follows: Research methodology, threats to validity, background, results and discussion, related work, conclusion, and future work.

2. Research Methodology

2.1. Methods Overview

This section provides an overview of the research methodology. Following the guidelines, [75,82] our process was conducted as follows: the first step was to define the research questions and determine the scope (see Table 1). The second step involved searching designated digital libraries using a carefully constructed search string. All collected papers were then screened by applying inclusion and exclusion criteria. Afterward, a quality assessment was performed to categorize the collected studies. Next, relevant papers were grouped, and keywords were classified for each study.

 doi:10.20944/preprints202604.0038.v1

Finally, data extraction and a structured literature review were carried out to complete the systematic mapping study. Figure 1 illustrates the main steps of the SMS. In this research, we applied the PICOC (Population, Intervention, Comparison, Outcome, and Context) criteria to define the research questions [75]. These were formulated from five perspectives:

- Population: Machine learning-based systems, artificial intelligence, fairness techniques, methods, models, and bias.
- Intervention: Software engineering, software unfairness detection.
- Comparison: Research publication statistics, fairness approaches, fairness testing levels, evaluation metrics, solutions, biased algorithms, and datasets.
- Outcomes: Software fairness definitions and techniques used in MLS, fairness testing levels, biased algorithms and datasets, solutions for mitigating unfairness, and identification of research gaps.
- Context: Research content relevant to both academia and industry.

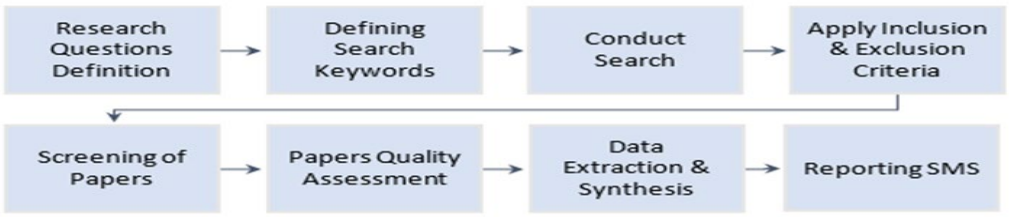


Figure 1. The methodological process followed in this study, illustrating the sequential steps for performing a systematic mapping study (SMS), from initial research question formulation to final reporting.

Table 1. Research Questions.

Id	Research questions	Rationale
RQ1	**What is the distribution of papers through venues and years?**	**Identify the kind of seminars for collected papers, published journals, conferences, timeline, and range of publishing dates.**
RQ2	**What are the different definitions of software fairness?**	**Explore software fairness definitions from primary studies.**
RQ3	**What types of problems are addressed?**	**Identify the type of addressed problem (detection, analysis, or evaluation).**
RQ4	**What different approaches of fairness testing are presented?**	**Describe different approaches used in fairness testing solutions, such as algorithms or tools. And explore fairness testing levels in MLS.**
RQ5	**Which datasets are used to detect the unfairness of MLS?**	**Identify some datasets/algorithms/models that are used to detect unfairness in MLS and explain the reasons behind their biases.**
RQ6	**What are the research gaps and trends discovered in the reviewed studies?**	**Explore the gaps in software fairness in MLS research topics.**

2.2. Search Strategy

To collect primary studies, we searched across several digital libraries, categorized as main and supplementary sources (see Table 2):

 doi:10.20944/preprints202604.0038.v1

Table 2. Digital Libraries.

Main Digital Libraries	Supplementary Databases/Websites
IEEE Xplore	
Springer Link	**Google Scholar**
ACM Digital Library	**Research Gate**
Wiley Online Library	
ArXiv	

Because the focus of this study is fairness in machine learning, the search keys were divided into two categories: one for fairness-related terms and their synonyms, and another for machine learning terms and alternatives. The search process began with the keywords ("fairness" and "machine learning"), which were then expanded with related terms. The search string was refined iteratively by both authors, using feedback from trial searches to improve the coverage of relevant studies [8].

2.3. Narrative for Study Screening and Selection

The initial search across all selected databases retrieved a total of 1,805 records, representing all studies containing at least one of the identified search terms related to fairness and machine learning.

In the first screening phase, duplicate entries across the digital libraries were removed. The inclusion criteria were then applied: studies had to be:

(a) published in academic journals, conferences, or magazines,
(b) written in English,
(c) full-text accessible, and
(d) published between 2010 and 2025.

After applying these filters, the pool was reduced from 1,805 to 1,370 studies (see Figure 2).

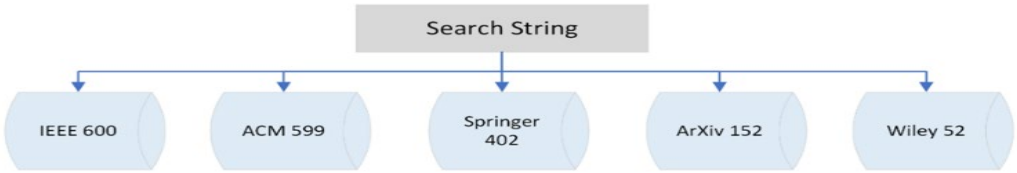
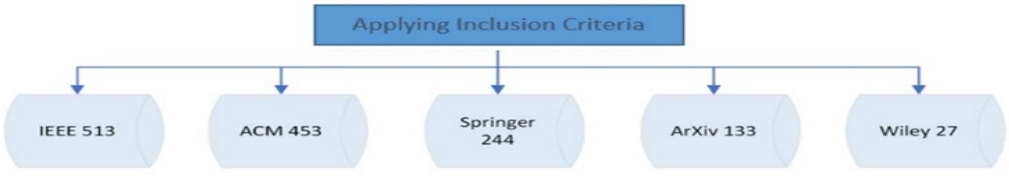


Figure 2. Overview of the initial search results generated by the search string, showing the number of publications retrieved from each source: IEEE (600), ACM (599), Springer (402), ArXiv (152), and Wiley (52).

The second phase involved title and abstract screening. At this stage, the exclusion criteria were applied:

(a) duplicate records not previously detected,
(b) studies on irrelevant topics, and
(c) non-published works.

Manual checks were performed to ensure that both conference papers and their extended journal versions were appropriately retained. After this step, the number of studies was further reduced (see Figure 3).



 doi:10.20944/preprints202604.0038.v1

Figure 3. Number of studies retained from each digital library after applying the inclusion criteria, showing IEEE (513), ACM (453), Springer (244), ArXiv (133), and Wiley (27).

The third phase involved full-text screening of the remaining articles. Papers were excluded if they did not directly address software fairness in MLS, if they lacked substantive discussion of unfairness detection, or if they did not provide sufficient methodological or empirical content to answer the predefined research questions.

Finally, a quality assessment (QA) was applied using the predefined QA checklist. Each paper was evaluated against three questions: (a) Does the study specify the goal of the research? (b) Does the study propose a solution for unfairness detection in MLS? (c) Does the study evaluate the proposed idea? Scores were assigned as Yes = 1, Partly = 0.5, and No = 0, with a cumulative threshold of 1.5. Studies scoring at or below this threshold were excluded. After all these steps, 67 primary studies remained and were included in the final synthesis of this systematic mapping study (see Figure 4).

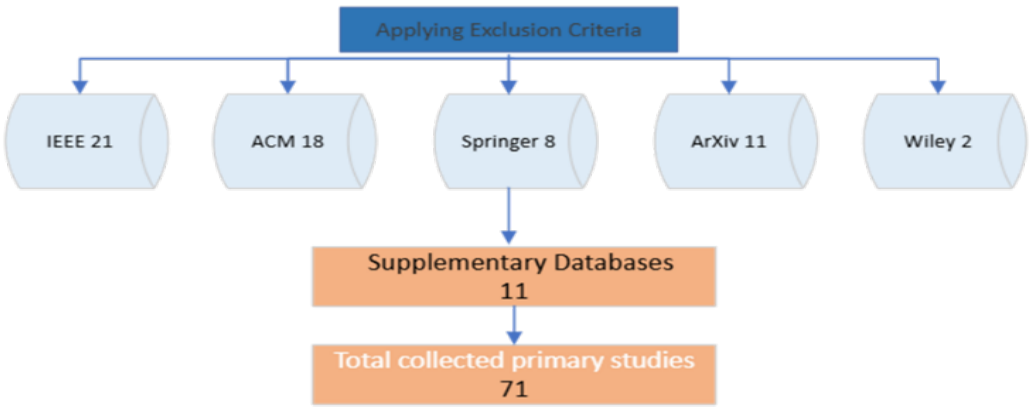


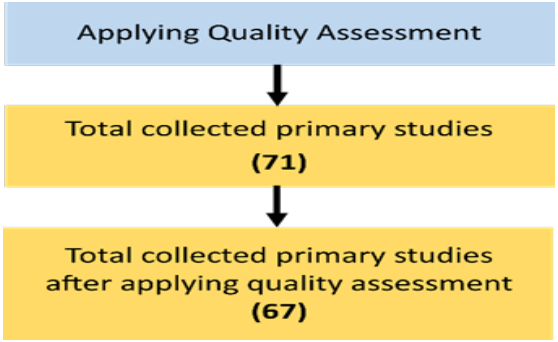
Figure 4. Number of studies remaining after applying the exclusion criteria across digital libraries (IEEE: 21, ACM: 18, Springer: 8, ArXiv: 11, Wiley: 2), followed by the addition of 11 studies from supplementary databases, yielding a total of 71 collected primary studies.

2.3. Quality Assessment and Data Analysis

To evaluate the literature, we applied a quality assessment procedure using predefined quality assessment (QA) questions. Each study was reviewed manually, and scores were assigned as follows: Yes = 1, Partly = 0.5, No = 0. The three QA criteria were:

- Does the study specify the goal of the research?
- Does the study propose a solution for unfairness detection in MLS?
- Does the study evaluate the proposed idea?

The cumulative score for each study was calculated. A threshold score of 1.5 was set; studies scoring at or below this threshold were excluded. After applying this assessment, the number of primary studies was reduced to 67 (see Figure 5, Appendix 1).



 doi:10.20944/preprints202604.0038.v1

Figure 5. Total number of collected primary studies before and after applying the quality assessment, showing a reduction from 71 initial studies to 67 studies that met the required quality criteria.

2.4. Data Extraction

In the data extraction stage, relevant information was collected from each selected study to answer the research questions. An Excel file was used to organize extracted information and generate charts. Keywords were defined for each research question (Table 3), and synonyms were used when necessary to avoid missing information. Extracted findings were synthesized and presented in the Results and Discussion section.

Table 3. Search Keywords.

Search keywords	RQs	Possible Values
Journals, Conferences, Year	RQ1	Journals, conferences, university names and types, published year
Definition	RQ2	Fairness definition
Problem Types	RQ3	Analyzing, reviewing, detecting, evaluating, or testing solutions.
Methodology	RQ4	Approaches, tools, algorithms, unit testing, input testing, or system testing.
Datasets/Algorithms/Models	RQ5	Bias/unfairness, datasets, models, algorithms
Research gaps	RQ6	Trend, future work, research gap

2.5. Threats to Validity

One risk for a threat to validity is that researchers working in this area of fairness might use different terminology, and possibly, we would miss important studies. To implement this, the search string was constructed with multiple possible alternative keywords, depending upon the research question and tailored to the search functionality of each digital library. When the full string was unable to retrieve results, I separated the keywords and tested them individually. This strategy minimized the risk of being incomplete and increased consistency in my dataset when collected.

The following section provides the background necessary to understand the scope and significance of the study. It outlines the theoretical foundations, previous research, and key concepts relevant to the topic.

3. Background

3.1. Machine Learning-Based Systems and Deep Learning

A machine learning (ML) model is a trained instance of a specific ML algorithm ML algorithms typically fall into four main types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

Supervised Learning: These algorithms create a mathematical model from a dataset containing inputs and corresponding outputs. The dataset is known as training data, and the labels represent the desired outcomes. Supervised learning is commonly applied in classification and regression tasks [68].

Unsupervised Learning: This type of algorithm analyzes a dataset with only inputs to find hidden structures or patterns. They cluster or group data on feature similarities. They are different from supervised methods in that there are no labeled data. Examples are clustering and association. A type of unsupervised learning is self-supervised learning as the model creates the labels from the

 doi:10.20944/preprints202604.0038.v1

input data to serve as the labels for learning. For example, self-supervised learning is common in natural language processing and computer vision [69].

Semi-Supervised Learning: This method of learning includes a small amount of labeled data and a large amount of unlabeled data during the training process, which brings considerable improvement in accuracy [68].

Reinforcement Learning: This strategy uses the interaction between the model and environment where the model takes action to maximize cumulative feedback or rewards. Reinforcement learning is popular in robotics, gaming, and navigation [69].

