

Article

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

Roa Alharbi^{1,2,*}  and Nouredine Abbadeni¹ 

¹ Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; nabbadeni@ksu.edu.sa

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 1805 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 in conference proceedings, accounting for 52% of the primary studies, followed by journals and workshop proceedings, which accounted for 42% and 6% of the primary studies. The literature encompasses multiple research themes, with 36% of the primary studies focusing on the analysis of existing fairness methods, 22% addressing bias mitigation strategies, 30% investigating testing techniques, and 12% proposing or evaluating evaluation frameworks. Fairness testing was conducted across multiple testing levels, including unit, integration, and system testing. Integration-level testing was the most prevalent, accounting for approximately 37.9% of the studies, followed by system-level testing at 27.3% and unit-level testing at 12.1%. Additionally, 22.7% of the studies applied fairness testing across more than one testing level. Frequently used datasets included COMPAS, Adult Census Income, and German Credit. Widely adopted tools, such as IBM AI Fairness 360, Themis, and Aequitas, were also identified. Overall, the systematic mapping study (SMS) highlights the progress made in fairness research while emphasizing the need for stronger integration of fairness into practical machine learning development.



Academic Editor: Hongyu Zhang

Received: 16 March 2026

Revised: 13 April 2026

Accepted: 22 April 2026

Published: 27 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

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

1. Introduction

Machine learning-based systems (MLSs) 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 MLSs as critical decision-support instruments in high-stakes contexts. Because MLS outputs can significantly affect people and groups, ensuring that predictions and recommendations are credible, as well as ensuring the transparency of the process and outcome, fairness, and equitable treatment, is essential for societal trust and responsible/ethical deployment. When MLSs generate biased outcomes, it can lead to social injustices by amplifying historical inequities and perpetuating discrimination in critical domains, such as hiring, lending, and medical prioritization, where procedural fairness and substantive outcomes are crucial.

Machine learning itself can be broadly categorized into supervised, unsupervised, semi-supervised, and reinforcement learning methods, each with distinctive ways of learning from data and making predictions [1]. Furthermore, systematic approaches for conducting structured reviews in software engineering and MLS research have been established and guide studies, such as the approach published by the authors of [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 definition stated that fairness is achieved when protected attributes (e.g., race, gender, and 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; rather, metrics need to be considered and understood in relation to the application context and the kinds of harm identified. 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 [6].

Fairness challenges can manifest at all stages of the MLS lifecycle. In the pre-processing 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 required, adjusted to avoid disparate impacts across groups [7]. Fairness testing approaches are classified according to the ML lifecycle stage at which the tested artifact (data, model internals, or outputs) is modified or assessed, rather than where the fairness intervention is applied [8]. Separating 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 an ingrained practice from start to finish [9].

While fairness-related skills and tools have become increasingly visible in contemporary AI research and practice, the literature continues to face fundamental challenges. Prior surveys and reviews have provided valuable insights into fairness metrics, bias mitigation strategies, and algorithmic perspectives on discrimination, helping to consolidate foundational knowledge in the field [8,10,11]. More recent surveys further extend this understanding to complex deep learning and foundation models, including large language models, by organizing fairness evaluation methods across embeddings, model outputs, and generated content [11]. 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, and 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 type. 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 publications. Researchers in these communities often apply different fairness definitions and evaluation protocols, complicating cross-study synthesis. The commonly used datasets and benchmarks, each with known limitations, further motivate 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 study emphasizes breadth with structure: it documents where research is published (publication/year), 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 evidence. By consolidating these elements into a single, organized account, this study provides a platform for understanding the field’s current state and for charting principled directions for future work [12].

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, such reviews 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 publication types and years within a single, unified framework. Moreover, earlier reviews do not 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 this section of the study deliberately avoids detailed tutorials on ML algorithms, data modalities, and deep learning architectures—topics that are covered comprehensively in Section 3—the present section maintains the essential foundations required to motivate a fairness-centered SMS. Specifically, this section 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 [7], 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 discussion in Section 3, so that the rest of the study can maintain a focused review of the societal context, the fairness problem, the literature limitations, and this 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 MLSSs. 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 sources of published studies in the field of unfairness detection in MLSSs.
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 MLSSs.
5. Provide researchers with datasets, algorithms, and models for detecting unfairness in MLSSs 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: related work, background, research methodology, threats to validity, results and discussion, conclusions, and future work.

2. Related Work

Several previously conducted surveys and reviews have studied fairness in machine learning systems (MLSSs) from different perspectives. Caton et al. provided a comprehensive overview of approaches to reduce bias and categorized them into pre-processing, in-processing, and post-processing approaches [13]. 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 as a guide for researchers [8].

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 [14]. Other reviews, e.g., Richardson et al. [15], 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 [16] using the German Credit dataset to study gender biases. Chen et al. [11] conducted a bibliometric review of fairness testing up to 2023 by looking at testing workflows, such as test input generation and the identification of an oracle. Their study defined testing taxonomy at three levels (data, program, and model levels), 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), datasets, or testing workflows, without synthesizing the definitions, problem types, testing pathways, datasets, and emerging research gaps into a single unified framework. Furthermore, most of the reviews concluded prior to 2025, and thus did not afford consideration of the advances in deep learning fairness, causal inference methods, and interdisciplinary perspectives in their analyses. Table 1 compares our study with related systematic studies regarding their timeframes, number of primary studies, addressed research questions, and covered software engineering topics.

Table 1. Related studies timeframes, number of primary studies, research questions and topics.

Study	Timeframe	PS	Research Questions	SE Topics
Software Unfairness Detection in Machine Learning-based Systems (SMSs)	2010–2025	60–70	What is the distribution of studies through publication types and years? What are the different definitions of software fairness? What types of problems are addressed? What different approaches to fairness testing are presented? Which datasets are used to detect the unfairness of MLs? What are the research gaps and trends discovered in the reviewed studies?	Software testing, machine learning, and unfairness detection
Mehrabi et al. [8]	2013–2020	100–120	What are the major sources of bias in machine learning systems (data, algorithms, and human processes)? How are fairness notions defined and categorized across disciplines? What mitigation techniques exist at different stages of the ML pipeline? What open challenges remain for fair ML research and deployment?	Data collection, model training
Caton et al. [13]	2008–2019	60–70	How is fairness defined across disciplines (computer science, law, and ethics)? What are the main mathematical fairness notions and trade-offs? How do fairness concepts differ between group and individual perspectives?	Conceptual and model levels
Richardson & Gilbert [15]	2015–2020	80–90	What technical and organizational fairness solutions exist? How can fairness interventions be categorized into a unifying framework? Which stages of AI systems are addressed by current fairness solutions?	Model training and post-processing
Chen et al. [11]	2010–2023	220–230	What fairness definitions and metrics are used in ML testing? What testing techniques are applied to detect fairness issues? How are fairness testing approaches mapped to ML system artifacts and execution stages? What trends and open challenges exist in fairness testing research?	Testing and evaluation

Our research directly tackles this gap through a systematic mapping study (SMS) of software unfairness detection in MLs, covering all the relevant literature from 2010 to 2025. Unlike previous review studies that 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 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 MLSs.

The following section provides the background necessary to understand the scope and significance of this 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 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, and supervised learning is commonly applied in classification and regression tasks [6]. Unsupervised learning analyzes a dataset with only inputs to find hidden structures or patterns, where the algorithm clusters or groups data based on feature similarities and is different from supervised methods in that there is no labeled data; examples are clustering and association, and a type of unsupervised learning is self-supervised learning, as the model creates the labels from the input data to serve as the labels for learning, for example, self-supervised learning is common in natural language processing and computer vision [17].

Semi-supervised 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 [6]. Reinforcement learning uses the interaction between the model and the environment, where the model acts to maximize cumulative feedback or rewards, and reinforcement learning is popular in robotics, gaming, and navigation [17]. Neural networks (NNs) are inspired by neurons in the human brain, where one neuron passes on information to another neuron, and each neuron has inputs to process before passing information to the next neuron; the same is true for an NN, where it receives input through the input layer, processes under one or more hidden layers, and receives output from the output layer [18].

Deep neural networks (DNNs) have multiple layers of connected nodes that build upon the previous layer and the neurons to create a more refined prediction, where the process of passing information forward through the layers is called forward propagation and the process of moving through the network in reverse is called backpropagation, which gives weight to the neurons for the network so that estimates are minimized to the errors; a DNN uses both tasks to learn to estimate accurately and correct the errors, and 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” [19]. In traditional ML algorithms, the features used for making a prediction are selected by the researcher or user, whereas for DNN algorithms, the features used as input for prediction are learned automatically by the DNN algorithm from the datasets during the training process, and with DNNs, the dataset input into the DNN algorithm can be a large amount of data; as the DNN is processing the data, it can learn the features for heterogeneous or homogeneous data that are relevant to the data provided, and then the DNN can estimate 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 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 [19].

Recurrent neural networks (RNNs), including their most well-known variant, the Long Short-Term Memory (LSTM) model, 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 that utilize 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 [19,20].

More recently, transformer architecture has become the leading standard in NLP and related fields. Transformers leverage self-attention mechanisms and are thus better able to represent long-range dependencies compared with 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 are 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 ML applications, they are also often concentric focuses of fairness detection and mitigation research, which situates technical advancement in the design of architectures in direct association with ethical considerations in machine learning [21,22].

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 (SVMs) [15]. In regression problems, the goal is to predict one or more continuous values. The algorithms used for regression include Linear Regression and Multi-Layer Perceptron [23]. 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 continuous value [17]. Sequence Prediction is when a system predicts (or forecasts) one or more future values based on past data it has observed. This method is practically used in natural language processing, bioinformatics, time series, and forecasting [24].

Descriptive Tasks (Clustering, Association, and Anomaly Detection): Clustering consists of grouping items into clusters in which high similarity exists between objects (or instances) in the same cluster and low similarity exists 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 [11]. In addition, clustering is an area where neural networks (NNs) may be applied, and has received more prominence in recent studies.

The association task focuses on discovering relationships between variables within a dataset. Association is widely used in association-based systems, such as recommendation systems. Algorithms commonly used with association include Apriori, Eclat, and FP-Growth [24].

Anomaly detection consists of 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 and an error depends on context, for example, fraud detection in banking, intrusion detection in cybersecurity, etc. [25].