Neural Networks (NNs): An NN is inspired by neurons in the human brain, where one neuron passes on information to another neuron. Each neuron has inputs to process before passing information to the next neuron. The same for an NN, where it receives input through the input layer, processes under the one or more hidden layers, and receives output from the output layer [11].

Deep Neural Networks (DNNs): A DNN has multiple layers of connected nodes that build upon the previous layer and the neurons to create a more refined prediction. The process of passing information forward through the layers is called forward propagation. The process of moving through the network in reverse is called backpropagation, which gives weights to the neurons for the network, so estimates are minimized to the errors. The DNN uses both tasks to learn to estimate accurately and correct the errors. More simply, deep learning consists of: "many hidden layers of neural networks, performing complex manipulation over excessively large amounts of structured and unstructured data…any form of data such as images, text, sound, and time series, using sufficient amounts of training data to improve estimates" [12].

In traditional ML algorithms the features used for making a prediction are selected by the researcher or user. For DNN algorithms, the features used as input for prediction are learned automatically by the DNN algorithm from the datasets during the training process. With DNN, the inputs dataset input into the DNN algorithm can be a large amount of data and as it is processing the data, it can learn the features for heterogeneous or homogeneous data that are relevant to the data provided, and the DNN estimates weights and biases to minimize error between the DNN prediction and the expected outcome.

3.2. Deep Learning Tasks and Architectures

Research on deep learning has generated a number of prominent architectures that have been successfully deployed in a number of applications. Convolutional Neural Networks (CNNs) are the dominant model in computer vision, acquiring local and hierarchical features from image data through convolutional and pooling layers. CNNs have achieved state-of-the-art performance in a number of tasks including image classification, object detection, and medical image analysis, and they are the most deployed models when addressing fairness related to demographic bias in vision tasks [20].

Recurrent Neural Networks (RNNs), including their most well-known variant the Long Short-Term Memory (LSTM) models, purposefully study sequential or time-series data. RNNs and LSTMs are especially well suited for modeling sequential dependencies as they generate contextual representations of the data and thus become important for advancing research in natural language processing (NLP), speech detection, and other fields having time-dependent data. Fairness concerns with RNNs and LSTMs arise strongly when sequential dependencies encode historical or social inequities, as seen in predictive text and risk checklists. [12,13]

More recently, transformer architecture has become the leading standard in NLP and related fields. Transformers leverage self-attention mechanisms and thus are better able to represent long-range dependencies compared to standard RNNs, which is appropriate for building large-scale models such as BERT and GPT. However, while these models represent a considerable advancement in model performance, and as scale and data training is elevated there is an increase in concerns about hidden biases that manifest when leveraging large text corpora. Given that CNNs, RNNs/LSTMs, and transformers are foundational components for many modern MLS applications they are also

 doi:10.20944/preprints202604.0038.v1

often the concentrical focuses of fairness detection and mitigation research, which situates the technical advancement in the design of the architecture directly association with ethical considerations in machine learning [8,20].

3.3. ML and DL Tasks in Data Mining

Predictive Tasks: Classification, Regression, and Sequence Prediction.

In classification problems, the system assigns input data a class based on algorithms, for example, Decision Trees, k-Nearest Neighbors, or Support Vector Machines (SVM) [74]. In regression problems, the goal is to predict one or more continuous values. There are algorithms to use for regression are Linear Regression, Multi-Layer Perceptron, etc. [38]. Classification and regression models are classified as supervised learning algorithms because these algorithms require labeled data; a training dataset that contains labeled data, and every label is a class or continuo (continuous value) [69]. Sequence Prediction is when a system predicts (or forecasts) one or even more future values based on past data it has observed. This method is practically used in Natural Language Processing, bioinformatics, time series, and forecasting [83].

Descriptive Tasks: Clustering, Association, and Anomaly Detection.

Grouping items into clusters in which high similarity exists among objects (or instances) in the same cluster and low similarity in different clusters. Clustering is an unsupervised learning strategy. Common algorithms for clustering are Hierarchical Agglomerative Clustering, K-means++, K-Medoids, and Gaussian Mixture Models [81]. In addition, clustering is an area where neural networks (NNs) may be applied and in recent work has received more prominence.

Association is focused on discovering relationships among variables within a dataset. Association is widely used in association-based systems, such as recommendation systems. Algorithms commonly used with the association not limited to but include Apriori, Eclat, and FP-Growth [83].

Anomaly detection is identifying rare, usually unlabeled observations. Anomalies may represent something significant (e.g., some important incident or event) or may just be an error. In some cases, discriminating between what is important or an error depends on context, for example, fraud detection in banking; intrusion detection in cybersecurity; etc. [14].

3.4. Types of Datasets

Datasets are often categorized as structured or unstructured.

Structured Datasets: Think of a structured dataset as a table of rows and columns, in which the columns are referred to as features and the rows are instances. Structured datasets usually exist as a number of different file types, including CSV files, spreadsheets, XML or JSON files, or in a relational database like MySQL. Structured datasets are generally used in regression and classification tasks. In a more specific way, structured datasets can be classified as categorical datasets, numerical datasets, time series datasets, and spatial datasets [73]. Examples include Kaggle datasets and the UCI Machine Learning Repository.

Unstructured Datasets: These datasets lack a predefined structure and may include text, images, audio, or video data. They are frequently used in domains such as natural language processing (NLP), speech recognition, and computer vision. Unstructured datasets can be categorized into text, image, audio, and video datasets [40]. For example, Common Crawl is widely used in NLP tasks, while Reddit datasets are often used in sentiment analysis.

3.5. Software Fairness

Software fairness is a sensitive and important attribute that can be measured in any decision-making application such as search engines and recommendation systems. These systems are widely used in sensitive areas like healthcare, banking, law, and e-commerce. Such systems use algorithmic

 doi:10.20944/preprints202604.0038.v1

predictions based on big data to guide decisions and predict outcomes. Despite these benefits, it is necessary to ensure software fairness and avoid biased results.

Although software fairness is considered a desirable quality attribute, achieving it is not an easy task. MLS fairness is regarded as a main objective in artificial intelligence (AI) [9]. Researchers are focusing on applying techniques and tools to measure fairness in MLS. Numerous algorithms and technologies have been proposed to evaluate fairness in MLS using specific metrics [81].

3.6. Testing Fairness in MLS

MLS fairness can be affected in three stages: pre-processing, in-processing, and post-processing. In each stage, fairness is measured to ensure the production of a fair model. For instance, in the pre-processing phase, fairness is evaluated on the data before it is put into the model. In the in-processing phase, fairness is evaluated by testing algorithms and iterating to build a fair model. Finally, in the post-processing phase, test cases are evaluated that check for fairness in the predictions of MLS [14].

There are many fairness metrics that are quite familiar for evaluating fairness in MLS, including disparate impact, independence, demographic parity, equalized odds, fairness through awareness, and positive and negative class balance. Each metric will have its own formal conditions and definitions that must be satisfied in the MLS predictions [5].

3.7. Fairness and Bias in MLS

There are multiple sources of bias in MLS, and they can be summarized into three categories. They are, Data to Algorithm, Algorithm to User, and User to Data. [17]

In the Data to Algorithm category bias is derived from data that was not measured or reported correctly. Additionally, errors of data representation or errors reaching conclusions based on results would negatively impact the functionality of machine learning algorithms.

In the Algorithm to User category, bias can be created through design aspects in an algorithm, such as the use of specific optimization functions, and custom rules. Additionally, the manner in which results are ranked or displayed may include biases impacting the fairness of the model. Furthermore, the possible use of bias evaluation tools (Adience and IJB-A) may also lead to unfairness or bias in the fairness reiterated in the model, as both tools have reported issues with skin color.

In the Case of User to Data, any data source used to train MLS, is potentially biased due to being user-generated. Thus, any bias in user behavior or thought processes may be reflected in the data generated [70].

Types of bias in MLS can be summarized in the following points, [35] as shown in Figure 6.

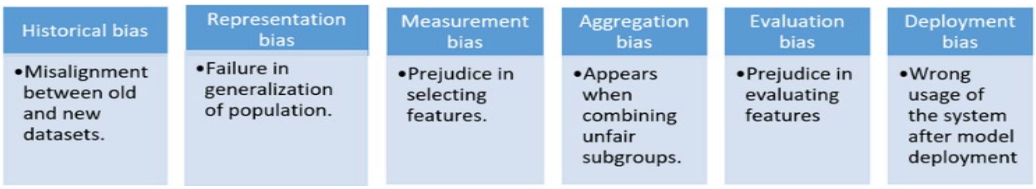


Figure 6. Overview of common types of bias in machine learning, including historical bias, representation bias, measurement bias, aggregation bias, evaluation bias, and deployment bias, along with examples of how each bias can arise.

Fairness evaluation in MLS is concerned with bias identification and mitigation along various stages of the model development process. The evaluation of fairness metrics in pre-processing stages focus on the quality and balance of input data, measuring variables such as statistical parity, disparate impact, or representation bias in order to identify and mitigate any imbalances prior to training. In the in-processing stage, fairness is measured by metrics such as equal opportunity, equalized odds, and demographic parity which are typically used as components in an ML algorithm, whether they are being explicitly included in the training process via a fairness-aware algorithm, or through

 doi:10.20944/preprints202604.0038.v1

regularizing the objective function. Finally, the post-processing stage of fairness evaluation trip includes looking at the outputs of the model in a fair way by comparing and contrasting group predictions, for example by calibration by group, or using individual fairness to avoid predictions being disproportionately favored or disfavored. These metrics help guide interventions to improve fairness without significantly reducing model performance.

The following section outlines the key findings and provides an interpretation of the results in the context of existing literature.

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

4. Results and Discussion

RQ1: What is the distribution of papers through venues and years?

The distribution of the 67 primary studies across venue types is shown in Table 4 and Table 5. Conferences were the most common outlet (52%, n=35), followed by journals (42%, n=28), and workshops (6%, n=4) (see Figure 7). This indicates that over half of the studies were disseminated at conferences, while a substantial proportion also appeared in journals.

Table 4. Primary Studies Which Are Published in Journals.

Venue	Study ID	Publisher
Journal of Data and Information Quality	PS26	ACM
Journal on Emerging Technologies in Computing Systems	PS3	ACM
Journal of Computing Sciences in Colleges	PS4	ACM
Journal of Artificial Intelligence Research	PS6	JAIR
IEEE Transactions on Industrial Informatics	PS8	IEEE
Journal of Machine Learning Research	PS13	JLMR
Neural Computing and Applications	PS20, PS28	Springer
Data Mining and Knowledge Discovery	PS21, PS57	Springer
Data Science and Engineering	PS23	Springer
International Journal of Data Science and Analytics	PS24	Springer
DBLP CoRR journal	PS26, PS60	DBLP
Knowledge-Based Systems	PS36	Elsevier
International Journal of Intelligent Systems	PS37	WILEY
IEEE Intelligent Systems	PS40	IEEE
Algorithms	PS44	MDPI
Advances in Neural Information Processing Systems	PS45	NeurIPS

 doi:10.20944/preprints202604.0038.v1

IEEE Transactions on Software Engineering	PS55	IEEE
International Journal of Crowd Science	PS58	DBLP
ACM Transactions on Software Engineering and Methodology	PS59	ACM
Journal of Technology in Human Service	PS63	DBLP
Electronics	PS64	MDPI
Expert Systems	PS66	WILY
ACM Transactions on Knowledge Discovery from Data	PS67	ACM
Venue	Study ID	Publisher
Journal of Data and Information Quality	PS26	ACM

*PS = Primary Study (see References [26–92]).

Table 4. Primary Studies Which Are Published in Conferences/Workshops.