3.4. Types of Datasets

In the context of unfairness detection research, datasets are discussed not by their general structure, but by their role in revealing, measuring, and testing bias. Most fairness studies utilize structured tabular datasets containing sensitive attributes (e.g., gender, race, and age) and decision outcomes. These datasets enable the computation of group and individual fairness metrics. The most commonly used datasets to evaluate bias in classification and regression settings are COMPAS, Adult Census Income, German Credit, and datasets from the UCI Machine Learning Repository.

Unstructured and semi-structured datasets have gained increasing importance with the rise of deep learning and foundation models. In natural language processing, large-scale text corpora, such as Common Crawl, Wikipedia, Reddit, and social media datasets, are frequently analyzed for representational and allocational harms, particularly in the study of stereotyping, toxicity, and demographic bias.

Across both types, prior work demonstrates that bias often arises from imbalanced group representation, historical and societal biases embedded in data sources, labeling practices, and feedback loops introduced during data collection and deployment. As a result, these datasets serve as a tool for fairness evaluation, mitigation, and auditing frameworks.

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, such as healthcare, banking, law, and e-commerce. Such systems use algorithmic 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) [3]. Researchers are focusing on applying techniques and tools to measure fairness in MLSs. Numerous algorithms and technologies have been proposed to evaluate fairness in MLSs using specific metrics [11].

3.6. Testing Fairness in MLSs

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 MLSs [25].

There are many fairness metrics that are commonly applied in the evaluation of fairness in MLSs, including disparate impact, independence, demographic parity, equalized odds, fairness through awareness, and positive and negative class balance. Each metric has its own formal conditions and definitions that must be satisfied in MLS predictions [5].

3.7. Fairness and Bias in MLSSs

There are multiple sources of bias in MLSSs, and they can be summarized into three categories. They are Data to Algorithm, Algorithm to User, and User to Data [26].

In the Data to Algorithm category, bias is derived from data that was not measured or reported correctly. Additionally, errors in data representation or incorrect conclusions based on errors in 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 the User to Data bias category, any data source used to train MLSSs is potentially biased due to being user generated. Thus, any bias in user behavior or thought processes may be reflected in the data generated [8].

Types of bias in MLSSs can be summarized in the points [27] shown in Figure 1.

Historical bias	Representation bias	Measurement bias	Aggregation bias	Evaluation bias	Deployment bias
• Misalignment between old and new datasets.	• Failure in generalization of population.	• Prejudice in selecting features.	• Appears when combining unfair subgroups.	• Prejudice in evaluating features	• Wrong usage of the system after model deployment

Figure 1. Overview of common biases in machine learning, including historical, representation, measurement, aggregation, evaluation, and deployment biases, along with examples of how each bias can arise.

Fairness evaluation in MLSSs is concerned with bias identification and mitigation along various stages of the model development process. The evaluation of fairness metrics in pre-processing stages focuses on the quality and balance of input data, measuring variables, such as statistical parity, disparate impact, or representation bias, 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 regularizing the objective function. Finally, the post-processing stage of fairness evaluation 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.

4. Research Methodology

4.1. Methods Overview

This section provides an overview of the research methodology. Following the guidelines [28,29], our process was conducted as follows: The first step was to define the research questions and determine the scope (see Table 2). The second step involved searching designated digital libraries using a carefully constructed search string. All collected studies were then screened by applying inclusion and exclusion criteria. Afterward, a quality assessment was performed to categorize the collected studies. Next, relevant studies were grouped, and keywords were classified for each study. Finally, data extraction and a structured literature

review were carried out to complete the systematic mapping study. Figure 2 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 [28]. These were formulated from the following five perspectives:

- Population: Machine learning and AI systems, along with their models, datasets, and fairness-related artifacts, algorithmic bias, discrimination, or unfair decision-making behavior.
- Intervention: Software engineering-based approaches for unfairness detection, including testing, automated analysis, search-based techniques, and tool-supported evaluation of fairness in AI systems.
- Comparison: Research publication statistics, fairness approaches, fairness testing levels, evaluation metrics, solutions, biased algorithms, and datasets.
- Outcomes: Software fairness definitions and techniques used in MLSs, fairness testing levels, biased algorithms and datasets, solutions for mitigating unfairness, and identification of research gaps.
- Context: Research content relevant to academia and industry.

Table 2. Research questions.

Id	Research Questions	Rationale
RQ1	What is the distribution of studies through publication types and years?	Identify the kind of seminars for collected studies, published journals, conference proceedings, 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 to fairness testing are presented?	Describe different approaches used in fairness testing solutions, such as algorithms or tools, and explore fairness testing levels in MLSs.
RQ5	Which datasets are used to detect the unfairness of MLSs?	Identify some datasets/algorithms/models that are used to detect unfairness in MLSs 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.

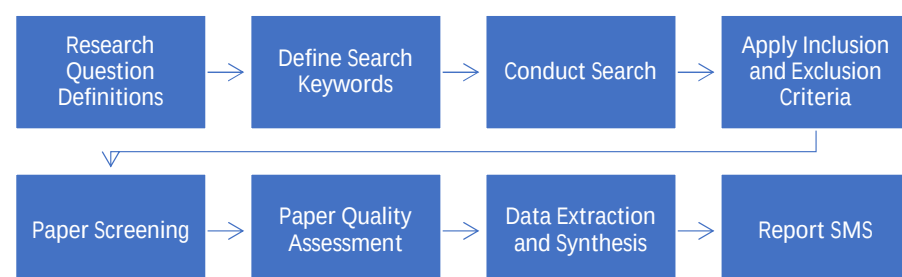


Figure 2. The methodological process followed in this study. The figure illustrates the sequential steps for performing a systematic mapping study (SMS), from initial research question formulation to final reporting.

4.2. Search Strategy

To collect primary studies, we searched across several digital libraries and categorized the databases into main and supplementary sources (see Table 3).

Table 3. Digital libraries.

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

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 [21].

The following search string is used for searching in the following designated search engines and digital libraries:

Software: “fairness OR unbiased OR biased OR Non-Discriminatory”.

Machine Learning: “Machine learning OR deep learning OR neural network OR data mining OR mining software repositories”.

4.3. Narrative for Study Screening and Selection

The initial search across all selected databases retrieved a total of 1805 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 following inclusion criteria were then applied:

- (a) Published in academic journals, conference proceedings, or magazines;
- (b) Written in English;
- (c) Full-text accessible;
- (d) Published between 2010 and 2025.

Figure 3 shows that after applying the search filters, the pool was reduced from 1805 to 1370 studies.

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

- (a) Duplicate records not previously detected;
- (b) Studies on irrelevant topics;
- (c) Non-published works.

Authors typically refine the study selection process by removing undetected duplicate records, excluding studies irrelevant to the research scope through title and abstract screening, and discarding non-published or non-peer-reviewed works. These steps ensure that only unique, relevant, and scientifically validated studies are included, thereby improving the reliability and rigor of the final study set.

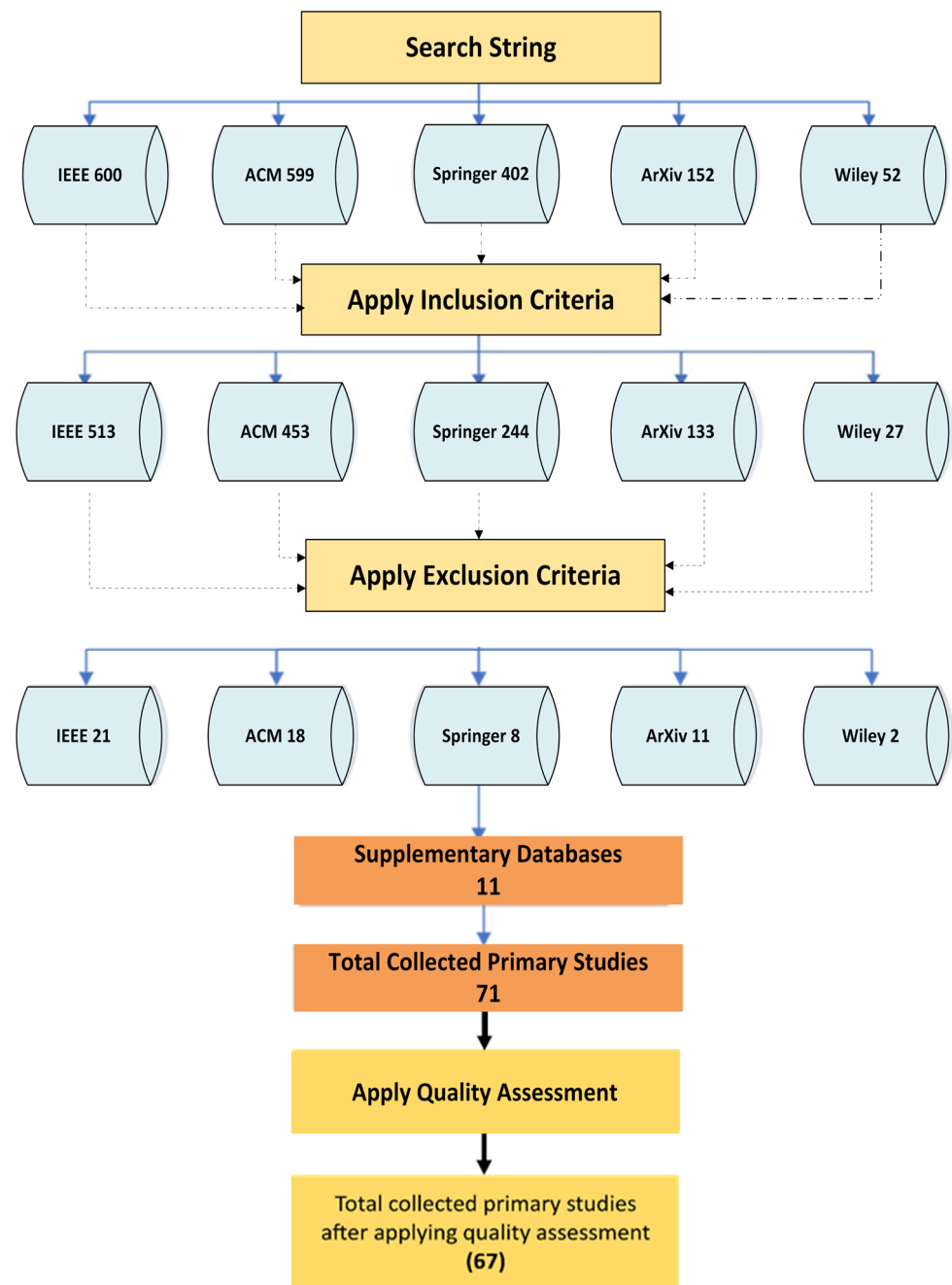


Figure 3. Overview of the study selection process, illustrating the number of studies retrieved from each digital library, followed by the application of the inclusion and exclusion criteria, the supplementary database search, and the quality assessment, leading to the final set of primary studies.