Venue	Study ID	Publisher
The World Wide Web Conference	PS2, PS5	ACM
International Conference on Computer, Control, and Communication	PS7	IEEE
IEEE/ACM International Workshop on Software Fairness	PS9	IEEE/ACM
IEEE International Symposium on Technology and Society	PS10	IEEE
IEEE TrustCom 2020	PS11	IEEE
IEEE Conference on Decision and Control	PS12	IEEE
International Conference on Testing Software and Systems	PS14	SPRINGER
International Conference on the Quality of Information and Communications Technology	PS15, PS17	SPRINGER
International Conference on Computer-Aided Verification	PS16	SPRINGER
Joint European Conference on Machine Learning and Knowledge Discovery in Databases	PS18	SPRINGER
Companion Proceedings of the Web Conference 2021	PS22	ACM
Proceedings of the 23rd ACM SIGKDD	PS25	ACM
Proceedings of the Conference on Fairness, Accountability, and Transparency	PS27, PS30, PS38	ACM
Proceedings of the AAAI/ACM Conference on AI	PS29	ACM
ACL Workshop on Gender Bias for Natural Language Processing	PS31	ACL

 doi:10.20944/preprints202604.0038.v1

European Software Engineering Conference and Symposium on the Foundations of Software Engineering	PS32	ACM
ACM Conference (Conference'17)	PS33	ACM
Annual Meeting of the Association for Computational Linguistics	PS34	ACL
INNS Big Data and Deep Learning Conference	PS35	SPRINGER
Proceedings of the 36th International Conference on Machine Learning	PS41, PS42	PMLR
AAAI Conference on Artificial Intelligence	PS43	AAAI
International Conference on Software Engineering	PS46, PS47, PS51, PS52, PS54, PS56	IEEE/ACM
International Workshop on Equitable Data and Technology	PS48	ACM
IEEE International Conference on Software Testing, Verification, and Validation Workshop	PS49	IEEE
ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering	PS50	ACM
IEEE/ACM 7th International Workshop on Metamorphic Testing	PS53	IEEE/ACM
ACM SIGSOFT International Symposium on Software Testing and Analysis	PS61	ACM
International Conference on Evaluation and Assessment in Software Engineering	PS65	EASE

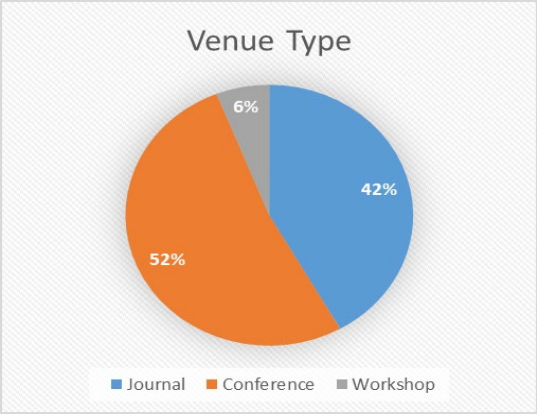


Figure 7. Distribution of primary studies by publication venue type, showing that 52% were published in conferences, 42% in journals, and 6% in workshops.

Figure 8 presents the yearly trend of publications. Research on fairness in MLS began to increase notably after 2019, with a peak in 2022 (n=14 studies, 21%) (see Figure 8). This sharp rise reflects the growing attention to fairness as an urgent challenge in MLS.

 doi:10.20944/preprints202604.0038.v1

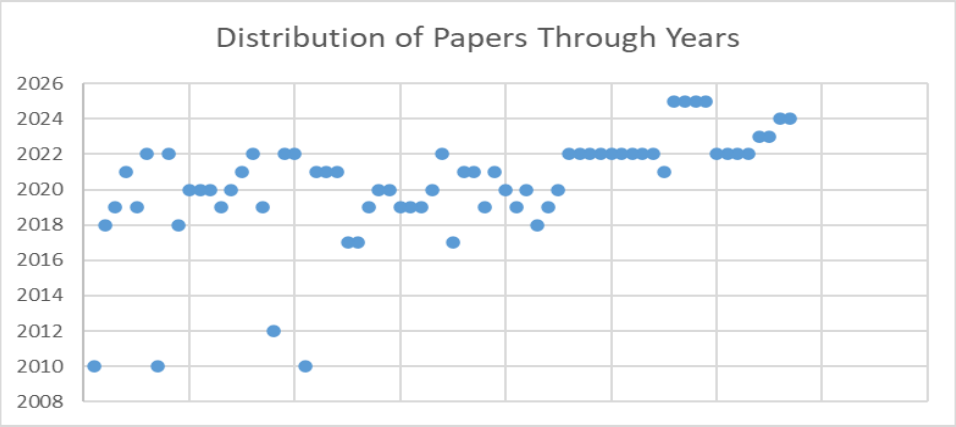


Figure 8. Distribution of the primary studies across publication years, illustrating the increasing research interest over time, with papers ranging from 2009 to 2025.

The dominance of conferences suggests that fairness detection in MLS remains an emerging and exploratory research area, where rapid dissemination of preliminary findings is prioritized. Journals, by contrast, tend to host more mature and detailed contributions, explaining why 42% of studies were published there. The comparable proportions highlight the complementary roles of both venues in shaping the field.

The increase after 2019 aligns with the intensifying global discourse on ethical AI and responsible machine learning, and indicates that fairness within computing, a previously niche concern, has now entered the larger fold of research within computing. The diversity of repositories in which research about fairness detection is published: conferences, journals, and workshops indicate the spirit of collaboration about fairness detection is interdisciplinary in nature. Contributions have been made from software engineering, data mining, and algorithm designing communities highlighting the diversity in technical communities exploring the topic (see Table 6).

Table 1. Primary Studies Which Are Published as Empirical Studies.

Venue	Study ID	Publisher
Empirical Software Engineering	**PS19, PS62**	**SPRINGER**

Furthermore, much of this research is anchored in decision-making domains with high societal impact, such as hiring, lending, and healthcare, where fairness is critical to prevent discriminatory outcomes. The presence of studies in these areas suggests that fairness detection is not only a theoretical challenge but also a pressing applied problem, driving the need for robust methods to detect and mitigate bias in MLS.

RQ2: What are the different definitions of software fairness?

The 67 primary studies put forth various definitions of software fairness and revealed the diversity of approaches in the field. The definitions of fairness were often found tucked away in the background or methodology sections of the papers; however, they illustrated various ways of operationalizing fairness in MLS. As shown in Table 7, we categorized definitions into four large categories of fairness which were general fairness, individual fairness, group fairness, and fairness metrics.

Table 7. summarizes the extracted fairness definitions, categorizing them according to their concepts and metric.

Category	Concept/Metric	Fairness Definition	Study ID
General Fairness	**Fairness**	**Ethical principle ensuring equitable and unbiased treatment across individuals/groups.**	**PS66**
	Fairness-aware Model	**A model that avoids discrimination and promotes fairness.**	**PS66**

 doi:10.20944/preprints202604.0038.v1

Individual Fairness	Fairness Degree	Max difference in predictions for pairs differing only in sensitive attributes.	PS19
	Fairness Through Unawareness	Fairness by excluding sensitive attributes from decision-making.	PS38
	Counterfactual Fairness	Prediction remains unchanged if the individual belongs to a different group.	PS38
	Algorithmic Bias	Bias from mathematical rules favoring certain attributes.	PS63
	Individual Fairness	Similar individuals should receive similar outcomes.	PS67, PS10, PS23, PS26
	Individual Discrimination	Discrimination between individuals differs only in protected attributes.	PS32, PS11, PS32
Group Fairness	Group Fairness	Equal outcomes across demographic groups (e.g., gender, race).	PS10, PS23, PS26
	Fairness Constraints	Constraints like demographic parity, equal opportunity, disparate impact.	PS23, PS26
Fairness Metrics	Demographic Parity	Outcome is independent of the protected attribute.	PS15, PS25
	Equal Opportunity	Equal true positive rates across groups.	PS25, PS6, PS39
	Predictive Parity	Equal positive predictive value across groups.	PS15, PS35
	Disparate Impact	Ratio of favorable outcomes between groups.	PS25, PS6
	Average Absolute Odds	Average difference in false/true positive rates across groups.	PS6, PS35
	Theil Index	Measures inequality in prediction outcomes.	PS6
	Calibration	Predicted score should reflect actual outcomes equally across groups.	PS35, PS38

Group fairness was the most popular approach, cited in 28 studies (42%). These definitions discuss groups, and equitable performance across groups that include sensitive attributes like gender, race, or age. The studies S15 and PS25 use demographic parity and predictive parity. [40,50] That is both equal outcomes that are independent of the protected features. The study PS23 referred to group fairness as fairness in output distribution across the sensitive groups [23].

Definitions of individual fairness were found in 18 studies (27%), based on the principle that similar individuals should be treated similarly. For example, PS10 noted that fairness requires that valid inputs that differ only in a protected attribute should receive identical classifications, while PS32 defined unfairness as discrimination between individuals that differ only in protected features [10,32].

A small number of studies (11 studies, 16%) proposed general fairness concepts, like fairness-aware models (PS66), degree of fairness (PS19), and fairness through unawareness (PS38) [19,38,66]. These concepts are often not precise categories but appear to be conceptual framing, mostly framed before researchers selected a metric-based evaluation of fairness.

 doi:10.20944/preprints202604.0038.v1

Finally, in 10 studies (15%), researchers evaluated fairness by using operational definitions in terms of specific fairness metrics. The metrics included disparate impact (PS25, PS6), equal opportunity (PS25, PS39), mean absolute odds (PS6, PS35), and Theil index (PS6) [6,25,35,39]. The other definitions included counterfactual fairness (PS38, PS56) and defining unfairness as the inverse of algorithmic bias (PS63) [38,56,63].

Overall, the results indicate that fairness in MLS studies is defined inconsistently in literature, with no definition dominating the field. The preponderance of the group fair metrics demonstrates that researchers appear to focus primarily on comparing populations instead of individuals, which aligns with the increased attention society has paid to the practical implications of discrimination through unfairness hiring or lending decisions (PS20, PS23) [20,23]. Individual fairness appears to be less well represented, though perhaps in spirit it is more important theoretically. However, researchers may find it difficult to define similarity between individuals practically.

The emergence of alternative and hybrid definitions also illustrates the efforts researchers are making to improve fairness evaluation. In the PS39 study, the authors indicated that anti-classification (unawareness), calibration (equity in individual fairness), and classification parity (group fairness) are similar but different and thus represent multidimensional approach [39]. The PS35 study emphasized the probabilistic nature of fairness, while the PS56 study analyzed prediction difference in pairwise definition [35,56]. These refinements indicate researchers are experimenting with definitions of fairness to characterize fairness in different contexts.

Ultimately, the results indicate a fragmented landscape, where fairness is conceived in multiple and sometimes conflicting ways. While this diversity captures the interdisciplinary nature of fairness research, it also makes comparisons and benchmarking across studies problematic. Compelling evidence suggests a strong need for greater consolidation and uniformity of definitions of fairness, while maintaining flexibility for researchers to employ definitions that work for the context.

RQ3: What types of problems are addressed?

The primary studies focused on four primary types of problem in MLS fairness detection: analyzing methods, developing bias mitigation methods, proposing methods for fairness testing, and analyzing methods. As shown in Figure 9 and detailed in Table 8, these categories exemplify how researchers conceptualize and operate the fairness problem.

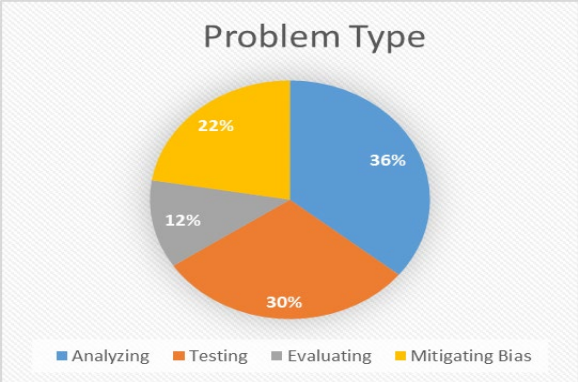


Figure 9. Distribution of primary studies by problem type, showing that 36% focus on analyzing bias, 30% on testing, 12% on evaluating, and 22% on mitigating bias.

Table 8. Distribution of Problem Types and Number of Studies.

Problem Type	Problem Description	Number of Studies	Studies IDs
Analyzing	**The process of systematically examining unfairness detection proposed methods to**	**24**	**PS1, PS4, PS5, PS6, PS12, PS16, PS18, PS20, PS21, PS22, PS24, PS25,**

 doi:10.20944/preprints202604.0038.v1

	understand their effectiveness and limitations.		**PS26, PS28, PS36, PS37, PS41, PS42, PS45, PS46, PS58, PS61, PS66**
Mitigating Bias	**The process of implementing strategies or algorithms that ensure fair outcomes and mitigate bias in unfairness detection systems.**	**15**	**PS3, PS7, PS8, PS13, PS15, PS23, PS27, PS31, PS34, PS43, PS48, PS53, PS59, PS64, PS67**
Testing	**The process of proposing testing solutions for checking whether the model produces fair outcomes and follows fairness metrics.**	**20**	**PS2, PS10, PS11, PS14, PS17, PS19, PS32, PS35, PS39, PS49, PS50, PS51, PS52, PS54, PS55, PS56, PS60, PS62, PS63, PS65**
Evaluating	**The process of comparing different approaches, analyzing outcomes, and validating results.**	**8**	**PS9, PS29, PS30, PS33, PS40, PS44, PS47, PS57**

Analyzing methods was the most popular type of problem type, represented by 24 studies (36%). These studies analyzed methods for unfairness detection systematically and identified strengths and weaknesses of the methods. For instance, PS5 optimized classification models by changing the fairness measures of the classifiers, and PS6 analyzed the impact of different mitigated methods on performance and fairness [5,6].