The third phase involved full-text screening of the remaining articles. Studies 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. Manual checks were performed to ensure that both studies published in conference proceedings and their extended journal versions were appropriately retained. After this step, the number of studies was further reduced from 1370 to 61, and with the addition of the primary studies from the supplementary databases, we collected a total of 71 primary studies.

Finally, a quality assessment (QA) was applied to evaluate the literature using the predefined QA checklist. Each study 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? Each study was reviewed manually, and scores were assigned as follows: Yes = 1, Partly = 0.5, or No = 0.

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 all these steps were completed, 67 primary studies remained and were included in the final synthesis of this systematic mapping study (see Appendix A).

4.4. Data Extraction

In the data extraction stage, relevant information was collected from each selected study to answer the research questions. Data extraction was performed independently by two authors using a predefined extraction form. An Excel file was used to systematically organize the extracted information and generate charts. Keywords were defined for each research question (Table 4), and synonyms were used when necessary to avoid missing relevant information. After extraction, the results from both authors were compared to identify discrepancies. Any disagreements were discussed between the two authors to reach a consensus. This procedure helped ensure the accuracy, validity, and reliability of the extracted data. The extracted findings were synthesized and are presented in Section 5.

Table 4. Search keywords.

Search Keywords	RQs	Possible Values
Journals, Conferences, Year	RQ1	Journals, conference proceedings, university names and types, and 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, and algorithms.
Research Gaps	RQ6	Trends, future work, and research gaps.

4.5. Threats to Validity

One risk that presents a threat to validity is that researchers working in this area of fairness might use different terminology, which could lead to important studies being overlooked. Therefore, the search string was constructed with multiple 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, the keywords were separated and tested individually. This strategy minimized the risk of incomplete search results and increased consistency in the dataset collection process.

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

5. Results and Discussion

RQ1: What is the distribution of studies through publication types and years?

The distribution of the 67 primary studies across publication types is shown in Tables 5–7. Conference proceedings were the most common outlet (52%, $n = 35$), fol-

lowed by journals (42%, $n = 28$) and workshop proceedings (6%, $n = 4$) (Figure 4). This indicates that over half of the studies were published in conference proceedings, while a substantial proportion also appeared in journals. Conference papers often present early or exploratory work with lighter methodological rigor, while journal articles typically report more mature studies with higher methodological quality. As a result, journals generally achieve greater long-term citation impact than conference publications.

Table 5. Primary studies published in journals.

Publication	Study ID	Publisher
Journal of Computing Sciences in Colleges	PS5	ACM
Journal of Artificial Intelligence Research	PS6	AI Access Foundation
International Journal of Intelligent Systems	PS7	Wiley
Journal of Machine Learning Research	PS11	JLMR
IEEE Transactions on Industrial Informatics	PS12	IEEE
Neural Computing and Applications	PS13, PS25	Springer
IEEE Intelligent Systems	PS18	IEEE
Data Science and Engineering	PS19	Springer
IEEE Transactions on Software Engineering	PS24	IEEE
Expert Systems	PS28	Wiley
Journal of Technology in Human Service	PS31	ArXiv
International Journal of Data Science and Analytics	PS35	Springer
Knowledge-Based Systems	PS37	ArXiv
International Journal of Crowd Science	PS39	ArXiv
Journal on Emerging Technologies in Computing Systems	PS43	ACM
DBLP CoRR journal	PS48, PS63	ArXiv
Journal of Data and Information Quality	PS49	ACM
Algorithms	PS55	MDPI
Advances in Neural Information Processing Systems	PS56	ArXiv
Data Mining and Knowledge Discovery	PS61	Springer
ACM Transactions on Software Engineering and Methodology	PS62	ACM
Electronics	PS65	MDPI
ACM Transactions on Knowledge Discovery from Data	PS67	ACM

Table 6. Primary studies published in conference/workshop proceedings.

Conference/Workshop	Study ID	Publisher
The World Wide Web Conference	PS2, PS44	ACM
IEEE/ACM International Workshop on Software Fairness	PS3	IEEE/ACM
IEEE TrustCom 2020	PS9	IEEE
IEEE International Symposium on Technology and Society	PS20	IEEE
IEEE Conference on Decision and Control	PS10	IEEE
ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering	PS8	ACM
Proceedings of the Conference on Fairness, Accountability, and Transparency	PS14, PS32, PS50	ACM
International Conference on Testing Software and Systems	PS15	Springer
International Conference on the Quality of Information and Communications Technology	PS16, PS34	Springer
INNS Big Data and Deep Learning Conference	PS17	Springer
European Software Engineering Conference and Symposium on the Foundations of Software Engineering	PS21	ACM
International Workshop on Equitable Data and Technology	PS26	ACM
ACM Conference (Conference '17)	PS27	ACM
Proceedings of the 23rd ACM SIGKDD	PS29	ACM

Table 6. Cont.

Conference/Workshop	Study ID	Publisher
International Conference on Software Engineering	PS30, PS40, PS41, PS57, PS58, PS59	IEEE/ACM
AAAI Conference on Artificial Intelligence	PS33	AAAI
Companion Proceedings of the Web Conference 2021	PS36	ACM
IEEE International Conference on Software Testing, Verification, and Validation Workshop	PS38	IEEE
ACM SIGSOFT International Symposium on Software Testing and Analysis	PS42	ACM
Proceedings of the 36th International Conference on Machine Learning	PS45, PS54	PMLR
International Conference on Computer-Aided Verification	PS45	Springer
International Conference on Computer, Control, and Communication	PS46	IEEE
Joint European Conference on Machine Learning and Knowledge Discovery in Databases	PS47	Springer
Proceedings of the AAAI/ACM Conference on AI	PS51	ACM
ACL Workshop on Gender Bias for Natural Language Processing	PS52	ACL
Annual Meeting of the Association for Computational Linguistics	PS53	ACL
IEEE/ACM 7th International Workshop on Metamorphic Testing	PS60	IEEE/ACM
International Conference on Evaluation and Assessment in Software Engineering	PS66	EASE

Table 7. Primary studies published as empirical studies.

Publication	Study ID	Publisher
Empirical Software Engineering	PS22, PS64	Springer

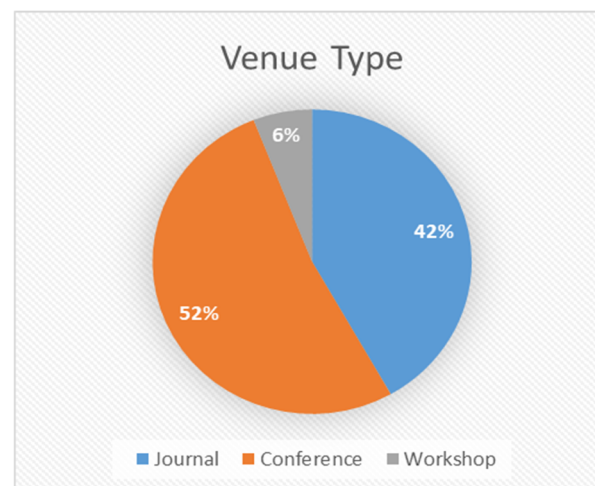


Figure 4. Distribution of primary studies by publication type, showing that 52% were published in conference proceedings, 42% in journals, and 6% in workshop proceedings.

Figure 5 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%). This sharp rise reflects the growing attention to fairness as an urgent challenge in MLs.

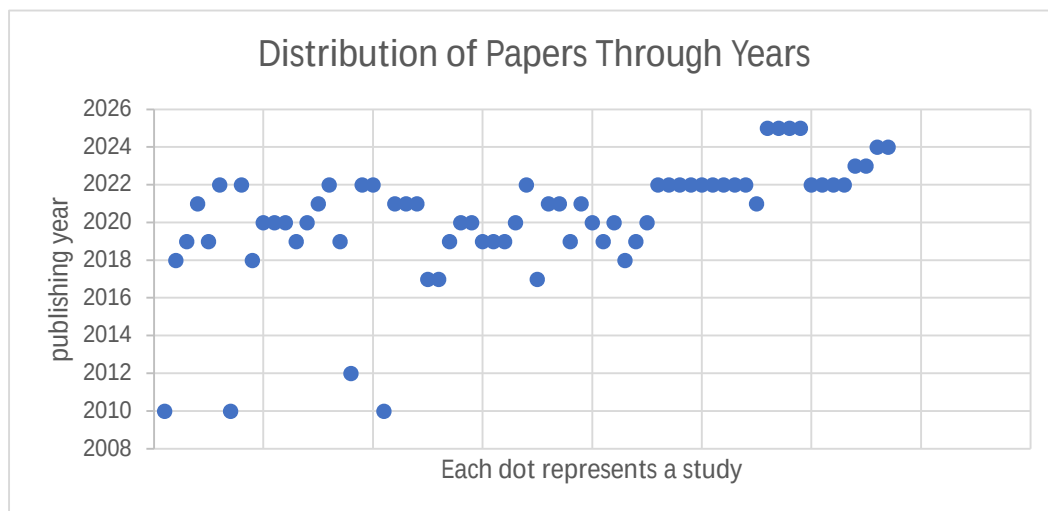


Figure 5. Distribution of the primary studies across publication years, illustrating the increasing research interest over time, with studies ranging from 2009 to 2025.

The dominance of studies published in conference proceedings suggests that fairness detection in MLSs 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 in journals. The comparable proportions highlight the complementary roles of both publication types 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.

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 8, we categorized definitions into four large categories of fairness which were general fairness, individual fairness, group fairness, and fairness metrics.

Table 8. 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.	PS28
	Fairness-aware Model	A model that avoids discrimination and promotes fairness.	PS28
	Fairness Degree	Max difference in predictions for pairs differing only in sensitive attributes.	PS22
	Fairness Through Unawareness	Fairness by excluding sensitive attributes from decision-making.	PS14
	Counterfactual Fairness	Prediction remains unchanged if the individual belongs to a different group.	PS14
	Algorithmic Bias	Bias from mathematical rules favoring certain attributes.	PS31
Individual Fairness	Individual Fairness	Similar individuals should receive similar outcomes.	PS10, PS19, PS49, PS67
	Individual Discrimination	Discrimination between individuals differs only in protected attributes.	PS9, PS21
Group Fairness	Group Fairness	Equal outcomes across demographic groups (e.g., gender, race).	PS10, PS19, PS49
	Fairness Constraints	Constraints like demographic parity, equal opportunity, and disparate impact.	PS19, PS49
Fairness Metrics	Demographic Parity	Outcome is independent of the protected attribute.	PS29, PS34
	Equal Opportunity	Equal true positive rates across groups.	PS5, PS6, PS29,
	Predictive Parity	Equal positive predictive values across groups.	PS17, PS34
	Disparate Impact	Ratio of favorable outcomes between groups.	PS6, PS29
	Average Absolute Odds	Average difference in false/true positive rates across groups.	PS6, PS17
	Theil Index	Measures inequality in prediction outcomes.	PS6
	Calibration	Predicted score should reflect actual outcomes equally across groups.	PS14, PS17

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 S15 and PS29 studies use demographic parity and predictive parity [30,31], which ensure equal outcomes that are independent of the protected features. These are both equal outcomes that are independent of the protected features. Study PS19 referred to group fairness as fairness in output distribution across the sensitive groups [32].

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 PS21 defined unfairness as discrimination between individuals that differ only in protected features [33,34].

However, group fairness may conflict with individual fairness, which requires that similar individuals receive similar treatment. For example, in hiring systems, two equally qualified candidates may receive different decisions to satisfy demographic parity across groups. In credit scoring, enforcing group fairness can result in loan approvals for less creditworthy individuals from one group while rejecting more creditworthy individuals from another. These cases illustrate the inherent trade-off between achieving fairness at the group level and ensuring equitable treatment at the individual level.

A small number of studies (11 studies, 16%) proposed general fairness concepts, like fairness-aware models, degree of fairness and fairness through unawareness [23,35,36]. These concepts are often not precise categories but appear to be conceptual framing, mostly framed before researchers selected a metric-based evaluation of fairness.

Finally, in 10 studies (15%), researchers evaluated fairness by using operational definitions in terms of specific fairness metrics. The metrics included disparate impact, equal opportunity, mean absolute odds, and Theil index [4,7,27,31]. The other definitions included counterfactual fairness and defining unfairness as the inverse of algorithmic bias [23,37,38].

Overall, the results indicate that fairness in MLS studies is defined inconsistently in literature, with no definition dominating the field. The preponderance the group fairness 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 [23,32].

Group fairness is widely adopted because it is easier to formalize and evaluate using aggregate metrics such as demographic parity, equalized odds, and predictive parity computed over clearly defined sensitive groups (e.g., gender or race). These metrics can be directly measured from datasets and model outputs, making them practical for empirical evaluation and benchmarking. In contrast, individual fairness depends on defining similarity between individuals, which is highly domain-specific and difficult to operationalize. 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 PS41 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 multi-dimensional approach [4]. The PS17 study emphasized the probabilistic nature of fairness, while the PS30 study analyzed prediction difference in pairwise definition [27,37]. 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 6 and detailed in Table 9, these categories exemplify how researchers conceptualize and operate the fairness problem.

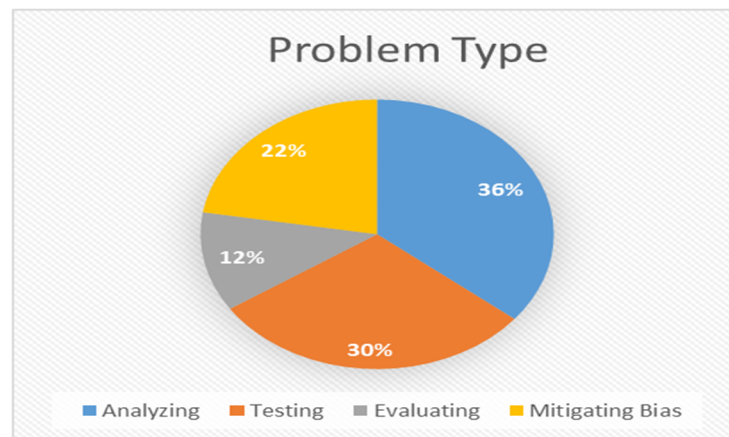


Figure 6. Distribution of primary studies by problem type, showing that 36% focus on analyzing bias, 30% on testing, 12% on evaluating, and 22% on mitigating bias.

Table 9. Distribution of problem types and number of studies.

Problem Type	Problem Description	Number of Studies	Study IDs
Analyzing	The process of systematically examining proposed unfairness detection methods to understand their effectiveness and limitations.	24	PS1, PS4, PS5, PS6, PS8, PS13, PS20, PS23, PS25, PS28, PS29, PS35, PS36, PS37, PS39, PS40, PS42, PS44, PS46, PS47, PS49, PS54, PS58, PS64
Mitigating Bias	The process of implementing strategies or algorithms that ensure fair outcomes and mitigate bias in unfairness detection systems.	15	PS7, PS11, PS12, PS19, PS26, PS33, PS34, PS43, PS50, PS52, PS53, PS56, PS65, PS60, PS67
Testing	The process of proposing testing solutions for checking whether a model produces fair outcomes and follows fairness metrics.	20	PS2, PS9, PS10, PS14, PS15, PS16, PS17, PS21, PS22, PS24, PS30, PS31, PS38, PS41, PS45, PS48, PS59, PS63, PS62, PS66
Evaluating	The process of comparing different approaches, analyzing outcomes, and validating results.	8	PS3, PS18, PS27, PS32, PS51, PS55, PS57, PS61

Analyzing methods was the most popular 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, PS44 optimized classification models by changing the fairness measures of the classifiers, and PS6 analyzed the impact of different mitigated methods on performance and fairness [5,7].

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) to check decision outcomes, and PS15 suggested a verification approach to develop fairness-based test cases [25,33].

Mitigating bias was the focus of 15 studies (22%). These contributions provided new methods and algorithms to reduce discrimination in MLs. For example, PS46 developed an alternative classification method and constructed a model that could develop unbiased models from biased training data, and PS19 proposed a fair outlier detection method that included more than one sensitive attribute (i.e., gender, race, religion, and nationality) [7,32].

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, PS3 summarized the various definitions of fairness and showed the definitions using case studies, while PS32 critically assessed the fairness criteria and mathematical techniques across diverse groups and individuals [3,39].

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 conference proceedings, 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 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 to 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 7. 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 within the scope of fairness within research.

Finally, the 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 [32,40]. There were also approaches to rank algorithms to minimize bias in recommendation systems, impute and resample data to create a balance in the dataset, and adjust word embeddings to minimize gender bias in natural language processing tasks [27,41,42].

Regarding levels of testing, input level testing was predominately carried out ($n = 25$, 38%); there were also 18 studies that provided examples of system level testing ($n = 18$, 27%) and eight 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, PS15 applied verification-based system testing, generating cases that examined accuracy and fairness [25]. At the unit level, PS35 performed dataset-focused testing to address sampling bias [42].

Input testing was frequently applied in comparative experiments, such as PS5, which tested sampling strategies across neural network classifiers [43].

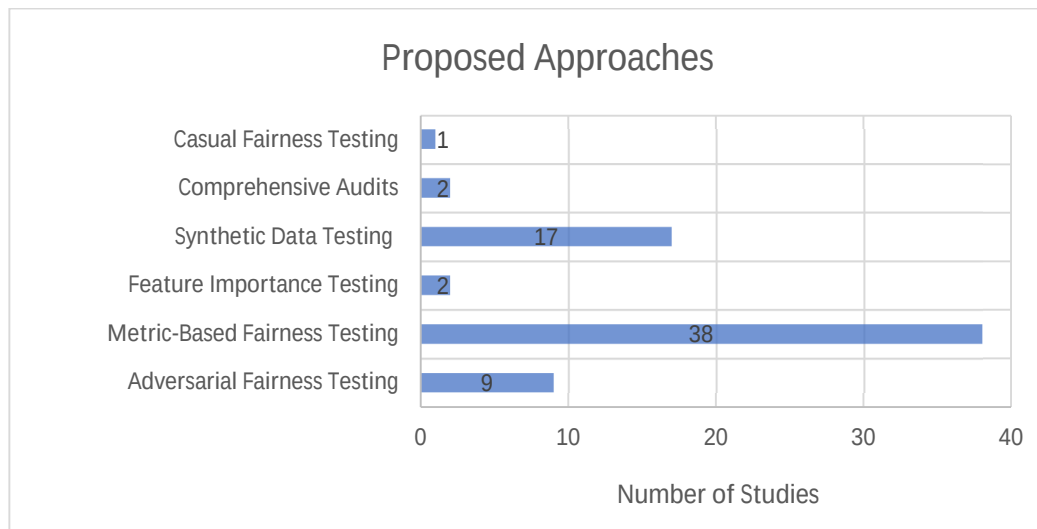


Figure 7. Summary of proposed fairness testing approaches across the reviewed studies, showing the distribution of the following methods: metric-based fairness testing (38 studies), synthetic data testing (17 studies), adversarial fairness testing (9 studies), comprehensive audits (2 studies), feature importance testing (2 studies), and causal fairness testing (1 study).

A summary of solutions associated with different problem types is provided in Table 10, covering optimization, reweighting, outlier detection, dataset augmentation, and hybrid white box/black box testing approaches.

Table 10. Summary of suggested solutions.

Problem Type	Suggested Solutions
Analyzing	Optimization methods and 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 causal 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 [7,31,41].

Importantly, part of the growing use of synthetic data (25%) is derived from the understanding that real historic data is likely embedded with historical bias. Consequently, synthetic data allows for artificial but related datasets to be built that facilitate testing fairness. Recent works increasingly rely on modern generative models such as Generative

Adversarial Networks (GANs) and diffusion-based models to generate balanced or augmented datasets. These approaches are primarily used to improve subgroup representation and to support controlled unfairness detection experiments, demonstrating a shift toward leveraging advanced generative techniques to address representational bias within fairness evaluation pipelines [23,44,45]. 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) [4,45]. Though not as prevalent as other methods, its value potentially lies chiefly in its ability to expose vulnerabilities that would otherwise go unnoticed.