Testing problem types represented 20 studies (30%) and studies in this category analyzed a model to test whether they produced fair outcomes based on some sort of fairness metrics. For example, PS10 used sensitive variables (i.e., skin color) as a variable to check decision outcomes, and PS14 suggested a verification approach which developed fairness-based test cases [10,14].

Mitigating bias was the focus of 15 studies (22%). These contributions provided new methods and algorithms to reduce discrimination in MLS. For example, PS7 developed an alternative classification method and constructed a model that could develop unbiased models from biased training data, and PS23 proposed a fair outlier detection method that included more than one sensitive attribute (i.e., gender, race, religion, and nationality) [6,23].

Evaluating definitions and approaches was the least favored category with eight studies (12%). These works compared the concepts, criteria, or methods of fairness in some way to validate whether their content was applicable. For instance, PS9 summarized the various definitions of fairness and showed the definitions using case studies, while PS30 critically assessed the fairness criteria and mathematical techniques across diverse groups and individuals [9.30].

The distribution of problem types highlights the multifaceted nature of fairness detection research in MLS. The predominance of analyzing studies (36%) suggests that the field is still consolidating knowledge and testing the validity of existing methods. This aligns with the observation from RQ1 that much of the work is disseminated through conferences, reflecting an exploratory stage of development.

The substantial proportion of testing studies (30%) demonstrates that researchers are increasingly focused on building systematic evaluation frameworks for fairness. This emphasis indicates a move toward practical verification tools that can be integrated into software testing pipelines. However, these studies are still fragmented across different fairness metrics, limiting comparability.

Bias mitigation, though smaller in share (22%), reflects the growing importance of intervention-oriented approaches. The fact that many of these studies introduce algorithms specifically designed to remove or reduce bias demonstrates a shift from merely diagnosing fairness problems to actively addressing them.

The limited number of evaluation-focused studies (12%) reveals a gap in the field. Without thorough cross-comparisons of fairness definitions, metrics, and mitigation strategies, it is difficult to

 doi:10.20944/preprints202604.0038.v1

establish consensus on best practices. This finding reinforces the need for standardized benchmarks and more comprehensive comparative evaluations in future work.

In summary, the analysis of problem types shows that fairness detection research is evolving along multiple lines—analysis, testing, mitigation, and evaluation—but with imbalances. While analysis and testing dominate, mitigation and especially evaluation remain underrepresented, pointing to important directions for further investigation.

RQ4: What different approaches of fairness testing are presented?

In the review of the 67 primary studies, quite a few types of fairness testing methods were identified, and distributions can be seen in Figure 10. The most prevalent method was metric-based fairness testing (n=38, 57%), followed by synthetic data testing (n=17, 25%), and adversarial fairness testing (n=9, 13%). All other methods combined less graphically represented testing and included feature importance testing (n=2, 3%), comprehensive audits (n=2, 3%), and causal fairness testing (n=1, 1%). The results indicate a strong presence of quantitative evaluation methods, similar to other types of testing withing the scope of fairness within research.

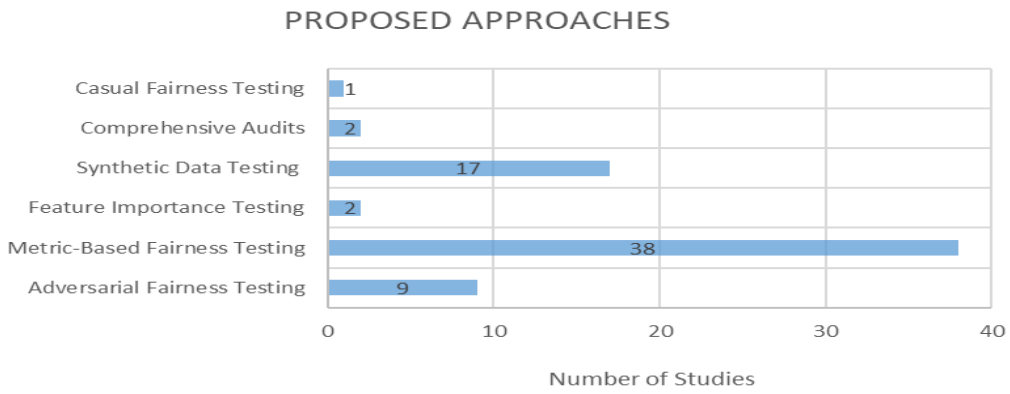


Figure 10. Summary of proposed fairness testing approaches across the reviewed studies, showing the distribution of methods: metric-based fairness testing (38 studies), synthetic data testing (17 studies), adversarial fairness testing (9 studies), comprehensive audits (2 studies), feature importance testing (2 studies), and causal fairness testing (1 study).

Finally, studies outlined actual tools and methods for bias detection and mitigation. For example, the IBM AI Fairness 360 tool kit, and fairkit-learn both have libraries to test fairness across datasets and algorithms (PS23, PS43) [23,43]. There were also approaches to rank algorithms to minimize bias in recommendation systems (PS15), impute and resample data to create a balance in dataset (PS24), and adjust word embeddings to minimize gender bias in natural language processing tasks SP35 [15,24,35].

With regard to levels of testing, predominately input level testing was carried out (n=25, 38%); there were also 18 studies that provided examples of system level testing (n=18, 27%) and 8 studies addressed unit level testing (n=8, 12%) of bias/fairness. There were another 15 studies (23%) that scaled the levels of testing with more than one level suggesting a promising multi-layered approach to testing. For instance, PS14 applied verification-based system testing, generating cases that examined both accuracy and fairness [14]. At the unit level, PS24 performed dataset-focused testing to address sampling bias [24]. Input testing was frequently applied in comparative experiments, such as PS4, which tested sampling strategies across neural network classifiers [4].

A summary of solutions associated with different problem types is provided in Table 9, covering optimization, reweighting, outlier detection, dataset augmentation, and hybrid white-box/black-box testing approaches.

 doi:10.20944/preprints202604.0038.v1

Table 9. Summary of Suggested Solutions.

Problem Type	Suggested Solutions
Analyzing	Optimization methods, Analysis of Fairness Metrics
Mitigating Bias	Mitigation methods include outlier detection, ranking, imputation, data massaging, dataset sampling, and data augmentation.
Testing	White/black box testing, test case generation tools, comprehensive audits, adversarial fairness testing, feature importance testing, metric-based fairness testing, synthetic data testing, and casual fairness testing.
Evaluating	Compare unfairness detection approaches, evaluate methods by using different benchmark datasets with suggested solutions, and compare multiple fairness metrics.

This major reliance (57%) on metric-based fairness testing reflects the extent to which the field relies on quantitative performance measures based on demographic parity, equal opportunity, predictive parity, etc. While extremely easy to measure and compare, the extent to which metric-based evaluations persist also suggests a degree of over-use, as they measure performance reported purely through statistical indicators - missing any deeper structural biases (PS6, PS15, PS25) [6,15,25].

Importantly, part of the growing use of synthetic data (25%) derived from the understanding that real historic data is likely embedded with embedded historical bias. Consequently, synthetic data allows for artificial but related datasets to be built that facilitate testing fairness (PS22, PS36, PS38) [22,36,38]. Similarly, somewhat in-line with this, the use of adversarial testing (13%) reflects attempts to stress test MLS models through provocative attempts at finding discriminatory cases (by design) (PS39, PS49) [39,49]. Though not as prevalent as other methods its value lies potentially chiefly in its ability to expose vulnerabilities that would otherwise go unnoticed.

The predominance to aggregate testing levels shows most studies are at least aware of measuring at the input level 38 since manipulation of datasets and controlled tests are highly useful for unveiling unfairness, however that 23% used more than one level suggests researchers are becoming aware of potentially using a more cohesive and integrated evaluation pipelines (see Figure 11). For example, deploying a combined input-level and system-level test could be conducted that engages both fairness evidence from either the data itself or measuring end-to-end performance (e.g., PS14, PS24) [14,24].

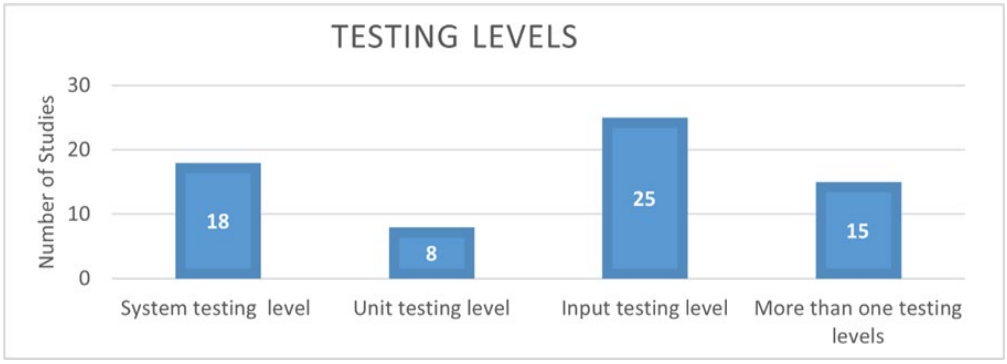


Figure 11. Distribution of primary studies across different testing levels, showing that 18 studies focus on system testing, 8 on unit testing, 25 on input-level testing, and 15 apply more than one testing level.

 doi:10.20944/preprints202604.0038.v1

The limited number of studies engaging with models that are able to do comprehensive auditing, causal testing or feature importance (collectively less than 10%) suggests considerable room for future inquiry into these developing areas. They are methodologies (underutilized) that could provide considerable contributions in their respective outputs - causal testing may highlight structural discrimination, audits may explore compliance with regulatory agreements, and feature importance may lend contributions in understanding attribute changes (PS38, PS63) [38,63]. Their inhibited engagements do not mean these areas are not relevant to the current or future engagement with fairness.

To that end, the results show that fairness testing in MLS is still somewhat metric driven and still principally remains focused at the input level though there are some studies also considering system-level degree of integration of the metrics or test paradigms. As a way forward, combining the vigilance of statistic metrics with the ability to intentionally deploy either adversarial, causal, or audit-based testing may offer more integrative and holistic ways of considering fairness, and improve the robustness and trustworthiness of MLS applications.

RQ5: Which datasets are used to detect the unfairness of MLS?

The primary studies primarily relied on limited sets of benchmark datasets to conduct fairness detection experiments. As seen in Figure 12 the most used dataset was the UCI Adult Census Income dataset (n=20, 30%), followed by the German Credit dataset (n=15, 22%), followed by the COMPAS dataset (n=11, 16%). Together these three datasets represented almost two-thirds of dataset usage in the reviewed studies.

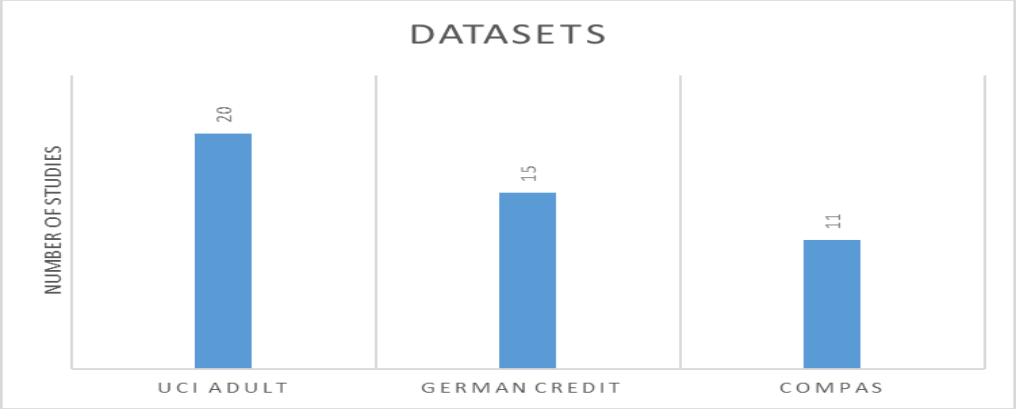


Figure 12. Distribution of the most commonly used datasets in the reviewed studies, showing that UCI Adult was used in 20 studies, German Credit in 15 studies, and COMPAS in 11 studies.

The benchmarks were most often used since they consist of sensitive demographic attributes (e.g., race, gender, and age), allowing for a fairness-related outcome to be formally assessed. For example, the adult dataset has included use cases to test classification models when gender representation is undesirably disproportionate (PS19, PS33) [19,33]. The German Credit dataset has sensitive variables (e.g., age, marital status), making it a unique example to test bias in financial decision-making scenarios (PS41, PS55) [41,55]. The COMPAS dataset is widely used by researchers examining fairness in the context of the criminal justice system and is cited in multiple works (PS28, PS48), despite its documented deficiencies regarding racial bias [28,48].