The predominance of aggregated testing levels demonstrates that most studies are at least aware of measuring at the input level, since manipulation of datasets and controlled tests are highly useful for unveiling unfairness; however, some studies used more than one level, which suggests that researchers are becoming aware of the potential of using more cohesive and integrated evaluation pipelines (Figure 8). 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 by measuring end-to-end performance [25,42].

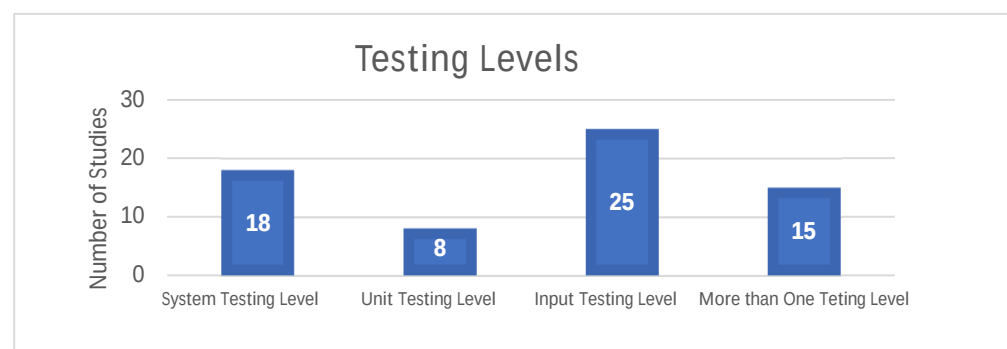


Figure 8. Distribution of primary studies across different testing levels, showing that 18 studies focus on system testing, 8 on unit testing, 25 on input-level testing, and 15 apply more than one testing level.

The limited number of studies engaging with models that are able to perform comprehensive auditing, causal testing, or feature importance (collectively less than 10%) suggests considerable room for future inquiry into these developing areas. These underutilized methodologies 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 [23,38]. Their inhibited engagements do not mean these areas are not relevant to current or future engagements with fairness.

To conclude, the results show that fairness testing in MLSs is still somewhat metric-driven and principally remains focused at the input level, though there are some studies also considering a 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 thus improve the robustness and trustworthiness of MLS applications.

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

The primary studies mainly relied on limited sets of benchmark datasets to conduct fairness detection experiments. Figure 9 shows that the most used dataset was the UCI Adult Census Income dataset ($n = 20$, 30%), followed by the German Credit dataset ($n = 15$, 22%) and the COMPAS dataset ($n = 11$, 16%). Together, these three datasets represented almost two-thirds of the dataset usage in the reviewed studies.

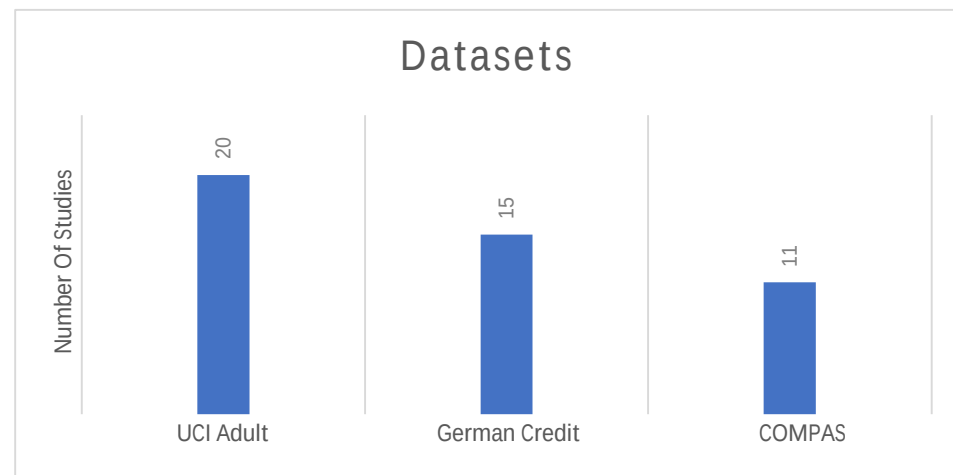


Figure 9. Distribution of the most commonly used datasets in the reviewed studies, showing that the UCI Adult Census Income dataset was used in 20 studies, the German Credit dataset in 15 studies, and the COMPAS dataset in 11 studies.

The benchmark datasets 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 Census Income dataset includes cases to test classification models when gender representation is undesirably disproportionate [46,47]. 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 [48,49]. The COMPAS dataset is widely used by researchers examining fairness in the context of the criminal justice system and is cited in multiple works, despite its documented deficiencies regarding racial bias [46,50].

Overall, the Adult Census Income, COMPAS, and German Credit datasets exhibit distinct but well-documented forms of bias rooted in their historical and institutional contexts. The Adult Census Income dataset reflects gender and racial bias arising from historical labor market inequalities and demographic underrepresentation, leading models to favor men and White individuals in income prediction tasks [46]. The COMPAS dataset is affected primarily by racial bias and label bias, where Black defendants are disproportionately misclassified as high risk due to differential error rates and arrest-based labels that reflect policing practices rather than true criminal behavior [50]. In contrast, the German Credit dataset suffers from selection bias and age- and gender-related biases, as it includes only approved loan applicants and encodes outdated financial and social norms from the 1970s, disadvantaging younger applicants and certain marital or household statuses [47]. Together, these datasets demonstrate that commonly used fairness benchmarks embed structural and contextual biases that can significantly influence fairness.

Alongside benchmark datasets, synthetic datasets were used [45,51], which allowed for controlled distribution 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 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 [32,40].

Ultimately, the study of dataset choice was a key theme influencing fairness evaluation results. For instance, PS27 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 PS19 and PS24, which indicated that the representativeness of a dataset is arguably significantly influential on the applicability of any derived fairness metrics [32,46,47].

The results indicate that fairness research in MLSs still largely relies on a few benchmark datasets, especially the Adult Census Income, German Credit, and COMPAS datasets. These popular datasets used in the service of comparability between and reproducibility of studies, but this reliance raises concerns. Each of these datasets has known biases (e.g., racial bias in COMPAS, gender bias in Adult Census Income, and demographic bias in German Credit). The reuse of these datasets 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; however, 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 indicates that fairness research has not been fully incorporated into sensitive real-world sectors where bias can have substantial consequences for 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. This study highlights an important tension in current methods that incorporate benchmarks to help guide which studies can be compared but limit fairness research. Therefore, there is still a requirement for more studies that attempt to maintain a mixture of benchmarks with more diverse, representative, and domain-related datasets that integrate aspects of fairness evaluation and a 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 MLSs. As shown in Figure 10, the most frequently mentioned issue was the influence of training dataset quality ($n = 27$, 40%), followed closely by pre-processing 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.

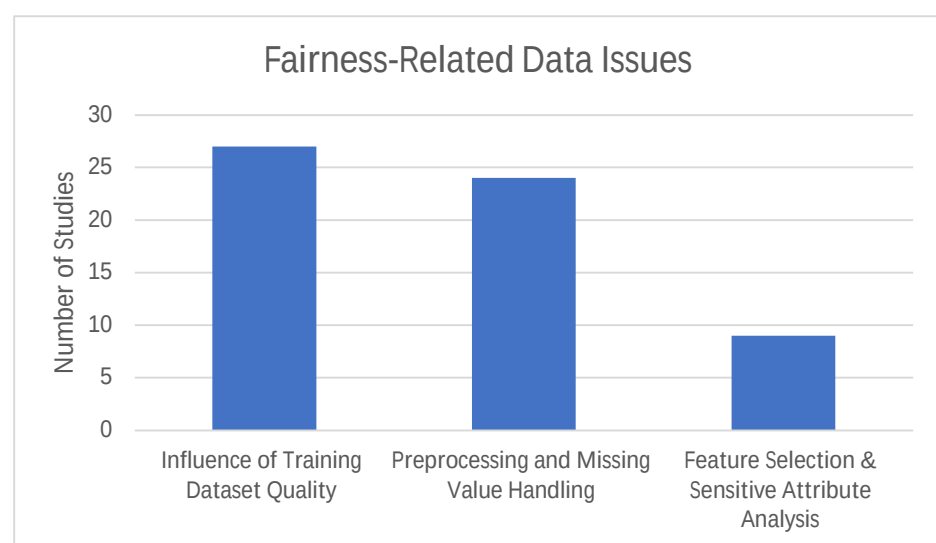


Figure 10. Fairness-related data issues identified across the reviewed studies, including the influence of training dataset quality (27 studies), pre-processing and missing-value handling (24 studies), and feature selection and sensitive attribute analysis (9 studies).

An ongoing gap across many studies is the lack of consistent fairness metrics. Many metrics exist—e.g., demographic parity, equal opportunity, predictive parity—but none fully capture fairness across contexts. PS7 and PS35 noted that researchers often applied metrics inconsistently across studies, and comparisons were difficult [12,44].

Another gap was the persistence of biased benchmark datasets. Datasets that we use often, such as the COMPAS or Adult Census Income datasets, are made with structural biases. PS27 and PS22 suggested that using the data without adjusting those datasets runs the risk of adding discrimination, even if you use fairness-aware algorithms [47,49].

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. PS25 and PS58 emphasized the necessity of the integrated frameworks aimed at fairness checks during the data collection and deployment processes [46,52].

Future research should address the lack of methodological consistency in fairness-aware machine learning by adopting standardized, multi-metric evaluation practices that explicitly justify metric selection and report trade-offs between competing fairness definitions. In addition, greater emphasis is needed on data-centric fairness, including systematic dataset auditing, bias documentation, and the use of reweighting, resampling, or synthetic data to mitigate structural biases in widely used benchmarks, such as the COMPAS and Adult Census Income datasets. Finally, there is a need for integrated, lifecycle-aware fairness testing frameworks that embed fairness checks across data collection, model development, deployment, and monitoring. Treating fairness as a first-class software quality attribute, supported by continuous testing and human-in-the-loop oversight, represents a promising direction for improving the reliability and real-world impact of fairness-aware systems [12,48,52].

Several current trends emerged in the literature. Multiple articles used fairness methods to deepen learning and natural language processing (NLP) models. PS52 and PS41 examined fairness-aware word embedding 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 [53,54].

Finally, interdisciplinary collaboration was highlighted in several studies. In PS55, 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 pre-processing issues were mentioned in over three-quarters of the reviewed studies (Figure 10) 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 MLSs, but they do not provide a complete integrated picture.

The themes emerging in deep learning and NLP research show that the field is reacting to the increasing influence of the models in more sensitive domains, such as language translation applications and recommendation systems, since MLSs are impacting society qua systemic racism and oppression. The use of fairness-aware embeddings and neural

regularizing techniques shows that fairness considerations are becoming more prominent and even integrated into the state of the art in machine learning [53,54]. The embrace of causal inference also indicates a movement toward more explanatory forms of modeling that focus on causal or structural sources of unfairness, and not just correlation [55].

The focus on collaborating with representatives from multiple disciplines means that fairness is seen not only as a technical issue, but also as a socio-technical issue [18,39]. 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 that a more thorough and holistic understanding of fairness in MLSs is on the horizon (Appendix B).

6. Conclusions

This study presents a systematic mapping study (SMS) of fairness in machine learning-based systems (MLSs), synthesizing the research published from 2010 to 2025. The SMS approached six research questions (RQs) systematically, investigating definitions of fairness; several types of problems, testing, and datasets; as well as identifying research gaps and emerging trends. 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 regarding research about MLSs.

The findings showed that fairness is a contested concept, with multiple operational definitions ranging from some notions 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 problem types (RQ3) showed that research was scattered across four categories: taking existing approaches to analyze, developing various methods for bias mitigation 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 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, and UCI Adult Census Income datasets) 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. RQ5 reveals that shortcomings in dataset quality, diversity, and transparency fundamentally limit fairness research outcomes. Addressing these challenges requires moving beyond algorithm-focused solutions toward rigorous, context-aware, and data-centric research practices, enabling more reliable and socially grounded fairness evaluations. 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.

Regarding 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, this study contributes 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 study cements this research space for future studies. The next steps will depend on building standardized fairness metrics, designing more representative datasets, and implementing robust fairness testing frameworks for all lifecycle aspects of MLSs. Ultimately, the challenges outlined above are not solely technical challenges but also societal challenges, which must be overcome to ensure that MLSs are trusted in real-life, high-stake applications, such as healthcare, finance, or justice. By standardized fairness metrics, we mean shared evaluation and reporting conventions rather than a single universal measure, such as consistently reporting a common set of baseline disparity metrics alongside task-specific ones and clearly stating underlying assumptions. Likewise, more representative datasets can be pursued through practical means compatible with privacy and access constraints, including federated or aggregate data collection, careful use of synthetic data to improve subgroup coverage, and transparent documentation of underrepresented populations.

Author Contributions: Conceptualization, R.A. and N.A.; methodology, R.A. and N.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 Nouredine 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 for Science and Technology are also gratefully acknowledged. During the preparation of this manuscript, the authors used their digital libraries to collect primary studies; the authors have reviewed and verified all outputs and take 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
GPT	Generative Pre-trained Transformer

Appendix A. Quality Assurance

Table A1. Quality assessment of collected primary studies. The full primary studies are numbered from S1 to S71. The primary studies after the quality assessment performed are numbered from PS1–PS67. The highlighted papers are excluded due to their score.

ID	Reference	Title	QA1	QA2	QA3	Score	ID After QA
S1	[1]	Using data mining techniques to discover bias patterns in missing data.	1	1	1	3	PS1
S2	[2]	Adaptive sensitive reweighting to mitigate bias in fairness-aware classification.	1	1	0.5	2.5	PS2
S3	[3]	Fairness definitions explained.	1	1	0	2	PS3
S4	[4]	Addressing Fairness, Bias, and Class Imbalance in Machine Learning: the FBI-loss.	1	1	1	3	PS4
S5		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	
S6	[5]	Improving machine learning fairness with sampling and adversarial learning.	1	1	0.5	2.5	PS5
S7		dalex: Responsible machine learning with interactive explainability and fairness in Python 3.9.	1	0.5	0	1.5	
S8	[7]	Impact of Imputation Strategies on Fairness in Machine Learning.	1	1	1	3	PS6
S9	[12]	How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness.	1	1	1	3	PS7
S10	[14]	MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software.	1	0.5	1	2.5	PS8
S11	[18]	Fairness Testing of Machine Learning Models Using Deep Reinforcement Learning.	1	1	1	3	PS9
S12	[19]	Flexible Regularization Approaches for Fairness in Deep Learning.	1	1	0.5	2.5	PS10
S13		The risk of machine learning bias (and how to prevent it).	1	0.5	0	1.5	
S14	[20]	Fairness constraints: A flexible approach for fair classification.	1	1	1	3	PS11
S15	[21]	Distribution Bias Aware Collaborative Generative Adversarial Network for Imbalanced Deep Learning in Industrial IoT.	1	0.5	0.5	2	PS12
S16	[22]	Image fairness in deep learning: problems, models, and challenges.	1	0.5	0.5	2	PS13

Table A1. Cont.

ID	Reference	Title	QA1	QA2	QA3	Score	ID After QA
S17	[23]	A comparative study of fairness-enhancing interventions in machine learning.	1	1	0.5	2.5	PS14
S18	[25]	Automatic fairness testing of machine learning models.	1	1	1	3	PS15
S19	[26]	Generating Biased Dataset for Metamorphic Testing of Machine Learning Programs.	1	0.5	0.5	2	PS16
S20	[27]	Fairness in machine learning.	1	0.5	0.5	2	PS17
S21	[30]	Fairness in deep learning: A computational perspective.	1	0.5	0.5	2	PS18
S22	[32]	FairLOF: Fairness in Outlier Detection.	1	1	1	3	PS19
S23	[33]	Fairness of Machine Learning Algorithms for the Black Community.	1	0.5	0.5	2	PS20
S24	[34]	Black box fairness testing of machine learning models.	1	1	1	3	PS21
S25	[35]	Search-based fairness testing for regression-based machine learning systems.	1	1	1	3	PS22
S26	[48]	Stable and fair classification.	1	0.5	0.5	2	PS23
S27	[49]	Automatic fairness testing of neural classifiers through adversarial sampling.	1	0.5	0.5	2	PS24
S28	[46]	Assessing gender bias in machine translation: a case study with Google Translate.	1	1	1	3	PS25
S29	[50]	Fair-SSL: Building fair ML Software with less data.	1	0.5	0.5	2	PS26
S30	[47]	Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms.	1	1	0.5	2.5	PS27
S31	[36]	Bias mitigation for fair automation of classification tasks.	1	1	1	3	PS28
S32	[31]	Algorithmic decision-making and the cost of fairness.	1	1	0.5	2.5	PS29
S33	[37]	White box Fairness Testing through Adversarial Sampling.	1	0.5	1	2.5	PS30
S34	[38]	Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation.	1	1	1	3	PS31
S35	[39]	50 years of test (un)fairness: Lessons for machine learning.	1	0.5	0.5	2	PS32
S36	[40]	Non-discriminatory machine learning through convex fairness criteria.	1	1	0.5	2.5	PS33
S37	[41]	Facing Many Objectives for Fairness in Machine Learning.	1	0.5	1	2.5	PS34

Table A1. Cont.

ID	Reference	Title	QA1	QA2	QA3	Score	ID After QA
S38	[42]	Analyzing the impact of missing values and selection bias on fairness.	1	1	1	3	PS35
S39	[43]	Automating fairness configurations for machine learning.	1	0.5	1	2.5	PS36
S40	[44]	Fairness in machine learning with tractable models.	1	1	0.2	2.5	PS37
S41	[45]	A combinatorial approach to fairness testing of machine learning models.	1	1	1	3	PS38
S42	[53]	Fairness in design: A framework for facilitating ethical artificial intelligence designs.	1	0.5	0.5	2	PS39
S43	[51]	FairNeuron: improving deep neural network fairness with adversary games on selective neurons.	1	1	1	3	PS40
S44	[52]	Neuronfair: Interpretable white box fairness testing through biased neuron identification.	1	1	0.5	2.5	PS41
S45	[54]	Fairrec: fairness testing for deep recommended systems.	0.5	1	0.5	2	PS42
S46	[55]	Trained biased number representation for ReRAM-based neural network accelerators.	1	1	1	3	PS43
S47	[56]	On convexity and bounds of fairness-aware classification.	1	1	0.5	2.5	PS44
S48	[57]	Verifying Fairness in Quantum Machine Learning.	1	1	1	3	PS45
S49	[58]	Classifying without discriminating.	1	1	0.5	2.5	PS46
S50	[59]	Fairness-aware classifier with prejudice remover regularizer.	1	1	0.5	2.5	PS47
S51	[60]	Three naive Bayes approaches for discrimination-free classification.	1	1	0.5	2.5	PS48
S52	[61]	A convex framework for fair regression.	1	1	0.5	2.5	PS49
S53	[62]	Classification with fairness constraints: A meta-algorithm with provable guarantees.	1	1	1	3	PS50
S54	[63]	Algorithmic Fairness from a Non-ideal Perspective.	1	1	0.5	2.5	PS51
S55	[64]	Equalizing gender biases in neural machine translation with word embedding techniques.	1	1	1	3	PS52
S56	[65]	Measuring Fairness of Text Classifiers via Prediction Sensitivity.	1	1	0.5	2.5	PS53
S57	[66]	Fair classification with noisy protected attributes: A framework with provable guarantees.	1	1	0.5	2.5	PS54

Table A1. Cont.