Alongside the benchmark datasets, synthetic datasets were used (PS36, PS58), [36,58] which allowed for controlled distributions of sensitive attributes and to test fairness constraints in varying conditions. The synthetic datasets were particularly useful for research where real-world data lacked the necessary diversity and where certain scenarios needed to be recreated for algorithm evaluation.

A few studies included domain-specific real-world datasets from healthcare (PS23, PS43) and social platforms to evaluate fairness in applied real-world decision-making situations; however, these studies were in the minority due to limited accessibility, privacy concerns, or a lack of standards [23,43].

 doi:10.20944/preprints202604.0038.v1

Ultimately, the study of dataset choice was a key theme influencing fairness evaluation results. For instance, PS33 highlighted that if a biased dataset is used, the resulting output will be discriminatory, even if the algorithms are theoretically fair. The study also included PS41 and PS55 which indicated that the representativeness of a dataset is arguably heavily influential on the applicability of any derived fairness metrics [33,41,55]

The results indicate that the fairness research agenda in MLS still largely relies on a small number of benchmark datasets – especially Adult, German Credit, and COMPAS. They are popular datasets used in the service of comparability and reproducibility among studies, but this reliance raises concerns. Each of these datasets has known biases (e.g., racial bias in COMPAS, gender bias in adult, and demographic bias in German Credit). Reuse of these would impose a narrow view of what is fair rather than opening vantage points to new contexts.

The use of synthetic datasets also carries evidence that researchers are trying to make up for these shortcomings by establishing a controlled testing environment, but synthetic datasets entail the risks of oversimplifying the complexity of the situation, leading to potential external validity reduction.

The small number of studies using healthcare and social media datasets indicate that we have not fully incorporated fairness research into sensitive real-world sectors where bias can have substantial consequences for real people. More often using healthcare and social media datasets can help improve the practical relevance of fairness detection studies, but this depends on some challenges such as privacy, access, and ethics.

To conclude, the choice of dataset has a decisive influence on the approach to and outcomes of fairness evaluation. We have highlighted an important tension in current methods to incorporate benchmarks that help guide which studies can be compared, but limit fairness research. There will still be a need for more work that tries to maintain a mixture of benchmarks with more diverse, representative, and domain related datasets aspects of fairness evaluation, and wider variety of social contexts.

RQ6: What are the research gaps and trends discovered in the reviewed studies?

The reviewed studies identified multiple research gaps that limit progress in fairness detection within MLS. As shown in Figure 13, the most frequently mentioned issue was the influence of training dataset quality (n=27, 40%), followed closely by preprocessing and missing value handling (n=25, 37%), and feature selection and sensitive attribute analysis (n=10, 15%). These findings reflect the importance of data-related factors in shaping fairness outcomes.

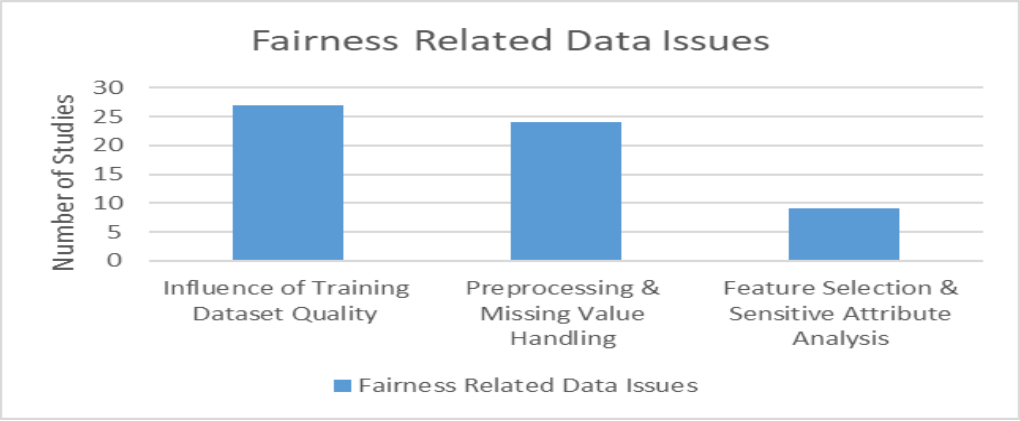


Figure 13. Fairness-related data issues identified across the reviewed studies, including the influence of training dataset quality (27 studies), preprocessing and missing-value handling (24 studies), and feature selection and sensitive attribute analysis (9 studies).

An ongoing gap across many papers was having no consistent fairness metrics. Many metrics exist, e.g., demographic parity, equal opportunity, predictive parity—but none fully captures fairness

 doi:10.20944/preprints202604.0038.v1

across contexts. PS22 and PS37 noted that researchers often applied metrics inconsistently across studies, and comparisons were difficult [22,37].

Another gap was the persistence of biased benchmark datasets. Datasets that we use often like COMPAS or Adult Census Income datasets are made with structural biases. PS33 and PS41 suggested that using the data without adjusting those data sets runs the risk of adding discrimination, even if you use fairness-aware algorithms [33,41].

The need for testing frameworks was emphasized. Although IBM AI Fairness 360 and Themis offer useful functionality, frameworks do not holistically address fairness across all stages of the MLS lifecycle. PS28, and PS46, emphasized the necessity of the integrated frameworks aimed at fairness checks during the data collection to deployment [28,46].

There are a number of current trends which emerged. Multiple articles used fairness methods to deepen learning and natural language processing (NLP) models. PS49 and PS54 examined fairness-aware word embeddings and regularization of neural networks as a method of managing bias in large-scale systems. Another emerging trend was the use of causal inference and fairness testing, which allows researchers to investigate the impact of sensitive attributes with more rigor [49,54].

Finally, interdisciplinary collaboration was highlighted in a number of studies. In PS18 and PS44 the authors argued that fairness requires input from computer science, ethics, law, and policy to ensure that solutions are implementable and aligned with collective societal values.

The analysis of gaps shows that fairness research is still grappling with data-driven limitations, particularly the reliance on benchmark datasets with known biases and the absence of standardized evaluation metrics. These limitations reduce the reproducibility and comparability of results across studies. The fact that dataset quality and preprocessing issues were mentioned in over three-quarters of the reviewed studies (see Figure 13) underscores the urgency of improving dataset curation and handling methods.

The key takeaway about the emerging need to have testing frameworks built into the entire MLS lifecycle is the acknowledgment that fairness cannot be achieved by stand-alone interventions. Interventions must be instituted across entire pipelines, including the data collection stage, pre-processing stage, model design stage, and when the model is deployed in the wild. Current tools (e.g., AI Fairness 360, Themis), provide valuable first steps to tackle bias in MLS, but do not give the complete integrated picture.

The themes we are seeing in deep learning, and NLP show that the field is reacting to the increasing influence of the models in more sensitive domains, like language translation applications and recommendation systems, since MLS are impacting society qua systemic racism & oppression. The use of fairness aware embeddings and neural regularizing techniques (PS49, PS54) show that fairness considerations are becoming more prominent and even integrated into the state of the art in machine learning [49,54]. The embrace of causal inference (PS61) also indicates a movement toward more explanatory forms of modeling that focuses on causal or structural sources of unfairness, and not just correlation [61].

The focus on collaborating representatives from multiple disciplines (PS18, PS44) means that fairness is seen not only as a technical issue, but also as a socio-technical issue [18,44]. Ethicists, lawyers, and policymakers need to be involved in this process to ensure that fairness definitions and metrics are appropriate to real-world constraints and societal appetite.

In brief, RQ6 indicates that the field is evolving, but the stakes remain high given inconsistent metrics, biased datasets, and a lack of consistent evaluation tools. However, encouraging developments, particularly in deep learning, causal inference and collaboration across disciplines, suggest a more thorough and holistic understanding of fairness in MLS is on the horizon (see Appendix 2).

5. Related Work

Several previously conducted surveys and reviews studied fairness in machine learning systems (MLS) from different perspectives. Caton et al. provided a comprehensive overview of approaches to

 doi:10.20944/preprints202604.0038.v1

reduce bias and categorized them into pre-processing, in-processing, and post-processing approaches [73]. Their review examined fairness metrics, explored the complexity of obtaining unbiased results, and provided suggestions for future research. Mehrabi et al. similarly reviewed real-world applications of bias in sensitive areas such as hiring and lending, summarized types, and sources of bias, and produced a taxonomy of definitions and measures of fairness for a guide to researchers [70].

Other complementary reviews have examined datasets and empirical foundations. Le Quy et al. provided a review of fairness-aware datasets used in fairness-aware research studies and examined tabular data with Bayesian networks to show relationships between protected attributes and locations of unfairness [50]. Other reviews, e.g., Richardson et al., [74] emphasized algorithmic bias and fairness interventions and critiques of techniques and tools for bias mitigation and pointed out differences between theory and practice.

Apart from surveys, individual studies have also added more pointed knowledge. For example, a master's thesis by Maha Alkatheri introduced software ontology for fairness-aware test case generation, [71] using the German Credit dataset to study gender biases. Chen et al. [81] conducted a bibliometric review of fairness testing until 2023 by looking at testing workflows such as test input generation and identifying an oracle. Their study defined a testing taxonomy at three levels (data level, program level, model level), and set out several exciting directions for future research, such as how to use public datasets and open-source tools.

While these studies provide useful descriptions, they often examine fairness only from one angle, whether it be bias mitigation methods (input testing), data sets or testing workflows, without synthesizing the definitions, problem type, testing pathway, datasets, and emerging research gaps into a single unified framework. Furthermore, most the reviews concluded prior to 2025, thus did not afford consideration of the advances in deep learning fairness, causal inference methods, and interdisciplinary perspectives, in their analyses.

Our research directly tackles this gap through a systematic mapping study (SMS) of software unfairness detection in MLS covering all relevant literature from 2010 to 2025. Unlike previous review studies which discussed fairness issues in isolation, our review synthesizes fairness definitions, research challenges, testing strategies, datasets, algorithms, and future trends into one cohesive whole. To accomplish this, we provided a rigorous SMS methodology that included specific classification criteria, database search strategies, and quality assessment protocols in order to ensure systematic coverage. This allows our research to be able to extend prior works and be updated within the updated knowledge typology with respect to a current perspective on the state of fairness research in MLS.

6. Conclusions

This study presents a systematic mapping study (SMS) of fairness in machine learning based systems (MLS), synthesizing the research published from 2010 to 2025. The SMS approached six research questions (RQ) systematically, investigating definitions of fairness, several types of problems, types of testing, datasets, as well as identifying what research gaps and emerging trends were addressing. This study collected evidence based on sixty-seven primary studies and provides the most thorough overview of how fairness has been defined, measured, and challenged in terms of research about MLS.

The finding's showed fairness is a contested concept with multiple operational definitions ranging from some notion of group fairness, individual fairness, and counterfactual definitions. These perspectives bring richness to the field; however, the absence of standardized and universally accepted metrics that permit comparability across studies continues to remain a barrier (RQ2).

The analysis of the problem types (RQ3) showed that research was scattered across four categories: taking existing approaches to analyze; developing various methods for bias mitigating strategies; implementing fairness with multiple methods; and testing definitions. The most common approaches in fairness testing (RQ4) entailed metric-based evaluation, synthetic data generation, and

 doi:10.20944/preprints202604.0038.v1

adversarial testing, at either the input, system, or unit level. Although all the practical toolkits (IBM AI Fairness 360, Themis, Aequitas, etc.) have touted fairness determination methods - none have provided an end-to-end solution to help shape fairness through the MLS process.

Datasets emerged as a major driver of fairness outcomes (RQ5). Each of the major benchmark datasets (e.g., COMPAS, German Credit, UCI Adult) is still used regularly, despite the acknowledgement that bias exists in these datasets and that they may not be adequate for constructing fairness-aware systems. Synthetic datasets should allow for controlled contextual testing but can lead to distrust as they cannot justify outcomes based on people's realities. The absence of diverse, representative, and unbiased datasets remains one of the greatest barriers to developing fairness research.

In regard to research gaps (RQ6), the SMS identified three major gaps: (1) no standard fairness definitions and metrics, (2) deviance from biased benchmark datasets, and (3) lack of fairness testing tools that are practical for deployment. At the same time, there are several emerging trends: fairness-aware deep learning, fairness in natural language processing, the use of causal inference-based approaches, and increasing calls for interdisciplinary teams and cooperation across computer assistance, ethics, and policy.

Overall, we have made a contribution to the research field by providing a structured synthesis of fairness in MLS research beyond previous reviews. By mapping definitions, approaches, datasets, and gaps into one coherent framework, this paper cements this space for future research. Next steps will depend on building standardized fairness metrics, designing more representative datasets, and implementing robust fairness testing frameworks for all lifecycle aspects of MLS. Ultimately, the challenges outlined above are not solely technical challenges but also societal challenges, which must be overcome to ensure that MLS are trusted in real-life high stakes applications such as health care, finance, or justice.