ID	Reference	Title	QA1	QA2	QA3	Score	ID After QA
S58	[67]	Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems.	1	0.5	0.5	PS2	PS55
S59	[68]	Ensuring fairness beyond the training data.	1	0.5	0.5	2	PS56
S60	[69]	Fairness-aware configuration of machine learning libraries.	1	1	0.5	2.5	PS57
S61	[70]	Training data debugging for the fairness of machine learning software.	1	0.5	0.5	2	PS58
S62	[71]	Explanation-guided fairness testing through genetic algorithm.	1	1	0.5	2.5	PS59
S63	[72]	Fairness evaluation in deepfake detection models using metamorphic testing.	1	1	1	3	PS60
S64	[73]	Algorithmic fairness datasets: the story so far.	1	0.5	0.5	2	PS61
S65	[74]	Fair enough: Searching for sufficient measures of fairness.	1	0.5	0.5	2	PS62
S66	[75]	Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking.	1	1	1	3	PS63
S67	[76]	Search-based automatic repair for fairness and accuracy in decision-making software.	1	1	0.5	2.5	PS64
S68	[77]	Machine Learning-Based Multilevel Intrusion Detection Approach.	1	1	0.5	2.5	PS65
S69	[78]	Testing Individual Fairness in Graph Neural Networks.	1	0.5	0.5	2	PS66
S70	[79]	GFairHint: Improving Individual Fairness for Graph Neural Networks via Fairness Hint.	1	1	1	3	PS67
S71		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	

Appendix B. Research Gap Summary

Table A2. 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. - Investigate measures for fairness. - Improve the definitions of fairness in data analysis. - Study other fairness metrics beyond individual fairness. - Much research relies on specific fairness metrics while ignoring others.	PS4 PS19 PS22 PS32 PS37 PS55 PS61
	Fairness Across ML Tasks	- Include fairness constraints in supervised (regression, recommendation) and unsupervised (set selection, ranking) tasks. - Evaluate fairness in natural language understanding, resource allocation, representation learning, and causal learning. - Evaluate fairness in regression-based systems. - Identify limitations in fair regression. - Evaluate fairness in CNNs and DNNs. - Extend DNN fairness testing to CNNs.	PS10 PS11 PS15 PS42 PS49 PS47 PS55
	Bias Mitigation Techniques	- Develop more effective mitigation strategies beyond simple ML modifications. - Investigate retraining as a solution to genetic fairness testing. - Explore hyper-parameter configurations that lead to high fairness. - Improve semi-supervised techniques to achieve fairness with limited labeled data. - Investigate ways to adjust the learning process to account for biases.	PS12 PS26 PS28 PS51 PS57 PS59
	Sensitive Attributes and Social Constructs	- Study numerical attributes and groups of attributes as sensitive. - Consider income, race, religious beliefs, age, and nationality. - Explore multiple and continuous sensitive features. - Study other social constructs and stereotypes.	PS46 PS48 PS52
	Fairness Testing Frameworks	- Extend reinforcement learning-based testing for fairness. - Extend Themis for algorithmic bias testing. - Apply FairRec for multi-attribute group fairness testing. - Conduct white box and black box fairness testing. - Explore other measures of fairness in white box testing. - Develop techniques for black box testing to detect individual discrimination.	PS15 PS22 PS31 PS38 PS42
	Fairness in Real-World Contexts	- Evaluate fairness–accuracy trade-offs in real-world scenarios. - Evaluate 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, such as privacy and explainability.	PS50 PS62

Table A2. Cont.

Category	Sub-Category	Research Gap	Primary Study ID
2. Feature Selection and Data Handling		• Improve the quality of datasets with test case generation approaches.	PS1
		• Evaluate feature selection in large datasets.	PS6
		• Investigate advanced feature selection techniques.	PS7
		• Add datasets, algorithms, and imputation strategies.	PS20
		• Conduct pre-processing of missing training data.	PS33
		• Conduct pre-processing of training data.	PS35
		• Study other datasets and fairness–accuracy trade-offs.	PS56
		• Study other datasets and fairness–accuracy trade-offs.	PS57
		• Explore different characteristics in training data.	PS65
3. Model Optimization and Training Efficiency		• Identify new ways for faster training processes.	PS2
		• Train CNNs with low-precision weights.	PS3
		• Improve algorithms for efficiency and equity.	PS12
		• Improve stability by shifting decision boundaries.	PS28
		• Implement models in industrial IoT for imbalanced learning.	PS29
		• Improve quantum decision models with fairness guarantees.	PS45
		• Enhance the scalability of fairness testing for large datasets and complex graphs.	PS66
4. Testing and Evaluation Frameworks		• Improve white box models for robustness.	PS21
		• Explore test case generation approaches.	PS24
		• Conduct DNN white box testing challenges.	PS27
		• Test different equivalence classes for retraining.	PS30
		• Study bias kernel detected via verification algorithms.	PS41
		• Investigate neuron coverage as a distortion metric.	PS45
		• Explore causal techniques, post-processing, interpretability, and calibration.	PS53
		• Extend counterfactual thinking to text and image processing.	PS62
5. Reinforcement Learning and Reward Functions		• Study other definitions of reward functions in black box testing.	PS9
		• Extend RL-based testing frameworks.	
		• Determine the ideal G-ratio for fairness testing.	
6. Ethics, Law, and Societal Impact		• Identify types of laws or regulations.	PS5
		• Consider ethical difficulties in statistical machine translation.	
		• Algorithms raise complex questions for researchers and policymakers.	
		• Facilitate more discussion on fairness, including technical and cultural causes.	
		• Investigate transparency in model complexity and fairness trade-off.	
		• Investigate algorithmic choices and social context.	
		• Develop fair and unified solutions.	
7. Emerging Applications and Techniques		• Promote interdisciplinary collaborations (CS, statistics, and cognitive science).	PS14
		• Investigate deep clustering, adversarial training, and attacks.	PS23
		• The study proposed methods with image compression and deepfake techniques.	PS54
		• Investigate noise models for non-binary attributes.	PS60
		• Conduct long-term studies on bias mitigation effects.	

References

1. Tremblay, M.C.; Dutta, K.; Vandermeer, D. Using data mining techniques to discover bias patterns in missing data. *J. Data Inf. Qual.* **2010**, *2*, 2. [\[CrossRef\]](#)
2. Krasanakis, E.; Spyromitros-Xioufis, E.; Papadopoulos, S.; Kompatsiaris, Y. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*; ACM: New York, NY, USA, 2018.
3. Verma, S.; Rubin, J. Fairness definitions explained. In *IEEE/ACM International Workshop on Software Fairness (Fairware)*; IEEE: New York, NY, USA, 2018.
4. Ferrari, E.; Bacciu, D. Addressing Fairness, Bias and Class Imbalance in Machine Learning: The FBI-loss. *arXiv* **2021**, arXiv:2105.06345. [\[CrossRef\]](#)
5. Amend, J.J.; Spurlock, S. Improving machine learning fairness with sampling and adversarial learning. *J. Comput. Sci. Coll.* **2021**, *36*, 14–23.
6. Foidl, H.; Felderer, M. Risk-based data validation in machine learning-based software systems. In *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation*; ACM: New York, NY, USA, 2019.
7. Caton, S.; Malisetty, S.; Haas, C. Impact of Imputation Strategies on Fairness in Machine Learning. *J. Artif. Intell. Res.* **2022**, *74*, 1011–1035. [\[CrossRef\]](#)
8. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 115. [\[CrossRef\]](#)
9. Pessach, D.; Shmueli, E. A review on fairness in machine learning. *ACM Comput. Surv.* **2022**, *55*, 51. [\[CrossRef\]](#)
10. Das Jui, T.; Rivas, P. Fairness issues, current approaches, and challenges in machine learning models. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 3095–3125. [\[CrossRef\]](#)
11. Chen, Z.; Zhang, J.M.; Hort, M.; Harman, M.; Sarro, F. Fairness testing: A comprehensive survey and analysis of trends. *ACM Trans. Softw. Eng. Methodol.* **2024**, *33*, 137. [\[CrossRef\]](#)
12. Valdivia, A.; Sánchez-Monedero, J.; Casillas, J. How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *Int. J. Intell. Syst.* **2021**, *36*, 1619–1643. [\[CrossRef\]](#)
13. Caton, S.; Haas, C. Fairness in machine learning: A survey. *arXiv* **2020**, arXiv:2010.04053. [\[CrossRef\]](#)
14. Chen, Z.; Zhang, J.M.; Sarro, F.; Harman, M. MAAT: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*; ACM: New York, NY, USA, 2022.
15. Richardson, B.; Gilbert, J.E. A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *arXiv* **2021**, arXiv:2112.05700. [\[CrossRef\]](#)
16. Alkatheri, M.S. Testing Machine-Learning-Based Software for Fairness: Ontology-Guided Test Cases Generation. Doctoral Dissertation, King Saud University, Riyadh, Saudi Arabias, 2022.
17. Riccio, V.; Jahangirova, G.; Stocco, A.; Humbatova, N.; Weiss, M.; Tonella, P. Testing machine learning based systems: A systematic mapping. *Empir. Softw. Eng.* **2020**, *25*, 5193–5254. [\[CrossRef\]](#)
18. Xie, W.; Wu, P. Fairness Testing of Machine Learning Models Using Deep Reinforcement Learning. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*; IEEE: New York, NY, USA, 2020.
19. Olfat, M.; Mintz, Y. Flexible Regularization Approaches for Fairness in Deep Learning. In *2020 59th IEEE Conference on Decision and Control (CDC)*; IEEE: New York, NY, USA, 2020.
20. Zafar, M.B.; Valera, I.; Gomez-Rodriguez, M.; Gummadi, K.P. Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.* **2019**, *20*, 2737–2778.
21. Zhou, X.; Hu, Y.; Wu, J.; Liang, W.; Ma, J.; Jin, Q. Distribution Bias Aware Collaborative Generative Adversarial Network for Imbalanced Deep Learning in Industrial IoT. *IEEE Trans. Ind. Inform.* **2022**, *19*, 570–580. [\[CrossRef\]](#)
22. Tian, H.; Zhu, T.; Liu, W.; Zhou, W. Image fairness in deep learning: Problems, models, and challenges. *Neural Comput. Appl.* **2022**, *34*, 12875–12893. [\[CrossRef\]](#)
23. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.P.; Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2019.
24. Ahuja, R.; Chug, A.; Gupta, S.; Ahuja, P.; Kohli, S. Classification and clustering algorithms of machine learning with their applications. In *Nature-Inspired Computation in Data Mining and Machine Learning*; Springer International Publishing: Cham, Switzerland, 2019; pp. 225–248.
25. Sharma, A.; Wehrheim, H. Automatic fairness testing of machine learning models. In *IFIP International Conference on Testing Software and Systems*; Springer: Cham, Switzerland, 2020.
26. Nakajima, S.; Chen, T.Y. Generating biased dataset for metamorphic testing of machine learning programs. In *IFIP International Conference on Testing Software and Systems*; Springer: Cham, Switzerland, 2019.

27. Oneto, L.; Chiappa, S. Fairness in machine learning. In *Recent Trends in Learning from Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*; Springer International Publishing: Cham, Switzerland, 2020.
28. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; Version 2.3; EBSE Technical Report; EBSE: Durham, UK, 2007; Volume 5.
29. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*; BCS Learning & Development: Swindon, UK, 2008.
30. Du, M.; Yang, F.; Zou, N.; Hu, X. Fairness in deep learning: A computational perspective. *IEEE Intell. Syst.* **2020**, *36*, 25–34. [\[CrossRef\]](#)
31. Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2017.
32. Deepak, P.; Abraham, S.S. FairLOF: Fairness in Outlier Detection. *Data Sci. Eng.* **2021**, *6*, 485–499. [\[CrossRef\]](#)
33. Kiemde, S.M.A.; Kora, A.D. Fairness of Machine Learning Algorithms for the Black Community. In *2020 IEEE International Symposium on Technology and Society (ISTAS)*; IEEE: New York, NY, USA, 2020.
34. Aggarwal, A.; Lohia, P.; Nagar, S.; Dey, K.; Saha, D. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*; ACM: New York, NY, USA, 2019.
35. Perera, A.; Aleti, A.; Tantithamthavorn, C.; Jiarapakdee, J.; Turhan, B.; Kuhn, L.; Walker, K. Search-based fairness testing for regression-based machine learning systems. *Empir. Softw. Eng.* **2022**, *27*, 79. [\[CrossRef\]](#)
36. Consuegra-Ayala, J.P.; Gutiérrez, Y.; Almeida-Cruz, Y.; Palomar, M. Bias mitigation for fair automation of classification tasks. *Expert Syst.* **2024**, *42*, e13734. [\[CrossRef\]](#)
37. Zhang, P.; Wang, J.; Sun, J.; Dong, G.; Wang, X.; Wang, X.; Dong, J.S.; Dai, T. White-box fairness testing through adversarial sampling. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*; ACM: New York, NY, USA, 2020.
38. Bantilan, N. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *J. Technol. Hum. Serv.* **2018**, *36*, 15–30. [\[CrossRef\]](#)
39. Hutchinson, B.; Mitchell, M. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2019.
40. Goel, N.; Yaghini, M.; Faltings, B. Non-discriminatory machine learning through convex fairness criteria. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 116. [\[CrossRef\]](#)
41. Villar, D.; Casillas, J. Facing Many Objectives for Fairness in Machine Learning. In *International Conference on the Quality of Information and Communications Technology*; Springer: Cham, Switzerland, 2021.
42. Wang, Y.; Singh, L. Analyzing the impact of missing values and selection bias on fairness. *Int. J. Data Sci. Anal.* **2021**, *12*, 101–119. [\[CrossRef\]](#)
43. Sun, H.; Yang, Y.; Li, Y.; Liu, H.; Wang, X.; Wang, W.H. Automating fairness configurations for machine learning. In *Companion Proceedings of the Web Conference*; ACM: New York, NY, USA, 2021.
44. Varley, M.; Belle, V. Fairness in machine learning with tractable models. *Knowl.-Based Syst.* **2021**, *215*, 106715. [\[CrossRef\]](#)
45. Patel, A.R.; Chandrasekaran, J.; Lei, Y.; Kacker, R.N.; Kuhn, D.R. A combinatorial approach to fairness testing of machine learning models. In *2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*; IEEE: New York, NY, USA, 2022.
46. Prates, M.O.R.; Avelar, P.H.; Lamb, L.C. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Comput. Appl.* **2020**, *32*, 6363–6381. [\[CrossRef\]](#)
47. Jones, G.P.; Hickey, J.M.; Di Stefano, P.G.; Dhanjal, C.; Stoddart, L.C.; Vasileiou, V. Metrics and methods for a systematic comparison of fairness-aware machine learning algorithms. *arXiv* **2020**, arXiv:2010.03986. [\[CrossRef\]](#)
48. Huang, L.; Vishnoi, N. Stable and fair classification. In *International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2019.
49. Zhang, P.; Wang, J.; Sun, J.; Wang, X.; Dong, G.; Wang, X.; Dai, T.; Dong, J.S. Automatic fairness testing of neural classifiers through adversarial sampling. *IEEE Trans. Softw. Eng.* **2021**, *48*, 3593–3612. [\[CrossRef\]](#)
50. Chakraborty, J.; Majumder, S.; Tu, H. Fair-SSL: Building fair ML Software with less data. In *Proceedings of the 2nd International Workshop on Equitable Data and Technology*; ACM: New York, NY, USA, 2022.
51. Gao, X.; Zhai, J.; Ma, S.; Shen, C.; Chen, Y.; Wang, Q. FairNeuron: Improving deep neural network fairness with adversary games on selective neurons. In *Proceedings of the 44th International Conference on Software Engineering*; ACM: New York, NY, USA, 2022.
52. Zheng, H.; Chen, Z.; Du, T.; Zhang, X.; Cheng, Y.; Ji, S.; Wang, J.; Yu, Y.; Chen, J. Neuronfair: Interpretable white-box fairness testing through biased neuron identification. In *Proceedings of the 44th International Conference on Software Engineering*; ACM: New York, NY, USA, 2022.
53. Zhang, J.; Shu, Y.; Yu, H. Fairness in design: A framework for facilitating ethical artificial intelligence designs. *Int. J. Crowd Sci.* **2023**, *7*, 32–39. [\[CrossRef\]](#)

54. Guo, H.; Li, J.; Wang, J.; Liu, X.; Wang, D.; Hu, Z.; Zhang, R.; Xue, H. Fairrec: Fairness testing for deep recommender systems. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis*; ACM: New York, NY, USA, 2023.
55. Wang, W.; Lin, B. Trained biased number representation for ReRAM-based neural network accelerators. *ACM J. Emerg. Technol. Comput. Syst.* **2019**, *15*, 15. [[CrossRef](#)]
56. Wu, Y.; Zhang, L.; Wu, X. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*; ACM: New York, NY, USA, 2019.
57. Guan, J.; Fang, W.; Ying, M. Verifying Fairness in Quantum Machine Learning. In *International Conference on Computer Aided Verification*; Springer: Cham, Switzerland, 2022.
58. Kamiran, F.; Calders, T. Classifying without discriminating. In *2nd International Conference on Computer, Control, and Communication*; IEEE: New York, NY, USA, 2009.
59. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2012.
60. Calders, T.; Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* **2010**, *21*, 277–292. [[CrossRef](#)]
61. Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; Roth, A. A convex framework for fair regression. *arXiv* **2017**, arXiv:1706.02409. [[CrossRef](#)]
62. Celis, L.E.; Huang, L.; Keswani, V.; Vishnoi, N.K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2019.
63. Fazelpour, S.; Lipton, Z.C. Algorithmic Fairness from a Non-ideal Perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*; ACM: New York, NY, USA, 2020.
64. Font, J.E.; Costa-Jussà, M.R. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv* **2019**, arXiv:1901.03116. [[CrossRef](#)]
65. Krishna, S.; Gupta, R.; Verma, A.; Dhamala, J. Measuring Fairness of Text Classifiers via Prediction Sensitivity. *arXiv* **2022**, arXiv:2203.08670. [[CrossRef](#)]
66. Celis, L.E.; Huang, L.; Keswani, V.; Vishnoi, N.K. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021.
67. Shrestha, Y.R.; Yang, Y. Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms* **2019**, *12*, 199. [[CrossRef](#)]
68. Mandal, D.; Deng, S.; Jana, S.; Wing, J.; Hsu, D.J. Ensuring fairness beyond the training data. In *Advances in Neural Information Processing Systems 33 (2020)*; NeurIPS Proceedings: San Diego, CA, USA, 2020; pp. 18445–18456.
69. Tizpaz-Niari, S.; Kumar, A.; Tan, G.; Trivedi, A. Fairness-aware configuration of machine learning libraries. In *Proceedings of the 44th International Conference on Software Engineering*; ACM: New York, NY, USA, 2022.
70. Li, Y.; Meng, L.; Chen, L.; Yu, L.; Wu, D.; Zhou, Y.; Xu, B. Training data debugging for the fairness of machine learning software. In *Proceedings of the 44th International Conference on Software Engineering*; ACM: New York, NY, USA, 2022.
71. Fan, M.; Wei, W.; Jin, W.; Yang, Z.; Liu, T. Explanation-guided fairness testing through genetic algorithm. In *Proceedings of the 44th International Conference on Software Engineering*; ACM: New York, NY, USA, 2022.
72. Pu, M.; Kuan, M.Y.; Lim, N.T.; Chong, C.Y.; Lim, M.K. Fairness evaluation in deepfake detection models using metamorphic testing. In *2022 IEEE/ACM 7th International Workshop on Metamorphic Testing (MET)*; IEEE: New York, NY, USA, 2022.
73. Fabris, A.; Messina, S.; Silvello, G.; Susto, G.A. Algorithmic fairness datasets: The story so far. *Data Min. Knowl. Discov.* **2022**, *36*, 2074–2152. [[CrossRef](#)]
74. Majumder, S.; Chakraborty, J.; Bai, G.R.; Stolee, K.T.; Menzies, T. Fair enough: Searching for sufficient measures of fairness. *ACM Trans. Softw. Eng. Methodol.* **2023**, *32*, 134. [[CrossRef](#)]
75. Wang, Z.; Zhou, Y.; Qiu, M.; Haque, I.; Brown, L.; He, Y.; Wang, J.; Lo, D.; Zhang, W. Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking. *arXiv* **2023**, arXiv:2302.08018. [[CrossRef](#)]
76. Hort, M.; Zhang, J.M.; Sarro, F.; Harman, M. Search-based automatic repair for fairness and accuracy in decision-making software. *Empir. Softw. Eng.* **2024**, *29*, 36. [[CrossRef](#)]
77. Ling, J.; Zhang, L.; Liu, C.; Xia, G.; Zhang, Z. Machine Learning-Based Multilevel Intrusion Detection Approach. *Electronics* **2025**, *14*, 323. [[CrossRef](#)]
78. Nasiri, R. Testing Individual Fairness in Graph Neural Networks. *arXiv* **2025**, arXiv:2504.18353. [[CrossRef](#)]
79. Xu, P.; Zhou, Y.; An, B.; Ai, W.; Huang, F. Gfairhint: Improving individual fairness for graph neural networks via fairness hint. *ACM Trans. Knowl. Discov. Data* **2025**, *19*, 67. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.