Author Contributions: Conceptualization, R.A. and N.A.; methodology, R.A.; software, R.A.; validation, R.A. and N.A.; formal analysis, R.A.; investigation, R.A.; resources, N.A.; data curation, R.A.; writing—original draft preparation, R.A.; writing—review and editing, R.A. and N.A.; visualization, R.A.; supervision, N.A.; project administration, N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: The author would like to express sincere gratitude to Dr. Noureddine Abbadeni for his continuous guidance, valuable feedback, and supervision throughout the development of this research. The support and resources provided by King Saud University and King Abdulaziz City of Science and Technology are also gratefully acknowledged. During the preparation of this manuscript, the author used their digital librarbie to collect primary studies; the author has reviewed and verified all outputs and takes full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
DL	Deep Learning
NN	Neural Network
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers

 doi:10.20944/preprints202604.0038.v1

GPT Generative Pre-trained Transformer

Appendix A: Quality Assurance

Table A1. Quality Assessment of Collected Primary Studies.

ID	Reference	Title	QA 1	QA 2	QA 3	Score	ID after QA
S1	Tremblay, Monica Chiarini, Kaushik Dutta, and Debra Vandermeer. Journal of Data and Information Quality (JDIQ) 2.1 ACM (2010)	Using data mining techniques to discover bias patterns in missing data.	1	1	1	3	PS1
S2	Krasanakis, Emmanouil, et al. Proceedings of the 2018 World Wide Web Conference. ACM 2018.	Adaptive sensitive reweighting to mitigate bias in fairness-aware classification	1	1	0.5	2.5	PS2
S3	Wang, Weijia, and Bill Lin. ACM Journal on Emerging Technologies in Computing Systems (JETC) 15.2 (2019): 1-17.	Trained biased number representation for ReRAM-based neural network accelerators.	1	1	1	3	PS3
S4	Thambawita, Vajira, et al. ACM Transactions on Computing for Healthcare 1.3 (2020): 1-29.	An extensive study on cross-dataset bias and evaluation metrics interpretation for machine learning applied to gastrointestinal tract abnormality classification.	1	0.5	0	1.5	
S5	Amend, Jack J., and Scott Spurlock. Journal of Computing Sciences in Colleges ACM 36.5 (2021): 14-23.	Improving machine learning fairness with sampling and adversarial learning.	1	1	0.5	2.5	PS4
S6	Baniecki, Hubert, et al. The Journal of Machine Learning Research 22.1 (2021): 9759-9765.	dalex: Responsible machine learning with interactive explainability and fairness in Python.	1	0.5	0	1.5	

 doi:10.20944/preprints202604.0038.v1

S7	Wu, Yongkai, Lu Zhang, and Xintao Wu. The World Wide Web Conference. ACM 2019.	On convexity and bounds of fairness-aware classification.	1	1	0.5	2.5	PS5
S8	Caton, Simon, Saiteja Malisetty, and Christian Haas. Journal of Artificial Intelligence ACM Research 74 (2022): 1011-1035.	Impact of Imputation Strategies on Fairness in Machine Learning."	1	1	1	3	PS6
S9	Kamiran, Faisal, and Toon Calders. 2009 2nd international conference on computer, control, and communication. IEEE, 2009.	Classifying without discriminating.	1	1	0.5	2.5	PS7
S10	DeBrusk, Chris. MIT Sloan Management Review (2018).	The risk of machine-learning bias (and how to prevent it.	1	0.5	0	1.5	
S11	Zhou, Xiaokang, et al. IEEE Transactions on Industrial Informatics (2022).	Distribution Bias Aware Collaborative Generative Adversarial Network for Imbalanced Deep Learning in Industrial IoT.	1	0.5	0.5	2	PS8
S12	Verma, Sahil, and Julia Rubin. 2018 ieee/acm international workshop on software fairness (fairware). IEEE, 2018.	Fairness definitions explained.	1	1	0	2	PS9
S13	KIEMDE, Sountongnoma Martial Anicet, and Ahmed Dooguy KORA. 2020.	Fairness of Machine Learning Algorithms for the Black Community	1	0.5	0.5	2	PS10
S14	Xie, Wentao, and Peng Wu. 2020.	Fairness Testing of Machine Learning Models Using Deep Reinforcement Learning	1	1	1	3	PS11
S15	Olfat, Matt, and Yonatan Mintz. 2020.	Flexible Regularization Approaches for Fairness in Deep Learning	1	1	0.5	2.5	PS12

 doi:10.20944/preprints202604.0038.v1

S16	Zafar, Muhammad Bilal, et al. The Journal of Machine Learning Research 20.1 (2019): 2737-2778.	Fairness constraints: A flexible approach for fair classification.	1	1	1	3	PS13
S17	Sharma, Arnab, and Heike Wehrheim, IFIP International Conference on Testing Software and Systems. Springer, Cham, 2020.	Automatic fairness testing of machine learning models.	1	1	1	3	PS14
S18	Villar, David, and Jorge Casillas. International Conference on the Quality of Information and Communications Technology. Springer, Cham, 2021.	Facing Many Objectives for Fairness in Machine Learning.	1	0.5	1	2.5	PS15
S19	Guan, Ji, Wang Fang, and Mingsheng Ying. International Conference on Computer Aided Verification. Springer, Cham, 2022.	Verifying Fairness in Quantum Machine Learning.	1	1	1	3	PS16
S20	Shin Nakajima and Tsong Yueh Chen (2019)	Generating Biased Dataset for Metamorphic Testing of Machine Learning Programs	1	0.5	0.5	2	PS17
S21	Kamishima, Toshihiro, et al. Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2012.	Fairness-aware classifier with prejudice remover regularizer.	1	1	0.5	2.5	PS18
S22	Perera, Anjana, et al. 2022	Search-based fairness testing for regression-based machine learning systems.	1	1	1	3	PS19

 doi:10.20944/preprints202604.0038.v1

S23	Tian, Huan, et al. Neural Computing and Applications (2022): 1-19.	Image fairness in deep learning: problems, models, and challenges.	1	0.5	0.5	2	PS20
S24	Calders, Toon, and Sicco Verwer. Data mining and knowledge discovery 21.2 (2010): 277-292.	Three naive Bayes approaches for discrimination-free classification.	1	1	0.5	2.5	PS21
S25	Sun, Haipei, et al. Companion Proceedings of the Web Conference 2021. 2021.	Automating fairness configurations for machine learning.	1	0.5	1	2.5	PS22
S26	Abraham, Savitha Sam. Data Science and Engineering 6.4 (2021): 485-499.	FairLOF: Fairness in Outlier Detection.	1	1	1	3	PS23
S27	Wang, Yanchen, and Lisa Singh International Journal of Data Science and Analytics 12.2 (2021): 101-119.	Analyzing the impact of missing values and selection bias on fairness.	1	1	1	3	PS24
S28	Corbett-Davies, Sam, et al. Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining. 2017.	Algorithmic decision making and the cost of fairness.	1	1	0.5	2.5	PS25
S29	Berk, Richard, et alarXiv preprint arXiv:1706.02409 (2017).	A convex framework for fair regression.	1	1	0,.5	2.5	PS26
S30	Celis, L. Elisa, et al. Proceedings of the conference on fairness, accountability, and transparency. 2019.	Classification with fairness constraints: A meta-algorithm with provable guarantees.	1	1	1	3	PS27
S31	Prates, Marcelo OR, Pedro H. Avelar, and Luís C. Lamb Neural Computing and Applications 32.10 (2020): 6363-6381.	Assessing gender bias in machine translation: a case study with Google translates.	1	1	1	3	PS28

 doi:10.20944/preprints202604.0038.v1

S32	Fazelpour, Sina, and Zachary C. Lipton. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.	Algorithmic Fairness from a Non-ideal Perspective	1	1	0.5	2.5	PS29
S33	Hutchinson, Ben, and Margaret Mitchell. Proceedings of the conference on fairness, accountability, and transparency. 2019.	50 years of test (un) fairness: Lessons for machine learning.	1	0.5	0.5	2	PS30
S34	Font, Joel Escudé, and Marta R. Costa-Jussa. arXiv preprint arXiv:1901.03116 (2019).	Equalizing gender biases in neural machine translation with word embedding techniques.	1	1	1	3	PS31
S35	Aggarwal, Aniya, et al. Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2019.	Black box fairness testing of machine learning models.	1	1	1	3	PS32
S36	Jones, Gareth P., et al. arXiv preprint arXiv:2010.03986 (2020).	Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms.	1	1	0.5	2.5	PS33
S37	Krishna, Satyapriya, et al. arXiv preprint arXiv:2203.08670 (2022).	Measuring Fairness of Text Classifiers via Prediction Sensitivity.	1	1	0.5	2.5	PS34
S38	Barocas, Solon, Moritz Hardt, and Arvind Narayanan. Nips tutorial 1 (2017)	Fairness in machine learning.	1	0.5	0.5	2	PS35
S39	Varley, Michael, and Vaishak Belle. Knowledge-Based Systems 215 (2021): 106715.	Fairness in machine learning with tractable models.	1	1	0.2	2.5	PS36

 doi:10.20944/preprints202604.0038.v1

S40	Valdivia, Ana, Javier Sánchez-Monedero, and Jorge CasillasInternational Journal of Intelligent Systems 36.4 (2021): 1619-1643.	How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness.	1	1	1	3	PS37
S41	Friedler, Sorelle A., et al. Proceedings of the conference on fairness, accountability, and transparency. 2019.	A comparative study of fairness-enhancing interventions in machine learning.	1	1	0.5	2.5	PS38
S42	Ferrari, Elisa, and Davide Bacciu. arXiv preprint arXiv:2105.06345 (2021).	Addressing Fairness, Bias, and Class Imbalance in Machine Learning: the FBI-loss.	1	1	1	3	PS39
S43	Du, Mengnan, et al. IEEE Intelligent Systems 36.4 (2020): 25-34.	Fairness in deep learning: A computational perspective.	1	0.5	0.5	2	PS40
S44	Huang, Lingxiao, and Nisheeth Vishnoi. International Conference on Machine Learning. PMLR, 2019.	Stable and fair classification.	1	0.5	0.5	2	PS41
S45	Celis, L. Elisa, et alInternational Conference on Machine Learning. PMLR, 2021.	Fair classification with noisy protected attributes: A framework with provable guarantees.”	1	1	0.5	2.5	PS42
S46	Goel, Naman, Mohammad Yaghini, and Boi Faltings. Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.	Non-discriminatory machine learning through convex fairness criteria.	1	1	0.5	2.5	PS43
S47	Shrestha, Yash Raj, and Yongjie Yang. Algorithms 12.9 (2019): 199.	Fairness in algorithmic decision-making: Applications in multi-winner voting, machine	1	0.5	0.5	2	PS44

 doi:10.20944/preprints202604.0038.v1

		learning, and recommender systems.					
S48	Mandal, Debmalya, et al. Advances in neural information processing systems 33 (2020): 18445-18456.	Ensuring fairness beyond the training data.	1	0.5	0.5	2	PS45
S49	Gao, Xuanqi, et al.	FairNeuron: improving deep neural network fairness with adversary games on selective neurons.	1	1	1	3	PS46
S50	Tizpaz-Niari, Saeid, et al.	Fairness-aware configuration of machine learning libraries.	1	1	0.5	2.5	PS47
S51	Chakraborty, Joymallya, Suvodeep Majumder, and Huy Tu.	Fair-SSL: Building fair ML Software with less data.	1	0.5	0.5	2	PS48
S52	Patel, Ankita Ramjibhai, et al.	A combinatorial approach to fairness testing of machine learning models.	1	1	1	3	PS49
S53	Chen, Zhenpeng, et al.	MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software.	1	0.5	1	2.5	PS50
S54	Li, Yanhui, et al.	Training data debugging for the fairness of machine learning software.	1	0.5	0.5	2	PS51
S55	Fan, Ming, et al.	Explanation-guided fairness testing through genetic algorithm.	1	1	0.5	2.5	PS52
S56	Pu, Muxin, et al.	Fairness evaluation in deepfake detection models using metamorphic testing.	1	1	1	3	PS53
S57	Zheng, Haibin, et al.	Neuronfair: Interpretable white-box fairness testing through biased neuron identification.	1	1	0.5	2.5	PS54

 doi:10.20944/preprints202604.0038.v1

S58	Zhang, Peixin, et al.	Automatic fairness testing of neural classifiers through adversarial sampling.	1	0.5	0.5	2	PS55
S59	Zhang, Peixin, et al.	White-box Fairness Testing through Adversarial Sampling	1	0.5	1	2.5	PS56
S60	Fabris, Alessandro, et al	Algorithmic fairness datasets: the story so far.	1	0.5	0.5	2	PS57
S61	Zhang, Jiehuang, Ying Shu, and Han Yu.	Fairness in design: A framework for facilitating ethical artificial intelligence designs.	1	0.5	0.5	2	PS58
S62	Majumder, Suvodeep, et al. "	Fair enough: Searching for sufficient measures of fairness.	1	0.5	0.5	2	PS59
S63	Wang, Zichong, et al. "	Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking.	1	1	1	3	PS60
S64	Guo, Huizhong, et al.	Fairrec: fairness testing for deep recommended systems.	0.5	1	0.5	2	PS61
S65	Hort, Max, et al.	Search-based automatic repair for fairness and accuracy in decision-making software.	1	1	0.5	2.5	PS62
S66	Bantilan, Niels.	Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation.	1	1	1	3	PS63
S67	Ling, Jiasheng, et al.	Machine Learning-Based Multilevel Intrusion Detection Approach	1	1	0.5	2.5	PS64
S68	Nasiri, Roya.	Testing Individual Fairness in Graph Neural Networks	1	0.5	0.5	2	PS65

 doi:10.20944/preprints202604.0038.v1

S69	Consuegra-Ayala et al.	Bias mitigation for fair automation of classification tasks.	1	1	1	3	PS66
S70	Paiheng Xu et al.	GFairHint: Improving Individual Fairness for Graph Neural Networks via Fairness Hint.	1	1	1	3	PS67
S71	Bahangulu et al.	Algorithmic bias, data ethics, and governance: Ensuring fairness, transparency, and compliance in AI-powered business analytics applications	0.5	0.5	0.5	1.5	

 doi:10.20944/preprints202604.0038.v1

Appendix B: Research Gaps Summary

Table B1. Research Gaps.

Category	Sub-category	Research Gap	Primary Study ID
1. fairness in machine learning	Fairness Metrics and Definitions	• Investigate incorporating fairness metrics into neural networks.	PS4
		• Measures for fairness.	PS18
		• Improve the definitions of fairness in data analysis.	PS22
			PS23
		• Study other fairness metrics beyond individual fairness.	PS30
			PS36
		• Much research relies on specific fairness metrics while ignoring others.	PS59
	Fairness Across ML Tasks	• Include fairness constraints in supervised (regression, recommendation) and unsupervised (set selection, ranking) tasks.	PS12
			PS13
		• Fairness in natural language understanding, resource allocation, representation learning, and causal learning.	PS15
			PS26
			PS44
		• Fairness in regression-based systems.	PS46
		• Discover limitations in fair regression.	PS61
		• Fairness in CNNs and DNNs.	
		• Extend DNN fairness testing to CNNs.	
		• More effective mitigation strategies beyond simple ML modifications.	
	Bias Mitigation Techniques	• Retraining as a solution to genetic fairness testing.	PS29
			PS47
		• Explore hyper-parameter configurations that lead to high fairness.	PS48
			PS50
		• Improve semi-supervised techniques to achieve fairness with limited labeled data.	PS52
			PS66
		• Investigate ways to adjust the learning process to account for biases.	
	Sensitive Attributes & Social Constructs	• Study numerical attributes and groups of attributes as sensitive.	
		• Consider income, race, religious beliefs, age, nationality.	PS7
			PS21
		• Explore multiple and continuous sensitive features.	PS31
		• Study other social constructs and stereotypes.	

 doi:10.20944/preprints202604.0038.v1

	Fairness Testing Frameworks	• Extend reinforcement learning-based testing for fairness. • Extend Themis for algorithmic bias testing. • FairRec for multi-attribute group fairness testing. • White-box and black-box fairness testing. • Explore other measures of fairness in white-box testing. • More techniques for black-box testing to detect individual discrimination. • Evaluate fairness-accuracy tradeoffs in real-world scenarios.	PS14 PS19 PS49 PS61 PS63
	Fairness in Real-World Contexts	• Fairness in temporal settings. • Study real-world data issues and their impact on fairness. • Add data documentation to future projects. • Extend fairness to ethical values like privacy and explainability.	PS27 PS57
2. Feature Selection & Data Handling		• Improve quality of datasets with test case generation approaches. • Feature selection in large datasets. • Investigating advanced feature selection techniques. • Adding datasets, algorithms, and imputation strategies. • Preprocessing of missing training data. • Pre-processing training data. • Study other datasets and fairness-accuracy tradeoffs.	PS1 PS6 PS24 PS37 PS41 PS43 PS45 PS51 PS64
3. Model Optimization & Training Efficiency		• Explore different characteristics in training data. • New ways for faster training processes. • Training CNNs with low-precision weights. • Improve algorithms for efficiency and equity. • Improve stability by shifting decision boundaries. • Implement models in industrial IoT for imbalanced learning. • Improve quantum decision models with fairness guarantees. • Enhance scalability of fairness testing for large datasets and complex graphs.	PS2 PS3 PS8 PS16 PS25 PS65 PS66
4. Testing & Evaluation Frameworks		• Improve white-box models for robustness. • Explore test case generation approaches. • DNN white-box testing challenges.	PS17 PS32 PS33

 doi:10.20944/preprints202604.0038.v1

	• Try different equivalence classes for retraining. • Study bias kernels detected by verification algorithms. • Neuron coverage as a distortion metric. • Explore causal techniques, post-processing, interpretability, calibration. • Extend counterfactual thinking to text and image processing	PS34 PS54 PS55 PS56 PS60
5. Reinforcement Learning & Reward Functions	• Study other definitions of reward functions in black-box testing. • Extend RL-based testing frameworks. • Determine ideal G-ratio for fairness testing.	PS11
6. Ethics, Law, and Societal Impact	• Types of laws or regulations. • Ethical difficulty in statistical machine translation. • Algorithms raise complex questions for researchers and policymakers. • More debates on fairness including technical and cultural causes. • Transparency in model complexity and fairness tradeoff. • Algorithmic choices and social context. • Fair unified solutions. • Interdisciplinary collaboration (CS, statistics, cognitive science).	PS28 PS30 PS38 PS39 PS58
7. Emerging Applications & Techniques	• Deep clustering, adversarial training, and attacks. • Study proposed methods with image compression and deepfake techniques. • Investigate noise models for non-binary attributes. • Long-term studies on bias mitigation effects.	PS20 PS42 PS53

References

1. Tremblay, Monica Chiarini, Kaushik Dutta, and Debra Vandermeer. "Using data mining techniques to discover bias patterns in missing data." Journal of Data and Information Quality (JDIQ) 2.1 ACM (2010). DOI: 10.1145/1805286.1805288.
2. Krasanakis, Emmanouil, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification." Proceedings of the 2018 World Wide Web Conference. ACM (2018). DOI: 10.1145/3178876.3186133.
3. Wang, Weijia, and Bill Lin. "Trained biased number representation for ReRAM-based neural network accelerators." ACM Journal on Emerging Technologies in Computing Systems (JETC) 15.2 (2019): 1-17. DOI: 10.1145/3318163.

 doi:10.20944/preprints202604.0038.v1

4. Amend, Jack J., and Scott Spurlock. "Improving machine learning fairness with sampling and adversarial learning." Journal of Computing Sciences in Colleges ACM 36.5 (2021): 14-23. DOI: 10.1145/3318163.
5. Wu, Yongkai, Lu Zhang, and Xintao Wu. "On convexity and bounds of fairness-aware classification." The World Wide Web Conference. ACM (2019). DOI: 10.1145/3308558.3313723.
6. Caton, Simon, Saiteja Malisetty, and Christian Haas. "Impact of Imputation Strategies on Fairness in Machine Learning." Journal of Artificial Intelligence Research 74 (2022): 1011-1035. DOI: 10.1613/jair.1.13197.
7. Kamiran, Faisal, and Toon Calders. "Classifying without discriminating." 2nd International Conference on Computer, Control, and Communication. IEEE (2009). DOI: 10.1109/IC4.2009.4909197.
8. Zhou, Xiaokang, Yiyong Hu, Jiayi Wu, Wei Liang, Jianhua Ma, and Qun Jin. "Distribution Bias Aware Collaborative Generative Adversarial Network for Imbalanced Deep Learning in Industrial IoT." IEEE Transactions on Industrial Informatics (2022). DOI: 10.1109/TII.2022.3161234.
9. Verma, Sahil, and Julia Rubin. "Fairness definitions explained." IEEE/ACM International Workshop on Software Fairness (Fairware). IEEE (2018). DOI: 10.1145/3194770.3194776
10. KIEMDE, Sountongnoma Martial Anicet, and Ahmed Dooguy. "Fairness of Machine Learning Algorithms for the Black Community." KORA (2020). DOI: 10.1109/ISTAS50296.2020.9462194.
11. Xie, Wentao, and Peng Wu. "Fairness Testing of Machine Learning Models Using Deep Reinforcement Learning." 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE (2020). DOI: 10.1109/TrustCom50675.2020.00023.
12. Olfat, Matt, and Yonatan Mintz. "Flexible Regularization Approaches for Fairness in Deep Learning." 2020 59th IEEE Conference on Decision and Control (CDC). IEEE (2020). DOI: 10.1109/CDC42340.2020.9304116.
13. Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. "Fairness constraints: A flexible approach for fair classification." The Journal of Machine Learning Research 1 (2019): 2737-2778. DOI: 10.48550/arXiv.1507.05259.
14. Sharma, Arnab, and Heike Wehrheim. "Automatic fairness testing of machine learning models." IFIP International Conference on Testing Software and Systems. Springer, Cham (2020). DOI: 10.1007/978-3-030-64847-1_10.
15. Villar, David, and Jorge Casillas. "Facing Many Objectives for Fairness in Machine Learning." International Conference on the Quality of Information and Communications Technology. Springer, Cham (2021). DOI: 10.1007/978-3-030-85347-9_8.
16. Guan, Ji, Wang Fang, and Mingsheng Ying. "Verifying Fairness in Quantum Machine Learning." International Conference on Computer Aided Verification. Springer, Cham (2022). DOI: 10.1007/978-3-030-67591-0_10.
17. Nakajima, Shin, and Tsong Yueh Chen. "Generating biased dataset for metamorphic testing of machine learning programs." IFIP International Conference on Testing Software and Systems. Springer, Cham (2019). DOI: 10.1007/978-3-030-31280-8_10.
18. Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. "Fairness-aware classifier with prejudice remover regularizer." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg (2012). DOI: 10.1007/978-3-642-33486-3_3.
19. Perera, Anjana, Aldeida Aleti, Chakkrit Tantithamthavorn, Jirayus Jiarpakdee, Burak Turhan, Lisa Kuhn, and Katie Walker. "Search-based fairness testing for regression-based machine learning systems." Empirical Software Engineering 27.3 (2022): 1-36. DOI: 10.1007/s10664-021-10006-3.
20. Tian, Huan, Tianqing Zhu, Wei Liu, and Wanlei Zhou. "Image fairness in deep learning: problems, models, and challenges." Neural Computing and Applications (2022): 1-19. DOI: 10.1007/s00521-022-07136-1.
21. Calders, Toon, and Sicco Verwer. "Three naive Bayes approaches for discrimination-free classification." Data Mining and Knowledge Discovery 21.2 (2010): 277-292. DOI: 10.1007/s10618-010-0190-x.
22. Sun, Haipei, Yiding Yang, Yanying Li, Huihui Liu, Xinchao Wang, and Wendy Hui Wang. "Automating fairness configurations for machine learning." Companion Proceedings of the Web Conference (2021). DOI: 10.1145/3442442.3452302.
23. Abraham, Savitha Sam. "FairLOF: Fairness in Outlier Detection." Data Science and Engineering 6.4 (2021): 485-499. DOI: 10.1007/s41019-021-00169-x.

 doi:10.20944/preprints202604.0038.v1

24. Wang, Yanchen, and Lisa Singh. "Analyzing the impact of missing values and selection bias on fairness." International Journal of Data Science and Analytics 12.2 (2021): 101-119. DOI: 10.1007/s41060-021-00259-z.
25. Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic decision making and the cost of fairness." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017). DOI: 10.1145/3097983.3098095.
26. Berk, Richard, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. "A convex framework for fair regression." Not published arXiv preprint arXiv:1706.02409 (2017). DOI: 10.48550/arXiv.1706.02409.
27. Celis, L. Elisa, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. "Classification with fairness constraints: A meta-algorithm with provable guarantees." Proceedings of the Conference on Fairness, Accountability, and Transparency (2019). DOI: 10.1145/3287560.3287586.
28. Prates, Marcelo OR, Pedro H. Avelar, and Luís C. Lamb. "Assessing gender bias in machine translation: a case study with Google Translate." Neural Computing and Applications 32.10 (2020): 6363-6381. DOI: 10.1007/s00521-019-04144-6.
29. Fazelpour, Sina, and Zachary C. Lipton. "Algorithmic Fairness from a Non-ideal Perspective." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020). DOI: 10.1145/3375627.3375828.
30. Hutchinson, Ben, and Margaret Mitchell. "50 years of test (un) fairness: Lessons for machine learning." Proceedings of the Conference on Fairness, Accountability, and Transparency (2019). DOI: 10.1145/3287560.3287600.
31. Font, Joel Escudé, and Marta R. Costa-Jussa. "Equalizing gender biases in neural machine translation with word embeddings techniques." arXiv preprint arXiv:1901.03116 (2019). Association for Computational Linguistics (ACL). DOI: 10.48550/arXiv.1901.03116.
32. Aggarwal, Aniya, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. "Black box fairness testing of machine learning models." Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (2019). DOI: 10.1145/3338906.3338937.
33. Jones, Gareth P., James M. Hickey, Pietro G. Di Stefano, Charanpal Dhanjal, Laura C. Stoddart, and Vlasios Vasileiou. "Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms." arXiv preprint arXiv:2010.03986 (2020). DOI: 10.48550/arXiv.2010.03986.
34. [34] Krishna, Satyapriya, Rahul Gupta, Apurv Verma, Jwala Dhamala. "Measuring Fairness of Text Classifiers via Prediction Sensitivity." arXiv preprint arXiv:2203.08670 (2022). Association for Computational Linguistics. DOI: 10.48550/arXiv.2203.08670.
35. Oneto, Luca, and Silvia Chiappa. "Fairness in machine learning." Recent Trends in Learning from Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019). Springer International Publishing (2020). DOI: 10.1007/978-3-030-43883-8_7.
36. Varley, Michael, and Vaishak Belle. "Fairness in machine learning with tractable models." Knowledge-Based Systems 215 (2021): 106715. DOI: 10.1016/j.knosys.2020.106715.
37. Valdivia, Ana, Javier Sánchez-Monedero, and Jorge Casillas. "How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness." International Journal of Intelligent Systems 36.4 (2021): 1619-1643. DOI: 10.1002/int.22354.
38. Friedler, Sorelle A., et al. "A comparative study of fairness-enhancing interventions in machine learning." Proceedings of the Conference on Fairness, Accountability, and Transparency (2019). DOI: 10.1145/3287560.3287589.
39. [39] Ferrari, Elisa, and Davide Bacciu. "Addressing Fairness, Bias and Class Imbalance in Machine Learning: the FBI-loss." arXiv preprint arXiv:2105.06345 (2021). DOI: 10.48550/arXiv.2105.06345.
40. Du, Mengnan, et al. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems 36.4 (2020): 25-34. DOI: 10.1109/MIS.2020.2995513.
41. Huang, Lingxiao, and Nisheeth Vishnoi. "Stable and fair classification." International Conference on Machine Learning. PMLR (2019). DOI: 10.48550/arXiv.1905.10874.

 doi:10.20944/preprints202604.0038.v1

42. Celis, L. Elisa, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. "Fair classification with noisy protected attributes: A framework with provable guarantees." International Conference on Machine Learning. PMLR (2021). DOI: 10.48550/arXiv.2102.12594.
43. Goel, Naman, Mohammad Yaghini, and Boi Faltings. "Non-discriminatory machine learning through convex fairness criteria." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. (2018). DOI: 10.1609/aaai.v32i1.11333.
44. Shrestha, Yash Raj, and Yongjie Yang. "Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems." Algorithms 12.9 (2019): 199. DOI: 10.3390/a12090199.
45. Mandal, Debmalya, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J. Hsu. "Ensuring fairness beyond the training data." Advances in Neural Information Processing Systems 33 (2020): NeurIPS Proceedings 18445-18456. DOI: 10.48550/arXiv.2010.06191.
46. Gao, Xuanqi, Juan Zhai, Shiqing Ma, Chao Shen, Yufei Chen, and Qian Wang. "FairNeuron: improving deep neural network fairness with adversary games on selective neurons." Proceedings of the 44th International Conference on Software Engineering (2022). DOI: 10.48550/arXiv.2204.02567.
47. Tizpaz-Niari, Saeid, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. "Fairness-aware configuration of machine learning libraries." Proceedings of the 44th International Conference on Software Engineering (2022). DOI: 10.48550/arXiv.2202.06196.
48. Chakraborty, Joymallya, Suvodeep Majumder, and Huy Tu. "Fair-SSL: Building fair ML Software with less data." Proceedings of the 2nd International Workshop on Equitable Data and Technology (2022). DOI: 10.48550/arXiv.2111.02038.
49. Patel, Ankita Ramjibhai, Jaganmohan Chandrasekaran, Yu Lei, Raghu N. Kacker, and D. Richard Kuhn. "A combinatorial approach to fairness testing of machine learning models." 2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE (2022). DOI: 10.1109/ICSTW55395.2022.00023.
50. Chen, Zhenpeng, Jie M. Zhang, Federica Sarro, and Mark Harman. "MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software." Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (2022). DOI: 10.1145/3540250.3549138.
51. Li, Yanhui, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. "Training data debugging for the fairness of machine learning software." Proceedings of the 44th International Conference on Software Engineering (2022). DOI: 10.1145/3510003.3510202.
52. Fan, Ming, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. "Explanation-guided fairness testing through genetic algorithm." Proceedings of the 44th International Conference on Software Engineering (2022). DOI: 10.1145/3510003.3510203.
53. Pu, Muxin, Meng Yi Kuan, Nyee Thoang Lim, Chun Yong Chong, and Mei Kuan Lim. "Fairness evaluation in deepfake detection models using metamorphic testing." 2022 IEEE/ACM 7th International Workshop on Metamorphic Testing (MET). IEEE (2022). DOI: 10.1145/3510003.3510204.
54. Zheng, Haibin, Zhiqing Chen, Tianyu Du, Xuhong Zhang, Yao Cheng, Shouling Ji, Jingyi Wang, Yue Yu, and Jinyin Chen. "Neuronfair: Interpretable white-box fairness testing through biased neuron identification." Proceedings of the 44th International Conference on Software Engineering (2022). DOI: 10.1145/3510003.3510205.
55. Zhang, Peixin, Jingyi Wang, Jun Sun, Xinyu Wang, Guoliang Dong, Xingen Wang, Ting Dai, and Jin Song Dong. "Automatic fairness testing of neural classifiers through adversarial sampling." IEEE Transactions on Software Engineering 48.9 (2021): 3593-3612. DOI: 10.1109/TSE.2021.3059472.
56. Zhang, Peixin, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. "White-box fairness testing through adversarial sampling." Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (2020). DOI: 10.1145/3377811.3380417.
57. Fabris, Alessandro, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. "Algorithmic fairness datasets: the story so far." Data Mining and Knowledge Discovery 36.6 (2022): 2074-2152. DOI: 10.1007/s10618-022-00829-7.

 doi:10.20944/preprints202604.0038.v1

58. Zhang, Jiehuang, Ying Shu, and Han Yu. "Fairness in design: A framework for facilitating ethical artificial intelligence designs." International Journal of Crowd Science 7.1 (2023): 32-39. DOI: 10.1108/IJCS-12-2022-0023.
59. Majumder, Suvodeep, Joymallya Chakraborty, Gina R. Bai, Kathryn T. Stolee, and Tim Menzie. "Fair enough: Searching for sufficient measures of fairness." ACM Transactions on Software Engineering and Methodology 32.6 (2023): 1-22. DOI: 10.1145/3571234.
60. Wang, Zichong, Yang Zhou, Meikang Qiu, Israat Haque, Laura Brown, Yi He, Jianwu Wang, David Lo, and Wenbin Zhang. "Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking." arXiv preprint arXiv:2302.08018 (2023). DOI: 10.48550/arXiv.2302.08018.
61. Guo, Huizhong, Jinfeng Li, Jingyi Wang, Xiangyu Liu, Dongxia Wang, Zehong Hu, Rong Zhang, and Hui Xue. "Fairrec: fairness testing for deep recommender systems." Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (2023). DOI: 10.1145/3597926.3597937.
62. Hort, M., Zhang, J. M., Sarro, F., & Harman, M. "Search-based automatic repair for fairness and accuracy in decision-making software." Empirical Software Engineering 29.1 (2024): 36. DOI: 10.1007/s10664-023-10123-4.
63. Bantilan, Niels. "Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation." Journal of Technology in Human Services 36.1 (2018): 15-30. DOI: 10.1080/15228835.2018.1428994.
64. Ling, Jiasheng, Lei Zhang, Chenyang Liu, Guoxin Xia, and Zhenxiong Zhang. "Machine Learning-Based Multilevel Intrusion Detection Approach." Electronics 14, no. 2 (2025): 323. DOI: 10.3390/electronics14020323.
65. Nasiri, Roya. "Testing Individual Fairness in Graph Neural Networks." arXiv preprint arXiv:2504.18353 (2025). DOI: 10.48550/arXiv.2504.18353.
66. Consuegra-Ayala, Juan Pablo, Yoan Gutiérrez, Yudivian Almeida-Cruz, and Manuel Palomar. "Bias mitigation for fair automation of classification tasks." Expert Systems 42, no. 2 (2025): e13734. DOI: DOI: 10.1111/exsy.13734.
67. Xu, Paiheng, Yuhang Zhou, Bang An, Wei Ai, and Furong Huang. "Gfairhint: Improving individual fairness for graph neural networks via fairness hint." ACM Transactions on Knowledge Discovery from Data 19, no. 3 (2025): 1-22. DOI: 10.1145/3714472
68. Foidl, Harald, and Michael Felderer. "Risk-based data validation in machine learning-based software systems." Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation (2019). DOI: 10.1145/3340482.3342743.
69. Riccio, Vincenzo, Gunel Jahangirova, Andrea Stocco, Nargiz Humbatova, Michael Weiss, and Paolo Tonella. "Testing machine learning based systems: a systematic mapping." Empirical Software Engineering 25.6 (2020): 5193-5254. DOI: 10.1007/s10664-020-09881-0.
70. Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning." ACM Computing Surveys (CSUR) 54.6 (2021): 1-35. DOI: 10.1145/3457607.
71. Alkatheri, Maha Saleh. "Testing Machine-Learning-based Software for Fairness: Ontology-Guided Test Cases Generation" (2022). DOI: 10.48550/arXiv.2205.00210.
72. Le Quy, Tuan, Abir Roy, Vasileios Iosifidis, Wei Zhang, and Eirini Ntoutsi. "A survey on datasets for fairness-aware machine learning." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2022). DOI: 10.1002/widm.1443.
73. Caton, Simon, and Christian Haas. "Fairness in machine learning: A survey." arXiv preprint arXiv:2010.04053 (2020). DOI: 10.48550/arXiv.2010.04053.
74. Richardson, Brianna, and Juan E. Gilbert. "A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions." arXiv preprint arXiv:2112.05700 (2021). DOI: 10.48550/arXiv.2112.05700.
75. Keele, Staffs. Guidelines for performing systematic literature reviews in software engineering. Vol. 5. Technical report, Ver. 2.3 EBSE Technical Report. EBSE (2007). DOI: 10.14236/ewic/EASE2008.8.
76. Pessach, Dana, and Erez Shmueli. "A review on fairness in machine learning." ACM Computing Surveys (CSUR) 55.3 (2022): 1-44. DOI: 10.1145/3487047.

 doi:10.20944/preprints202604.0038.v1

77. Kohavi, Ronny, and Barry Becker. "Adult data set." UCI Machine Learning Repository 5 (1996): 2093. DOI: 10.24432/C5NC77.
78. Hofmann, Hans. "Statlog (German credit data)." UCI Machine Learning Repository 10 (1994): C5NC77. DOI: 10.24432/C5NC77.
79. Ofer, Dan. "ProPublica (COMPAS)." Kaggle (2016). DOI: 10.34740/KAGGLE/DSV/100000.
80. Jui, Tonni Das, and Pablo Rivas. "Fairness issues, current approaches, and challenges in machine learning models." International Journal of Machine Learning and Cybernetics (2024): 1-31. DOI: 10.1007/s13042-023-01523-4.
81. Chen, Zhenpeng, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. "Fairness testing: A comprehensive survey and analysis of trends." ACM Transactions on Software Engineering and Methodology 33.5 (2024): 1–59. DOI: 10.1145/3571234.
82. Petersen, Kai, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. "Systematic mapping studies in software engineering." in 12th International Conference on Evaluation and Assessment in Software Engineering (EASE). BCS Learning & Development, 2008. DOI: 10.14236/ewic/EASE2008.8.
83. Ahuja, Ravinder, Aakarsha Chug, Shaurya Gupta, Pratyush Ahuja, and Shruti Kohli. "Classification and clustering algorithms of machine learning with their applications." in Nature-inspired computation in data mining and machine learning, pp. 225-248. Cham: Springer International Publishing, 2019.DOI: 10.1007/978-3-030-28553-1_11